琉球大学学術リポジトリ

教育分野におけるデータマイニングモデルと高校生の能力予測システムの開発

メタデータ	言語: 出版者: 琉球大学 公開日: 2021-06-21 キーワード (Ja): キーワード (En): 作成者: Phauk, Sokkhey メールアドレス: 所属:
URL	http://hdl.handle.net/20.500.12000/48610

Doctoral Dissertation of Engineering

Developing Educational Data Mining Models and Prediction System for Predicting High School Student Performance

March 2021

by

Sokkhey Phauk

Information Engineering

Graduate School of Engineering and Science

University of the Ryukyus

Doctoral Dissertation of Engineering

Developing Educational Data Mining Models and Prediction System for Predicting High School Student Performance

March 2021

by

Sokkhey Phauk

A dissertation submitted to the Graduate School
of Engineering and Science, University of the Ryukyus,
in partial fulfillment of the requirements for the degree of

Doctor of Engineering

Information Engineering
Interdisciplinary Intelligent Systems Engineering Course
Graduate School of Engineering and Science
University of the Ryukyus

Supervisor: Prof. Takeo Okazaki

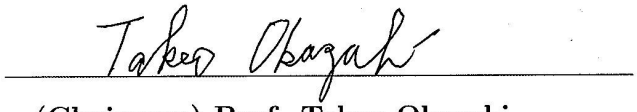
Reviewing Committee Members' Signatures

Dissertation Title: Developing Educational Data Mining Models and Prediction System for Predicting High School Student Performance.

Name: Sokkhey Phauk

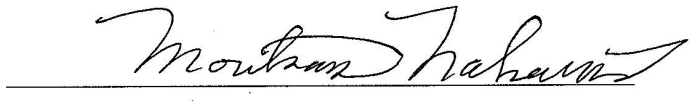
We, the undersigned, hereby, declare that we have read this dissertation and we have attended the dissertation defense and evaluation meeting. Therefore, we certify that, to the best of our knowledge this dissertation is satisfactory to the scope and quality as a dissertation for the degree of Doctor of Engineering, majoring in Data Mining under Interdisciplinary Intelligent Systems Engineering, Graduate School of Engineering and Science, University of the Ryukyus.

DISSERTATION REVIEW & EVALUATION COMMITTEE MEMBERS



(Chairman) Prof. Takeo Okazaki

(Committee) Prof. Satoshi Endo



(Committee) Prof. Morikazu Nakamura

Abstract

The challenging task in educational institutions is to enrich the quality of education by maximizing the high performance of students and minimizing the failure rate of poor-performing students. An effective solution to leverage this task is to know student learning patterns with highly influencing factors and get accurate early prediction of student learning outcomes at the timely stage for setting up policy for improvement. Educational data mining (EDM) is an emerging disciplinary field of data mining, statistics, and machine learning concerned with extracting useful knowledge and information for the sake of improvement and development in the education environment.

The study of this work is to propose developed techniques in EDM and integrate it into a web-based support system for predicting poor-performing students. A comparative study of prediction models was initially conducted. Subsequently, high performing models were further developed to get higher performance. The study proposed the developed models of hybrid machine learning models and optimization of deep belief network. The hybrid random forest (Hybrid RF) and the improved deep belief network (IDBN) generate the most successful classification. For the context of intervention, improving the learning outcomes, and enhancing model performance, the study introduced a novel feature selection method named MICHI. MICHI algorithm in which the combination of mutual information and Chi-square algorithms based on the ranked feature scores is introduced to select a dominant set and improve performance of prediction models.

The developed prediction models are integrated in a designed web-based application, called Academic Performance Prediction System (APPS). The system is designed for educational stakeholders and related individuals to get prediction of student performance at the early stage for intervention. The survey on system evaluation confirms the effectiveness of the prediction model and the usability of APPS. The designed APPS is reported to be helpful and useful for educational stakeholders such as teachers,

educational administrators, and policymakers, and related individuals to improve academic performance in educational institutions. Related individuals and educational institutions can figure out weak points of their educating, adapt teaching and learning methods, and provide an effective intervention to make improvements. Therefore, overall learning quality and learning performance can be improved greatly and reduce the failure rate of poor-performing students.

Author's Publication List

Journal Papers

1. Sokkhey, P., & Okazaki, T. (October 2019). Comparative Study of Prediction Models on High School Student Performance in Mathematics. *Journal of IEIE Transactions on Smart Processing and Computing*, 8(5), 394-404.

2. Sokkhey, P., & Okazaki, T. (January 2020). Hybrid Machine Learning Algorithms for Predicting Academic Performance. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 11(1), 32-41.

3. Sokkhey, P., & Okazaki, T. (June 2020). Multi-models of Educational Data Mining for Predicting Student Performance in Mathematics: A Case Study for High Schools in Cambodia. *Journal of IEIE Transactions on Smart Processing and Computing*, 9(3), 217-229.

4. Sokkhey, P., & Okazaki, T. (July 2020). Developing Web-based Support Systems for Predicting Poor-performing Students Using Educational Data Mining Techniques, *(IJACSA) International Journal of Advanced Computer Science and Applications*, 11(7), 23-32.

5. Sokkhey, P., & Okazaki, T. (August 2020). Development and Optimization of Deep Belief Network for Academic Prediction with Larger Datasets. *Journal of IEIE Transactions on Smart Processing and Computing*, 9(4), 298-311.

6. Sokkhey, P., & Okazaki, T. (August 2020). Study on Dominant Factors for Academic Prediction Using Feature Selection Methods. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 11(8), 492-502.

International Conference Proceedings (With Review)

1. Sokkhey, P., & Okazaki, T. (2019, June 23-26). Comparative Study of Prediction Models on High School Student Performance in Mathematics. *Proceeding of International Technical Conference on Circuit/Systems, Computer and Communication 2019 (ITC-CSCC 2019)*, Jeju, South Korea, 96-99.

2. Sokkhey, P., & Okazaki, T. (2020, October 05-06). Developing Academic Performance Prediction Systems for Predicting Poor-performing Students Using Educational Data Mining Techniques. *Proceeding of 22nd International Conference on Soft Computing, Artificial Intelligence and Applications (ICSCAIA 2020)*, Tokyo, Japan, 121-133.

Award

1. The Best Presentation Award. The 22nd International Conference on Soft Computing, Artificial Intelligence and Applications (ICSCAIA 2020), October 05-06, 2020, Tokyo, Japan.

Acknowledgment

I wish to express my gratitude and appreciation to people who have concerned, assisted, and supported along my doctoral degree journey and involved with contributing of this dissertation.

First and foremost, this dissertation would not have been possible without the guidance and the help of my advisor Prof. Takeo Okazaki. I am profoundly grateful for his mentoring, guiding, valuable support, patience, and full understanding. He is more than an advisor. I am also grateful to Prof. Satoshi Endo and Prof. Morikazu Nakamura for their patient in reading my manuscript, their advice and feedback, and support to all process of this work in spite of their busy schedules.

I acknowledge the financial support for my graduate studies to Japanese Government (Monbukagakusho: MEXT) Scholarship. It is a great opportunity for me to be awarded as a student of University of the Ryukyus and to study with researchers of such talent and intuition. Next, I would like to thank administrative and support staff in the Global Education Section, Information Engineering Office and University of the Ryukyus for their assistance and support, without whom my academic life would have been very difficult. Additionally, I would like to express my respect and thank to staff in the dormitory office for help, and assistance during my stay in Ryudai dormitory.

My deepest gratitude goes to my parents and my siblings for loving, caring, and supporting me spiritually throughout my life. Their unconditional sacrifices and pure love for education have greatly inspired me to pursue this academic dream. My heartfelt thanks go to my wife Sophealeaksmy who has been a constant supporter and a great motivator throughout this tough academic journey and my son Visal for being my daily smile and the greatest motivation to complete this task. Thank you for your care, endurance, and awaiting. I am also greatly blessed to my parents-in-law who have been encouraging and supporting me to accomplish this academic quest. I really appreciate

their support throughout these years to take care my wife and my son through their busy time. Dry food that they packed for me from home is the most delicious food. I am lucky to have them as parents-in-law. They are a part of this success.

Another gratefulness goes to my friends, seniors, juniors, researchers and staff in Ministry of Education, Youth, and Sport (MoEYS) in Cambodia in referencing , assisting, guiding in designing questionnaires, data collection, and domain understanding. Once again, I would like to thank all my friends who are teachers in various high schools and lecturers in higher educational institutions in Cambodia for their kindly help, strongly support, and tirelessly assisting in data collection. I am also particularly indebted to the academic staff and students in Cambodia who participated in this research. Without their participation and contribution, this research would not have been possible.

My regards and profound thanks also go to all of friends in my laboratory, especially my tutor Miyagi-san, and Tya-san for their friendly help and support, sharing laugh, sadness, and stress emotion during my life in Japan. Life at University of the Ryukyus would not have been the same without you guys. I will definitely miss all of you.

I am also grateful to my only Cambodian brother in Okinawa, David PICH, in guiding me on how to learn programming as an independent learner, in always introducing new learning platforms. Thanking for always visiting me and taking me around Okinawa. Academic life here would be more lonely and difficult without you.

Once again, I would like to thank to my colleague (Institute of Technology of Cambodia (ITC), Cambodia), family, friends, and all who have inspired me to step in this academic life, keep encouraging, admiring, and motivating to complete this journey.

"World and society are net, no one accomplish and succeed without help and support whether direct or indirect from beloved people"

Anonymous

Dedication

To my wife and parents-in-law

For having been waiting and the priceless support.

To my father

For have been inspiring, wishing, and waiting.

To my mother

May her soul Rest In Peace!

This achievement is dedicated to all of you.

CONTENTS

Page

Reviewing Committee Members' Signatures i

Abstract ii

Author's Publication List iv

Acknowledgment vi

Table of Contents x

List of Tables xvii

List of Figures xxi

List of Abbreviations xxii

CHAPTERS

1 Introduction 1

1.1 Domain of the Study Problem and Background 1

1.2 Mining Academic Performance 4

1.2.1 Poor Performance: Failure, Dropout, and Repeating Classes . . 5

1.2.2 Causes of Poor Academic Performance in Developing Countries 8

1.2.3 Successes in Academic Performance and Educational Settings . 10

1.3 Motivation of the Study 10

1.4 Statement of the Problem 12

1.5 Research Objectives 13

1.6 Dissertation Structure 14

2 Literature Review and State-of-the-art Methods 18
2.1 Overview 18
2.2 Applications of Educational Data Mining 19
2.3 Literature Reviews of Related Works 20
2.3.1 Features Affecting Academic Performance 20
2.3.2 Student Performance Prediction 23
2.3.3 Feature Selection Methods 27
2.3.4 Academic Prediction Systems 29
2.4 Summary Knowledge and State-of-the-art Methods 31
3 Educational Data Mining Methods 33
3.1 Introduction 33
3.2 Statistical Analysis Techniques 35
3.2.1 Structural Equation Modeling 36
3.2.2 Factor Analysis 36
3.2.3 Models of Structural Equation Modeling 38
3.2.4 Modification Fit Indexes 40
3.3 Machine Learning Algorithms 43
3.3.1 Logistic Regression 44
3.3.2 Decision Tree 46
3.3.3 Random Forest 49
3.3.4 Ensemble Methods 50
3.3.5 Support Vector Machine 51
3.3.6 Naive Bayes 55
3.3.7 *k*-Nearest Neighbors 56
3.3.8 Learning Vector Quantization 57
3.3.9 Multilayer Perceptron 59
3.4 Deep Learning Algorithms 63

3.4.1 Restricted Boltzmann Machine . . . 64
3.4.2 Deep Belief Network Architecture . . . 68
3.5 Feature Selection and Feature Extraction Methods . . . 71
3.5.1 Filter-Based Feature Selection . . . 74
3.5.2 Feature Extraction Using Principal Component Analysis . . . 78
4 Methodology for Academic Performance Prediction . . . 81
4.1 Introduction . . . 81
4.2 Domain Understanding . . . 83
4.3 Data Understanding . . . 84
4.3.1 The Related Factors Affecting Student Performance . . . 85
4.3.2 The Target Variable: Student Learning Performance Levels . . . 91
4.3.3 Synthetic Datasets . . . 91
4.3.4 Actual Dataset . . . 94
4.3.5 Evaluation on Utility of Synthetic Data . . . 97
4.4 Data Preprocessing . . . 102
4.4.1 Data Cleaning . . . 102
4.4.2 Feature Encoding . . . 103
4.4.3 Data Normalization . . . 103
4.4.4 Data Discretization . . . 103
4.4.5 Dimensional Reduction . . . 104
4.5 Data Mining or Modeling: Methods of EDM . . . 104
4.6 Model Evaluation Metrics . . . 104
4.6.1 Classification Accuracy . . . 105
4.6.2 Root Mean Square Error . . . 106
4.7 Model Deployment . . . 106
5 Comparative Study of Educational Data Mining Models . . . 108
5.1 Introduction . . . 108

5.2 The Comparative Prediction Models . . . 109
5.2.1 Statistical Analysis Techniques . . . 109
5.2.2 Machine Learning Algorithms . . . 110
5.2.3 Deep Learning Algorithms . . . 111
5.3 Comparative Study Using Synthetic Datasets . . . 111
5.3.1 Synthetic Datasets . . . 111
5.3.2 Experimental Results . . . 112
5.4 Comparative Study Using Actual Datasets . . . 116
5.4.1 Experimental Results . . . 117
5.5 Chapter Summary: Summary of Experimental Results . . . 120
6 The Proposed Hybrid Machine Learning Models . . . 121
6.1 Introduction . . . 121
6.2 The Concept for Model Improvement . . . 122
6.3 The Proposed Machine Learning Techniques . . . 123
6.3.1 The Principal Compoment Analysis . . . 124
6.3.2 Resampling Method . . . 125
6.4 Experimental Results . . . 127
6.4.1 Result of Baseline Models . . . 128
6.4.2 Experimental Results of the Proposed Models . . . 132
6.5 Chapter Summary: Summary of Experimental Results . . . 140
7 The Proposed Optimization of Deep Belief Network Model . . . 142
7.1 Introduction . . . 142
7.2 Characteristic Descriptions of Deep Belief Network Model . . . 143
7.3 Optimization Approach for Deep Belief Network Model . . . 145
7.3.1 Feature Selection . . . 146
7.3.2 Hyperparameter Optimization . . . 146
7.3.3 Regularization Methods . . . 152

7.4 Experimental Result . . . 153
7.4.1 Experimental Results of Phase I . . . 153
7.4.2 Experimental Results of Phase II . . . 155
7.4.3 Comparative Results and Discussion . . . 160
7.5 Chapter Summary: Summary of Experimental Results . . . 162
8 The Proposed FS Methods for Determining Dominant Factors . . . 163
8.1 Introduction . . . 163
8.2 Feature Selection Algorithms . . . 164
8.2.1 Existing Feature Selection Algorithms . . . 165
8.2.2 The Proposed Feature Selection Algorithm . . . 166
8.3 The Consideration on the Ranking of Feature Importance . . . 168
8.4 Classification Models . . . 173
8.4.1 *k*-Nearest Neighbors Model . . . 174
8.4.2 The Proposed Hybrid Machine Learning Models . . . 174
8.4.3 The Proposed Improved Deep Belief Network Model . . . 174
8.5 Experimental Results . . . 174
8.5.1 Experiment Results with ADS Dataset . . . 178
8.5.2 Experiment Results with GDS4 Dataset . . . 182
8.6 Discussion . . . 186
8.7 Chapter Summary: Summary of Experimental Result . . . 188
9 The Design and Implementation of APPS . . . 190
9.1 Introduction . . . 190
9.2 Academic Performance Prediction Systems Modeling . . . 191
9.3 Models for Development of the System . . . 193
9.3.1 Optimal Feature Subset . . . 193
9.3.2 Optimal Classifier . . . 194
9.4 The Design of Web Application System . . . 194

9.4.1 Component of APPS for Users . . . 194
9.4.2 The Design of APPS . . . 195
9.4.3 Web Application System Interface . . . 196
9.4.4 The System Usability . . . 199
9.4.5 The System Evaluation . . . 199
9.5 Chapter Summary: Results of APPS Development . . . 202
10 Summary and Conclusion . . . 203
10.1 Introduction . . . 203
10.2 Summary and Discussion . . . 203
10.3 Methodological Contributions . . . 206
10.3.1 Developed Optimal Prediction Models . . . 206
10.3.2 Proposed a Novel FS Algorithm . . . 206
10.4 Applicability Contributions . . . 207
10.4.1 The web-based APPS for Predicting Student Performance . . . 207
10.4.2 Intervention: Recommendation and Action Strategies . . . 208
10.4.3 Managerial Settings, Planning and Scheduling . . . 209
References . . . 211
Appendix . . . 238

LIST OF TABLES

Table		Page
1.1	System of Education in Cambodia	3
3.1	Evaluation of Fit Indexes	43
3.2	Comparison of Decision Tree Classifiers	48
4.1	Factors Affecting Student Performance	90
4.2	The Target Variables	91
4.3	The 5-point Likert Scale Measurement	95
4.4	Confusion Matrix	105
5.1	Performance of the Proposed Prediction Models	112
5.2	Accuracy Comparison of Comparative Models on dataset DS1	114
5.3	Accuracy Comparison of Comparative Models on dataset DS2	114
5.4	Accuracy Comparison of Comparative Models on dataset DS3	114
5.5	The Model Performance base on ACC	117
5.6	The Model Performance base on RMSE	119
6.1	Performance Comparison on Dataset ADS	128
6.2	Performance Comparison on Dataset GDS2	129
6.3	Performance Comparison on Dataset GDS3	130
6.4	Performance Comparison of Proposed Models on Dataset ADS	133
6.5	Performance Comparison of Proposed Models on Dataset GDS2	135
6.6	Performance Comparison of Proposed Models on Dataset GDS3	138

7.1 The Optimal Parameters 152
7.2 Performance Comparison of the Proposed Models of ADS 154
7.3 Performance Comparison of the Proposed Models of GDS1 156
7.4 Performance Comparison of the Proposed Models of GDS2 156
7.5 Performance Comparison of the Proposed Models of GDS3 156
7.6 Performance Comparison of the Proposed Models of GDS4 157
8.1 Comparison of Feature Selection Algorithms 165
8.2 The Comparative Ranking on the Importance of Predictive Features 172
8.3 Feature Ranking of the Proposed Feature Selection Methods 177
8.4 The Model Performance on Dataset ADS without Feature Selection 178
8.5 The KNN Performance on Dataset ADS with Feature Selection 179
8.6 Hybrid C5.0 Performance on Dataset ADS with Feature Selections 180
8.7 Hybrid RF Performance on Dataset ADS with Feature Selections 180
8.8 The IDBN Performance on Dataset ADS with Feature Selections 181
8.9 The Model Performance on Dataset GDS4 without Feature Selection 182
8.10 KNN Performance on Dataset GDS4 with Feature Selections 183
8.11 Hybrid C5.0 Performance on Dataset GDS4 with Feature Selections 184
8.12 Hybrid RF Performance on Dataset GDS4 with Feature Selections 184
8.13 IDBN Performance on Dataset GDS4 with Feature Selections 185
8.14 Comparison of Optimal Classifiers 187
9.1 Evaluation Survey Results. 200

LIST OF FIGURES

Figure **Page**

1.1 Conceptual model of student performance 5

1.2 Poor performance: Failure and repeating class rate 6

1.3 Self-System model of motivational development linking academic performance 7

1.4 Research summary (research roadmap) of this dissertation 14

2.1 The cycle of applying data mining in educational systems 20

2.2 Literature reviews and state-of-the-art methods 31

3.1 The Proposed baseline educational data mining techniques 35

3.2 Threshold model of SEM 40

3.3 Random forest algorithm 49

3.4 Ensemble methods with Bagging algorithms 50

3.5 Ensemble methods with Boosting algorithms 51

3.6 Support vector machine classification 52

3.7 k-nearest neighbor 57

3.8 The structure of neural network 60

3.9 Perceptron Structure 60

3.10 Scheme graph of a MLP network 61

3.11 Sigmoid and ReLu functions 62

3.12 Illustration of BM and RBM 65

3.13 Illustration of single step of CD or MCD algorithm 67

3.14 Layer by layer training of Deep Belief Network 70
3.15 Framework of supervised DBN Model 71
3.16 Feature engineering: Feature selection and feature extraction 72
4.1 Outline of CRISP-DM process 82
4.2 Systematic steps in data collection and preprocessing 84
4.3 Factors affecting student performance 85
4.4 Structural model for generating synthetic datasets 93
4.5 Samples of educational dataset for this prediction problem 97
4.6 Distibution of all features of synthetic and real datasets 99
4.7 Heatmap of the (a) Synthetic Data, and (b) Original Data 100
4.8 Comparison of C.I overlap indicate the utility of synthetic dataset. 101
5.1 Accuracies corresponding to the three datasets 113
5.2 RMSE corresponding to the three datasets 116
5.3 Performance of KNN on different values of parameter k 117
5.4 Boxplot indicate the accuracies of the proposed models 118
5.5 Boxplot indicate RMSE of the proposed models 119
6.1 The resampling algorithm of 10-fold cross validation 127
6.2 Performance of baseline models on data set ADS 128
6.3 Performance of baseline models on data set GDS2 130
6.4 Performance of baseline models on data set GDS3 131
6.5 ACC of proposed models on data set ADS 134
6.6 RMSE of proposed models on data set ADS 134
6.7 ACC of proposed models on data set GDS2 136
6.8 RMSE of proposed models on data set GDS2 137
6.9 ACC of proposed models on data set GDS3 139
6.10 RMSE of proposed models on data set GDS3 139
7.1 Performance of proposed models on data set ADS 154

7.2 Performance of proposed models on data set GDS1 158
7.3 Performance of proposed models on data set GDS2 159
7.4 Performance of proposed models on data set GDS3 159
7.5 Performance of proposed models on data set GDS4 160
7.6 Comparative performance based on accuracy of the proposed models 161
7.7 Comparative performance based on RMSE of the proposed models 161
8.1 Sample of feature importance by splitting rule in decision tree model 169
8.2 Feature Importance Rakning 170
8.3 Illustration of proposed framework 175
8.4 Performance of KNN on each FS sets with dataset ADS 179
8.5 Performance of Hybrid C5.0 on each FS sets with dataset ADS 179
8.6 Performance of Hybrid RF on each FS sets 180
8.7 Performance of IDBN on each FS sets with dataset ADS 181
8.8 Performance of KNN on each FS sets with dataset GDS4 183
8.9 Performance of Hybrid C5.0 on each FS sets with dataset GDS4 183
8.10 Performance of Hybrid RF on each FS sets with dataset GDS4 184
8.11 Performance of IDBN on each FS sets with dataset GDS4 185
8.12 ACC and RMSE comparison using ADS dataset 186
8.13 ACC and RMSE comparison using GDS4 dataset 187
9.1 The framework of the use of APPS for educational settings 191
9.2 Prediction modeling of APPS 192
9.3 The materials for educational stakeholder to use APPS 195
9.4 Shinny overview 196
9.5 A prototype of APPS interface 197
9.6 Operation flow chart in using APPS for academic performance prediction . . . 197
9.7 Prediction results identifying at-risk levels of poor-performing students 198
9.8 Summary of dominant set and ranking of highly influenced factors 198

9.9 Deployment of APPS 199
9.10 User feedback rating the characteristics of APPS 201

LIST OF ABBREVIATIONS

ACC	Classification Accuracy
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
APPS	Academic Performance Prediction System
CART	Classification and Regression Tree
CD	Contrastive Divergence
CHI	Chi-Square
DBN	Deep Beleif Network
DL	Deep Learning
DM	Data Mining
DNN	Deep Neural Network
DT	Decision Tree
EDM	Educational Data Mining
FA	Factor Analysis
FE	Feature Extraction
FS	Feature Selection
IDBN	Improved Deep Belief Network
IG	Information Gain
k-FCV	k-fold Cross Validation
KDD	Knowledge Discovering in Database
KNN	k-Nearest Neighbor
LR	Logistic Regression
LVQ	Learning Vector Quantization
MCD	Minimizing Contrastive Divergence

MI	Mutual Information
MICHI	FS method of combining MI and CHI algorithms
ML	Machine Learning
MLP	Multilayer Perceptronz
MoEYS	Ministry of Education, Youth, and Sport (Cambodia)
NB	Naive Bayes
PCA	Principal Component Analysis
PSEM	Predictive Structural Equation Modeling
RBM	Restricted Bolztmann Machine
RF	Random Forest
RMSE	Root Mean Square Error
SEM	Structural Equation Modeling
SMO	Sequential Minimial Optimization
Std.	Standard Deviation
STEM	Science, Engineering, Technology, and Mathematics
SVM	Support Vector Machine
SU	Symmetrical Uncertainty

CHAPTER 1

INTRODUCTION

1.1 Domain of the Study Problem and Background

Education is a significant mechanism in linking the economic and social development. Over the past decades, educational planners, policy makers and administrators all over the world have become increasingly concerned about the quality of education provided by the school systems. Ministry of Education Youth and Sport (MoEYS) in Cambodia has paid more attention to educational reform especially the school standard, curriculum, standard of school principals, teacher professional development, early childhood care education, lifelong learning and many more. In this case, to achieve the national plan of education, academic performance is the root factor that prepares students to do well later in life and be ready to compete in the global market. Cambodian schools, in both urban and rural areas have faced demanding problems on school management, community involvement, student's performance and school's achievement. Many schools have done well in their academic achievement while other schools have been struggling in doing so. With regard to the national examination reform 2014, more than 70 percent of nearly 90,000 of Cambodia's high school students who sat for that year national examination (K-12's national exam) failed to pass [1]. Under such circumstances, the poor performance causes the problem of under education and shortage of skilled manpower that has significant effect on the economic growth and the development of the country.

Cambodia is among the developing nations in Asia, with a relatively recent development of education system [2]. It was not until Cambodia made a surprising move from a socialist, centralized system to a capitalist, free-market economy model in the early 1990s that the emphasis on the role of education in producing a skilled labor force for socio-economic development has gained momentum ever since in the history of Cambodia's education. Over the years, student quality has, in particular, remained a chief national policy concern [3]. Low student achievement is a case in point. Therefore, as education has been viewed worldwide as an indispensable impetus to the nation's socio-economic development, an attention swift from the quantity to the quality within the current education development landscape is needed. resources to produce human resources with less quality.

From the experience the quality of education in Cambodia is still insufficient which needs a lot of improvement toward the whole educational system. After the Education Ministry implemented reforms on the education system and the national exam, the number of students who had successfully passed Cambodia's grade 12 exams was a slight increased [4]. In addition, the schools who performed well for their whole academic year, their students' national exam has also been high while some schools who did not perform well during their academic year, their national exam result was also down, which is caused by school factors . This is the main issue being faced by most of the schools in Cambodia. Numerous studies have focused on identifying factors that affect student performance.

To improve student quality in education, the Royal Government of Cambodia has placed a strong commitment and has launched a number of approaches to increase student achievement in national policy [3][4]. One of the immediate approaches is to enhance the quality of teaching and learning, and observe the learning behaviors that affect the student performance. Nonetheless, despite this policy promulgation, there is a substantial gap between policy and practice. In recent years, while MoEYS has placed improving teaching and learning quality at the core of the national policy as a

counter-measure to raise student achievement, education teaching and learning quality assessment has yet to exist institutions and nation-wide [3]. Throughout the years, the government's commitment has been predominantly placed on the establishment of the national accreditation body, i.e. the Accreditation Commitment of Cambodia in order to put the institutional accreditation in place. However, there are many others factors such as family environment and students' attitude toward learning behaviors that simultaneously affect students' achievement.

Cambodia's STEM (Science, Technology, Engineering and Mathematics) discipline has continuously improved through the years. However, it still has a long way to go in terms of mastery and application in the field. Educations and careers in STEM fields are playing an important role to help in developing and promoting economic growth in both developing countries and developed countries to move forward especially in the competitive ASEAN integration in 2015. Minister of MoEYS set out STEM discipline and digital learning for the year 2020-2030 for the development of Cambodia [5].

The system of education in Cambodia is generally divided into the grades or levels as in Table 1.1.

Table 1.1: System of Education in Cambodia

Education	School/Levels	Grades	Years
Primary	Primary	1-6	6
Middle	Junior High School	7-9	3
Secondary	High School	10-12	3
Vacation	Vocational/Technical School	T1-2/3	2-3
Tertiary	Bachelor	B1-4/5	4-5
Tertiary	Master	M1-2	2
Tertiary	Doctorate	D1-3	3

High school is the bridge providing background knowledge for their future to university. In that, higher performance in high school is critical point and backbone of their academic

life. High schools provide subject Mathematics, Physics, Chemistry, Biology, Khmer Literature, Geography, History, Earth Science, and foreign languages (English/French). The poor performance of students in high schools has become a obstacle for students in their high school performance as well as further study as it affects the secondary national exam and step to higher education. Mathematics is considered as the basic background for many science subjects, and give very strongly affect the national exam and for further study in higher education. For example, students who are poor in mathematics are much more likely to fail in diploma national exams in Cambodia [6]. Later, they found themselves harder to choose their wanted majors for higher study and hard to survive in the university journey. Effect of improving student performances in high school:

(i) Reducing failure rate and repeating grade rate of poor-performing students in high schools.

(ii) Increasing passing rate in national exam (K-12), diploma exam of final year of secondary school education.

(iii) STEM discipline and digital learnings in MoEYS strategies for development of the country.

1.2 Mining Academic Performance

A study conducted to understand this complex phenomenon concluded that there are several causes that could possibly contribute dropout, failing, and repeated class. These causes, however, first contribute to the student's poor academic performance which in turn provokes the student to poor performance [7]. Therefore, there is need to understand which factors cause students to under-perform in their academic work and end up failing exam, dropping out, and repeated class. Several factors have been identified and presented in a conceptual model as illustrated in Figure 1.1.

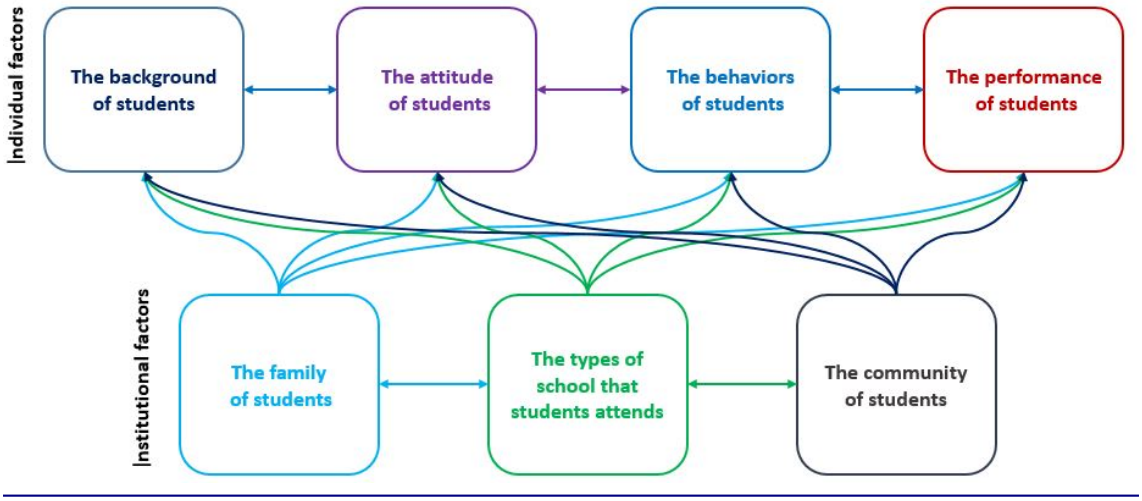


Figure 1.1: Conceptual model of student performance (adapted from Rumberger and Lim [7])

As illustrated in Figure 1.1, the factors that cause students to poor performance have been categorized as: individual students' characteristics; and institutional characteristics. Findings suggest that: no factor in itself causes students to drop out; the students' academic performance together with their behavior directly affects their decision to dropout; the process of dropping out of school is pointed to by an accumulation of poor academic performance together with bad behaviors [7]. These findings suggest that students who failed exam, drop out, and repeated classes of school because they have performed poorly in their academic work.

1.2.1 Poor Performance: Failure, Dropout, and Repeating Classes

Analyses in [7] show that poor performance of adolescent is not the result of any single risk factor, but rather of a combination and accumulation of various barriers and disadvantages that affect students throughout their lives. Low performers tend to have less perseverance, motivation and self-confidence in mathematics than better-performing students, and they skip classes or days of school more. Students who have skipped school at least once in the two weeks prior to the PISA test are almost three times more likely to be low performers

in mathematics than students who did not skip school [8]. This kind of at-risk students are highly possible to fail, drop out, or repeat classes due to their poor performance.

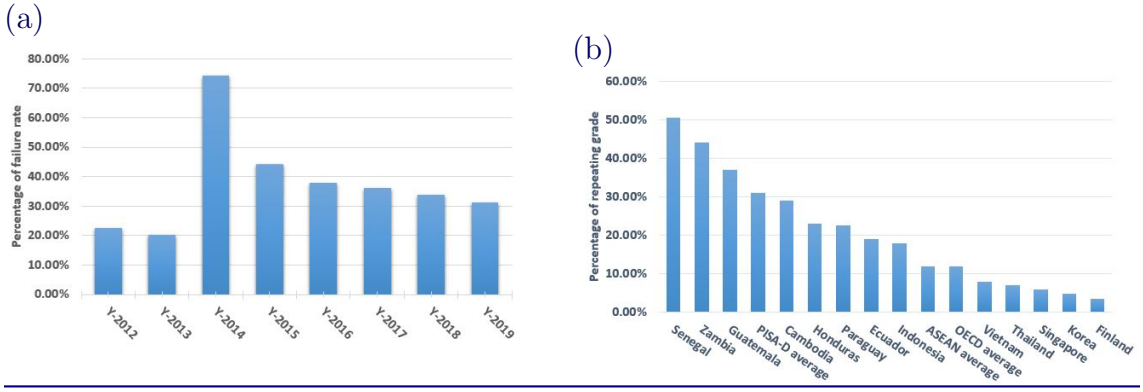


Figure 1.2: Poor performance: (a) Failure rate of Cambodia student in national exam (MoEYS), (b) Percentage of students who had repeated a grade in primary, lower secondary or upper secondary school (adapted from PISA 2015 for development databases [4])

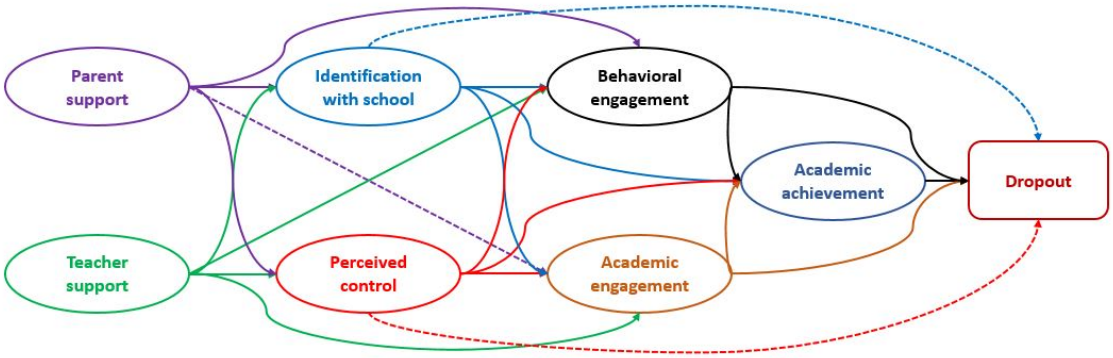


Figure 1.3: Self-System model of motivational development linking academic performance to dropping out, the solid lines show the direct contribution while the curved dotted lines show indirect contribution (adapted from Rumberger and Lim [7])

Grade repetition can be a costly policy, as it generally requires greater expenditure

on education and delays students' entry into the labor market [9]. In theory, repeating a grade gives students time to catch up with their peers if teachers believe they are not yet ready for more advanced coursework. If the curriculum is cumulative and further learning depends on a solid understanding of what has been previously learned, then promoting students regardless of their mastery of the content might place low-performing students in an increasingly difficult position at higher grades. Figure 1.2 (b) demonstrate the comparison of percentage of Cambodia and other countries of students who repeated class primary, lower secondary, and upper secondary schools.

Academic performance and dropping out are seen as related outputs of a combination of factors as illustrated in the self-system model of motivational development in Figure 1.3. As seen in the figure, the model links institutional factors including family support, and teacher support, to individual factors such as: perceived identification with school; and perceived control. Further, these factors are seen to influence the engagement behavior factors that in turn directly contribute to the two educational outcomes: academic achievement and dropping out. The two educational outcomes, the model suggests, are equally affected by the engagement factors as indicated by the solid lines. Additionally, academic performance directly contributes to dropping out [9]. Based on the model discussed, it can be concluded that the factors that cause dropping out also cause poor academic achievement to a large extent. This study adopted the view that poor academic performance is a cause of students dropping out of school. Specifically, in the developing world context, the study focuses on students in primary schools who drop out of school because they did not pass the standardized national exit examination. In this context, the students do not willingly dropout but are thrown out of the school system.

1.2.2 Causes of Poor Academic Performance in Developing Countries

Several reasons explain poor academic performance of students, as stated earlier. In Cambodia, most cognitive and non-cognitive factors affect student performance. These

reasons are said to include school based factors, home based factors, Government policies and an overlap of the three [10]. Other findings suggest that the causes of poor academic performance in schools are directly related to: school factors (lack of teaching materials, textbooks and trained teachers); teacher factors (lateness, absenteeism, using local language, and poor syllabus coverage); pupil characteristics (truancy, lack of help on studies at home, low interest in lessons); and parent characteristics (failure to provide breakfast, text books, basic school needs, and low interaction with teachers) [11].

Among the school factors, teacher effectiveness is the most important indicator in student achievement. Literature indicates that an effective teacher has qualities including knowledge and ability to organize the subject matter, skills of instruction, a positive attitude, clear communication with students, is respectful and fair, has concern for student learning, gives fair assignments and assessments and gives timely feedback. Additionally, the teachers' effectiveness is closely related to factors such as: adequate and relevant teaching and learning materials, and a good environment within the school.

Parents' socio-economic status has also been associated with poor academic achievement. For example, concepts are seen in Korea such as the *tiger mom* phenomenon or *academic zeal* of parents; that means parents have increased involvement in children's academic performance [12]. The quality and extent of parental involvement and investment in children's academic performance could be limited by the socio-economic status of the parents. Parents have to pay for extra classes to tutor the children. Children with parents who do not have money to pay for extra classes are therefore at risk of remaining behind in the academic progress [13].

Therefore, parental support and performance expectations could be among the most important factors that influence student behavior and academic performance [14]. Findings of a study on the relationship between parent expectations and end-of-grade performance in Maths and reading supported the concept that, higher expectations predict higher performance [15]. Further, communication and collaboration between

parents and teachers could improve parental expectations for their children and the expectations of the school settings [16].

So far in this section, we have looked at the many causes of school dropout or poor academic achievement in the literature. A large list of these causes have been extracted from the literature reviewed. These are categorized as domestic or home causes, individuals or student causes, and institutions or school causes. In most developing countries, the causes poor academic performance could effectively be grouped into sub-categories: school management and administration factors, teacher factors, student factors, parent factors, and environmental factors. Clearly, the list of the individual causes is large, and there is need to select only those causes that are most predictive of academic failure in a given environment. Initially, we conducted a survey in the study area to understand the causes of poor academic performance on the ground.

1.2.3 Successes in Academic Performance and Educational Settings

The current minister of MoEYS, H.E. Hang Chuon Naron (Dr.), and British Ambassador in Phnom Penh are trying their best to promote awareness of STEM education through STEM event. Surprisingly that STEM learning achievement research on Cambodia secondary school found that science and Maths were the lease favorite subject in STEM education. The strategies, planning, and scheduling of role of STEM and digital innovation's impact in Cambodia need pre-knowledge of students ability in Mathematics and science.

In the digital era and innovative decades, the MoEYS and Ministers of Cambodia set out the policy strategies of Science, Technology and Innovation (STI) for Year 2020-2030. The goal is to develop STEM discipline and digital learning to as skills to develop the country, Cambodia. That is the success of students in high schools role importantly as a strong backbone for these skills. Science, specially mathematics knowledge is concerned. The success of student in mathematics become an important thing to implement. Hence,

the study of predicting achievement of students in mathematics and mining the influence factors is a point in this dissertation.

1.3 Motivation of the Study

Cambodia is an developing country where STEM discipline, proposed by MoEYS is believed to be important for building and maintaining the development of the country. STEM education fosters a skill set that stresses critical thinking and problem-solving abilities. This type of skill set encourages innovation among those who possess it. Similarly, a country's economic development and stability are dependent on its ability to invent and develop new products. Technological innovation in the modern age is only obtainable through the expertise of specialists with knowledge of recent STEM research. Therefore, the role of STEM in developing country is important because a country's economy is completely dependent on new developments from technology and science.

High school background role importantly as background for their further study and also a bridge to their higher education. OECD (Organisation for Economic Co-operation and Development) studied of low-performing students [8]. The poor performance students who have not attained a baseline level of proficiency in at least one of the three core subjects: reading, mathematics, and science. The poor performance of students in high schools has become a worried-task for educators as it affects the secondary national exam and step to higher educations. Student performance in secondary school is considered as the basic background that give very strongly affect the national exam and for further study in higher education [17]. For example, poor-performing students are highly likely to fail or barely pass the national exam. Later, they found themselves harder to choose a major for higher study and hard to survive in the university journey. Early prediction of student performance has a high impact on in-time intervention, managerial setting, scheduling, and planning in education institutions. Therefore, early prediction has been

considered to be a powerful method for early identification of students who are in at-risk and need intervention and assistance.

An early survey conducted at the beginning of this study provided substantial motivation for it to be carried out. The focus in the this study is to establish the early prediction of high school student performance, analyze the learning behaviors and highly influenced factors for the views of education as benefits for stakeholders in Cambodia to improve academic performance, maximize the high performance of students and minimize the failure rate of poor-performing students. To accomplish this goal, the research apply advance techniques to develop of prediction models and the design and development of the academic performance prediction system.

1.4 Statement of the Problem

Analysis of academic performance is a challenging task since it relies on diverse hidden factors like domestic factors, individual factors, and schooling factors. We tried to observe all relevant features since the many most effective factors are not always found in school databases [18][19][20]. That is why all our attributes are from the various main factors and try to observe the contribution of these attributes to the success of student performance. In this research work, we aim not only to predict the students' performance but also to analyze the related features which influenced their achievement in mathematics.

In a dataset $\mathbf{D} \in \mathbb{R}^{n \times d}$, suppose we have $s_1, s_2, ..., s_n \in \mathbb{R}^d$ as a row vector. Think of s_i as a student such that $s_i = (F_{i1}, F_{i2}, .., F_{id})$ with each F_{ij} is the value of a feature (or attribute) or information describing that student. The objective of the study is to predict the student performance via the effect of vectors of feature set $(F_1, F_2, ..., F_n)$ on student performance $P(S)$ denoted as the mapping

$$S = (F_1, F_2, ..., F_n) \rightarrow P(S) \tag{1.1}$$

The academic institutions aim to enhance the performance of students by monitoring their learning behaviors and influencing factors. The process consumes time and efforts to gather such information for making decision especially for many students. Many works have revealed different approaches that help in providing a decision for improving student performance. This approach was called educational data mining (EDM) ranking from statistical analysis techniques, to machine learning techniques, and to deep learning frameworks. We will use this approach to in our analysis.

1.5 Research Objectives

The primary goal of educational systems is to enrich the quality of education by maximizing the best results and minimizing the failure rate of poor-performing students. This research aims at developing effective EDM techniques and proposing a web-based application for educational stakeholders to get prediction of student performance in the early state, so that the right intervention can be performed. Being able to predict student performance in the early stage enables a teacher or educational institution to adjust their teaching methods and provide appropriate assistance to poor-performing students. To be clear with the research objectives, the research work in this dissertation is driven as the following three main research questions follow by its primary objectives:

Question 1: Do the proposed classifiers or prediction models can generate the most successful results ?

- Comparative Study (Spot-Checking): Searching for the effective prediction models of EDM that popularly used in predicting academic performance and mining students' outcomes.
- Model Development: Developing the optimal models from the experimental results of comparative study to obtain the higher performance of prediction.

Question 2: How can we accurately detect student learning patterns or factors that highly affect (required and sufficient) their academic performance ?

- Feature Selection (Dominant Set): Determining the optimal feature set containing highly influencing factors called dominant factors that is required, sufficient, and effective on determining and control the academic success of students.

Question 3: How educational stakeholder and related individuals can utilize the proposed work to predict student learning outcomes for giving intervention and improvement of student performance ?

- Prediction System: Developing academic performance prediction system to predict students' outcomes and figuring out the poor performance students at the early stage for giving intervention.

1.6 Dissertation Structure

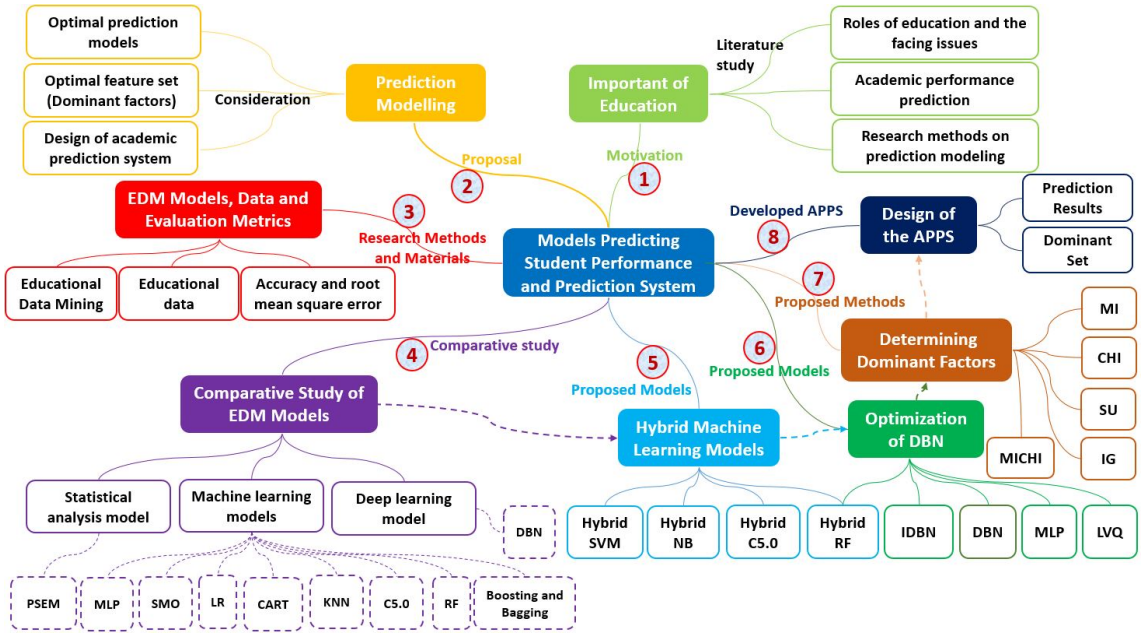


Figure 1.4: Research summary (research roadmap) of this dissertation

The remaining parts of this dissertation are organized as follows:

Chapter 2: Literature Reviews and State-of-the-art Methods

In this chapter, we discuss the previous research work that entails the consideration and analysis on factors affecting student performance, prediction models using for predicting students performance, and feature selection methods in ranking informative features and learning behaviors, and role of academic prediction system. The summary knowledge from literature review concerning with the domain understanding and state-of-the-art methods are stated.

Chapter 3: Educational Data Ming Techniques

This chapter introduces various methods of EDM that used for prediction modeling in this dissertation. The methods is ranked from statistical analysis methods, to machine learning techniques, and to deep learning framework that we will used for comparative study in this study. In this chapter, we also present primarily knowledge of effective feature selection methods to investigate and extract highly influencing factors and important learning behavior for academic performance prediction and recipe for recommendation/intervention in student performance. Feature extraction using principal component analysis will be used to improved the performance of machine learning models to be up to the high performance models.

Chapter 4: Methodology for Academic Performance Prediciton Modeling

In this chapter, we introduce steps in Cross-Industry Standard Process of Data Mining (CRISP-DM) for academic performance prediction modeling of in this dissertation. This chapter summarizes the six steps of CRISP-DM methodology: domain understanding, data understanding, data preparation, data modeling, model evaluation, and model deployment.

Chapter 5: Comparative Study of Educational Data Mining for Student Performance Prediction

In this chapter, we give a comparative study of prediction models that are popularly used in the research of predicting academic performance. Three classes of prediction models of EDM: statistical methods, machine learning, and deep learning, are used to analyze using three synthetic datasets and an actual dataset. The chapter aims at searching for the optimal models for further model development.

Chapter 6: The Proposed Hybrid Machine Learning Models for Student Performance Prediction

The optimal models from the comparative study are further developed to obtain the most successful classification result. In this chapter, we proposed a study of hybrid machine learning models as the optimization approach of model development. The study proposed a combination of baseline models with Principal Component Analysis (PCA), and a resampling method k-fold cross validation (k-FCV) to boost up our proposed methods.

Chapter 7: The Proposed Optimization of Deep Belief Network for Student Performance Prediction in the Context of Larger Datasets

This chapter proposes an improved Deep Belief Network (IDBN) for predicting student performance in the context of larger dataset. The proposed optimization approach comprise of enrich the dataset with features selection methods, finding optimal value of hyper parameters, and applying regularization method for weight decay. The purpose of this chapter is to give an optimal deep learning model, which can be alternatively used for predicting student performance context of larger dataset.

Chapter 8: The Proposed Feature Selection Methods for Determining Dominant Factors for Academic Performance Prediction

This chapter presents a study of feature selections and proposed a novel feature selection for selecting the dominant set that control the success or failure of academic performance. We proposed a novel feature selection method and compare with four existing popular feature selection methods. The chapter's objective is to introduce novel feature selection approach to select the dominant set for improving the performance of prediction and insight knowledge for intervention.

Chapter 9: Design Web-base Academic Performance Prediction System for Student Performance Prediction

In this chapter, we design an Academic Performance Prediction System (APPS) as web-based application for predicting student performance. The prediction models and feature selection methods are combined and we subsequently introduce the developed APPS for predicting student performance and identify the at-risk group of students. The system is designed for educational stakeholders to effectively use for predict student performance in the early stage, so that the right intervention can be implement for improvement.

Chapter 10: Summary and Conclusion

In this last chapter, we summarize the research works including the proposed EDM models and APPS and emphasize the experimental results of the research conducted in this dissertation. The chapter concludes the methodological contribution and applicability contribution of the conducted research.

CHAPTER 2

LITERATURE REVIEW AND STATE-OF-THE-ART METHODS

2.1 Overview

Numerous works on education research have investigated with various approaches through the decades. The classical research in education environment was observed using statistical approach to observe and discuss the relation of influencing features to the academic performance. The works were popularly done by considering the correlation of predictive feature to output variables or the inter-relation of among features and to output variables using traditional statistical approaches. The new trends of data mining called educational data mining was then applied to predict academic performance such as student performance, drop out rate or retention, and access the learning behaviors of students. Some studies have focused on mining the influencing features and learning behaviors that affect the students' success.

The present chapter gives the literature views of those works (related works to correlation and causation among features, related work to predict student performance,related works to ranking informative features and learning behaviors, and developing early warning/predicting systems). Additionally, we observed on extensive literature in order to investigate the performance of various models, and its task (relationship mining, prediction, clustering), and state the objectives of our study.

Purposively, from the literature reviews, we searched for the effective approaches for student prediction.

2.2 Applications of Educational Data Mining

Educational data mining (EDM) is a research field concerned with the application of data mining, machine learning, and statistics applied for exploring data in educational contexts [21]. Various managerial settings, planning, and scheduling in educational settings required effective techniques of EDM to uncover the knowledge and information of student learning patterns to give intervention and set up a policy to improve academic performance [22]. Romero and Ventura published the first literature survey of EDM that concerning with using data mining techniques in education settings in 2007 [23]. The literature was later improved in 2010 [24].

Various analysis techniques have been introduced for monitoring and anticipating academic performance to keep track of teaching, learning actions, and productive results. The main goal of EDM is to extract information from educational data to address important educational questions and support decision making [26]. Several studies on academic performance have been carried out using methods from the EDM discipline. Numerous tools have been applied according to the objectives of the studies. The distinction of characteristics of data, the complexity of data, the level of contribution signification, and limited performance of existing methods require advanced techniques of EDM [22][23][24][25][26][27]. EDM combines several interdisciplinary fields of study such as machine learning, statistics, data mining, information retrieval, psycho-pedagogy, cognitive psychology, recommender system methods and techniques to various educational datasets to resolve educational issues [24]. The cycle and applications of EDM are shown in Figure 2.1.

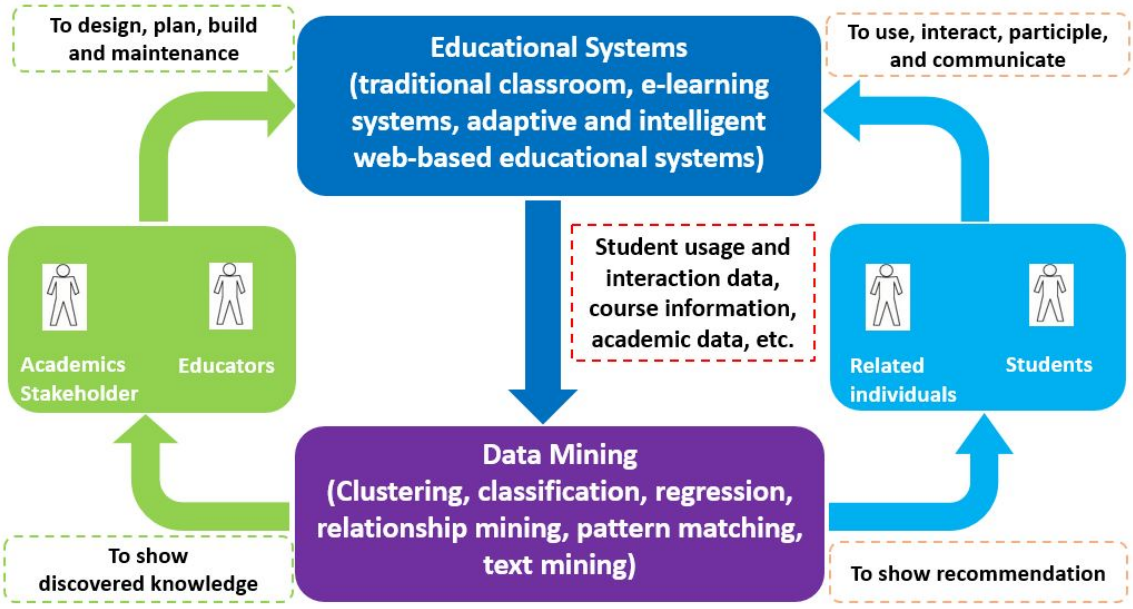


Figure 2.1: The cycle of applying data mining in educational systems (Educational data mining: EDM; adapted from [23])

Examples of these systems include: traditional classrooms, e-learning, and intelligent web-based educational systems. Students use, interact and participate with these systems. As they interact with the systems, data about their usage, interaction, and test marks are gathered. Different data mining techniques, such as clustering and prediction are then applied on the data in order to extract useful patterns that could be used to help the students achieve better performance. The discovered knowledge is beneficial to all the education stakeholders, especially teachers and students.

2.3 Literature Reviews of Related Works

2.3.1 Features Affecting Academic Performance

Mohamed et al. [28] proposed Factor Analysis (FA), and Structural Equation Modeling (SEM) to determine the structure and important factors influencing achievement in

mathematics. The data were obtained by giving questionnaires to lower secondary school students in Libyan Kuala Lumpur, Malaysia. The dominant factors observed in this study are teaching practices, teacher's attributes, classroom climate, students' attitude toward mathematics and students' anxiety. The study showed that teachers' characteristics and student's attitude toward mathematics are the most important variables for student's achievement in mathematics.

Uysal [29] presented a study of factors affecting the mathematics achievements of Turkish students in Program for International Student Assessment (PISA) 2012 using SEM. The research took place in Turkish high schools, where the PISA program was implemented. Mathematical structure equation model is analyzed, it has been seen that there is a medium and positive relationship between mathematics achievement and mathematics interest of student with mathematics self-concept. It has been determined that the variables mathematics anxiety has a negative and medium affect on mathematics achievement. It was determined that anxiety has a significantly negative effect on a student's performance. According to the results of research study there is no meaningful relationship between variables of teacher-student relation, classroom management, sense of belonging to mathematics achievement.

Nazim and Ahmad [30] proposed the application of confirmatory factor analysis (CFA) to assess the multidimensionality, reliability, validate and fitness of influential factors of 8th gradees students' mathematic achievement in Malaysia. CFA was used to obtain the best fit model in structural equation model. For the purpose of the study, secondary data of Trends in International Mathematics and Science Study (TIMSS) was used. The questionnaire were distributed to 5733 8th grade students in Malaysia. At the end of the study, all constructs has achieved multidimensionality, reliability, validity and fitness required. All independent latent constructs have significant effect towards both attitude and achievement variables thus have partial mediation role in linking to achievement except for school variable which is not significantly link to attitude

Tambunan [31] identified the most dominant factor of the teacher's role as a motivator that influences students' interest and motivation to perform in mathematics achievement. It is conducted in 8th grade of senior high school with 209 students, consisted of five state schools and two private schools from seven regencies in North Sumatera. The data collecting technique uses questionnaire about students' interest and motivation toward mathematics and teacher's role as motivator. Numerical data on mathematics achievement of students is obtained from school documents. The result of data with path analysis is obtained by dominant factor of teacher's role as motivator that is factor of delivery of learning goal and learning comfort equal to 6.10%, and 6.00% is influenced by the delivery of learning objectives and variations of learning approaches, 5.17% is influenced by the delivery of learning objectives, 5.06% is due to variations in the learning approach, 4.61% is influenced by learning comfort and variation of learning approach, and 4.26% influenced by pleasant class atmosphere.

Serhat and Ferit [32] proposed a study of analyzing factors affecting the success of students in science and mathematics. The main purpose of this study was to propose a model for how elementary school students' science and mathematics achievements in their schools and in Level Determination Exam (SBS) depend on the number of teachers and expert teachers in their schools. The sample of the study was 5672 elementary students for the purpose of the study, the number of teachers and expert teachers who worked in sample schools has been defined as independent variables, and students' science and mathematics achievements in their schools and in SBS exam have been defined as dependent variables. The data obtained from school administrations were analyzed using SEM to analyze relations among students' science and mathematics grades in their schools and science and mathematics achievements in SBS exam and the number of teachers and expert teachers in their school. As a result of the analysis, it has been observed that established model has acceptable fit indices and an increasing number of teachers and expert teachers have positive effects on students' science and mathematics achievements.

Kerlic and Askin [33] assessed the impact of students' background and parents' attitude towards their children on mathematics achievement across four countries which were participated to Trends in International Mathematics and Science Study (TIMSS) 2011. Since Republic of Korea, Singapore and Chinese Taipei were ranked as the first three countries in terms of the achievement in mathematics scores, they were selected for the comparison with Turkey. Multilevel logistic regression model was used to estimate coefficients and to model differences in mathematics achievement within and between schools for each country. Gender, student-parent relationship, possessions of computer, room and internet were taken as student level variables, while school composition by student economical background, discipline and safety of school climate were taken as school level variables. According to results, owning a desk was the most effective factor on achievement of students for both Turkey and Republic of Korea, at the student level. Furthermore, at school level, the most effective factor on achievement was found the school composition by student economical background for all countries.

2.3.2 Student Performance Prediction

Anuradha et al. [34] gave an analysis of classifying the performance of the student during the final year of an undergraduate degree. The analysis was made on students at a private institute in Tamil Nadu, India. The prediction was done using a dataset that was conducted using the information on the demographic and secondary school records. The models proposed in this study are C4.5, Naïve Bayesian (NB), k-Nearest Neighbors (KNN), and two rule's learner algorithms namely OneR, and JRip. The result of this study reveals that overall accuracy of the tested classifiers is ranged from 61% to 75%. In addition classification accuracy for the different classes reveals that the predictions are worst for distinction class and fairly good for the first class. The JRip produced highest classification accuracy for the distinction. Classification of the students based on the attributes revealed that prediction rates are not uniform among the classification

algorithms. Also showed that selected data attributes have found to be influenced the classification process.

Thai Nghe et al [35] has compared the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two academic institutes: Can Tho University and the Asian Institute of Technology, an international postgraduate institute in Thailand that draws students from 86 different countries. The total number of student records used for this prediction is 20492 and 936, respectively. Algorithms used for this prediction are C4.5, M5P Model Tree, and the Bayesian Network (BN). In this analysis, they concluded that for all classes of predictions C4.5 of decision tree yields better results than BN by 3 to 12%. The results of the experiment gave insight into techniques for accurately predicting student performance, compare the accuracy of data mining algorithms, and demonstrate the maturity of open source tools.

Kotsiantis et al [36] described a model to predict student results for a distance learning course in Hellenic Open University. Students' key demographic characteristics and their marks in a few written assignments can constitute the training set for a supervised machine learning algorithm. The learning algorithm could then be able to predict the performance of new students thus becoming a useful tool for identifying predicted poor performers. The scope of this work is to compare some of the state of the art learning algorithms. Two experiments have been conducted with six algorithms. The algorithms used for this prediction are C4.5, NB, Back Propagation (BP), KNN ($k = 3$), Sequential Minimal Optimization (SMO), and Logistic Regression (LR). A set of 510 students of the university was chosen for experimental purpose. It was found that the NB algorithm generates the best results with the accuracy of 72.48%.

Menaei-Bidgoli et al. [37] implemented data mining algorithms on "logged data" in educational web-based learning of students enrolled in a physics course in Michigan State University. The classification algorithms proposed in this study are decision tree

classifiers (C5.0, CART, QUEST, CRUISE) and non-tree classifier such as Quadratic Bayesian Classifier, KNN ($k = 1$), Parzen-Window, Multilayer Perceptrons (MLP). The system was tested with a data set of 227 students enrolled in a physics course in Michigan State University. The authors designed, implemented, and evaluated a series of pattern classifiers and compare their performance on an online course dataset. A combination of multiple classifiers leads to a significant improvement in classification performance. Furthermore, by learning an appropriate weighting of the features used via a Genetic Algorithm (GA), we further improve prediction accuracy. GA was demonstrated to successfully improve the accuracy of combined classifiers, about 10 to 12% when comparing to non-GA classifier.

Kovacic [38] explored the "socio-demographic" and "study environment" factors that results in student dropout in a polytechnic college in New Zealand. He used student enrollment data like age, gender, ethnicity for this purpose. The total number of student records used for purpose was 450. Algorithms used for this prediction are CHAID and Classification and Regression Trees (CART). It was found that CART obtained a higher degree of accuracy. Based on the results of Confusion Matrix and ROC curve he concluded that decision trees based on enrollment data alone are not sufficient to classify students accurately.

Mythili et al. [39] gave an analysis of student performance by applying data mining classification algorithms. The studied searched for effective data mining techniques for leveraging educational data. The various classification algorithms used in the study are J48, Random Forest (RF), MLP, Artificial Neural Network (ANN), IB1 of Nearest Neighbors classifier, and Decision Table. The results of such classification models were considered with accuracy level, confusion matrices and also the execution time. From the experimental results, conclusion could be reached that RF performance is better than that of different algorithms.

Daud et al. [40] utilized the techniques of EDM and learning analytics (LA) to

predict academic performance. For experimental purpose, the data of graduate and undergraduate students has been collected from different universities of Pakistan during the period (2004 to 2011). Initially, about 3000 student records were collected. After removing inconsistencies and duplication in the dataset, the study considered 776 students instances for experiments. Five classifiers of EDM were used in this prediction: Support Vector Machine (SVM), C4.5, Classification and Regression Tree (CART), BN, and NB. SVM classifier was found to be effective for the proposed problem and outperformed the rest models.

Sultana et al. [41] evaluated deep learning and data mining methods to predict student performance. The study aimed at surveying the performance of students at Saudi University. The prediction of student performance using EDM was carried out in this research work. Classification was done in order to predict students in different class categories like High, medium and low. The data was chosen for implementation was collected from the learning management system, the university database. The dataset consisted of 1100 records. The research employed deep learning techniques, such as Deep Neural Networks (DNN) and data mining techniques such as Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and NB. DNN and DT predicted the students' performance very well compared to other proposed techniques.

Sinthia and Balamurgan [42] proposed a study of analyzing student performance using data mining technique. MLP, k-means, and KNN were used as prediction models in this study. Participants in this study were the students of faculty of engineering at CHRIST (Deemed to be University) for the one department core subject. This course lasted for 6 months. Data was taken from 120 students who took the course in their second year of engineering. The students were observed based on the internal assignments and quizzes and the model exam given by the respective lecturers. MLP was good for analyzing the student's performance in considered environment and dataset, compared with KNN and k-means.

Zhang et al. [43] introduced Deep Belief Network (DBN) for grading student performance. The work tackled the task of Automatic Short Answer Grading (ASAG). While conventional ASAG research makes prediction mainly based on student answers referred as Answer-based, the study leveraged the information about questions and student models into consideration. The analysis considered the exploration of five machine learning classifier together with DBN. The five machine learning methods included NB, LR, DT, ANN, and SVM. The data utilized in the experiment were obtain from the University of the Cordilleras, Philippine. Experimental results showed that the proposed DBN produced the highest accuracy, which is the best model and followed by SVM, while NB is the worst.

2.3.3 Feature Selection Methods

Estrera et al. [44] predicted student performance for academic ranking in a university in the Philippine using the information records from high schools. Three prediction models of data mining were used in this prediction. The models used in the analysis are DT, NB, and KNN. The data used in the prediction was obtained from the information provided by the admitted students to a university using a survey questionnaire conducted on them. To get a better prediction and better understanding of learning behaviors for assessment of students' success, the authors proposed feature selection methods: Chi-square (CHI), Information Gain (IG), and Gain Ratio (GR) in this study. As a result, DT algorithm generated the highest accuracy of 90.67%.

Ramaswami and Bhaskaran [45] proposed a comparative study on feature selection methods to identify the factor affecting student performance and improving the predictive model of data mining. Six filter-based feature selection methods (Correlation-based (CB), Relief (REF), and Symmetrical Uncertainty (SU), CHI, IG, and GR) were used to improve the performance of base-line NB classifier. Optimal feature subsets from each feature selection method were subsequently carried out by using different classifiers: Voted

Perceptron, NB, OneR, and PART. The experimental outcomes showed the improvement of classification performance, reduction in computational time, and constructional cost in prediction phases.

Amjad et al. [46] used information systems record as features that contain their high school records, and the university records to predict the student performance in higher education. The dataset was extracted from a private university in the United Arab of Emirates. The dataset included 34 attributes and 56,000 records related to students' information. He introduced DT, RF, ANN, NB, LR, and Generalized Linear Models (GLM) as prediction models. The experimental results indicated that RF was the most appropriate prediction model in the prediction problem. IG was used to select the important feature in this study. It is revealed that the most important attributes that have a direct effect on the students' academic performance are belonged to four main categories, namely students' demographics, student previous performance information, course and instructor information, and student general information. The evidence from this study would assist the higher educational institutions by allowing the instructors and students to identify the weaknesses and factors affecting the students' performance, and act as an early warning system for predicting the students' failures and low academic performance.

Dimic et al. [47] studied behavior patterns of students in the blended learning environment. Dataset used in the study was created by integrating data from multiple sources into a form applicable for data mining technique application. Dataset of 225 instances was obtained. The experiment has focused on data preprocessing steps in data mining. Feature selection methods such as IG, SU, REF, CB, wrapper method, classifier subset evaluator methods were used to extract the most important features. The dependencies of features were computed using Mutual Information (MI). The prediction models: NB, Aggregating One-Dependence Estimators (AODE), DT, and SVM were used as prediction models using different feature subsets from each feature selection method.

The results indicated that the REF, wrapper method, and MI acted as the most successful features selection methods in selecting the optimal feature sets. The presented research concluded that selecting the subsets of lower cardinality of students' learning activities gives a significant improvement in predictive accuracy in a blended learning environment.

Zaffar and Savita [48] investigated the analysis of feature selection methods in improving the performance of prediction models for predicting student academic performance. The study utilized six feature selection methods: CB, CHI, Filtered, IG, Principal Component (PC), and REF. Fifteen classifiers were used: NB, Naïve Bayes Updateable (NBU), MLP, Simple Logistic (SL), SMO, DT, OneR, PART, JRip, Decision Stump (DS), J48, RF, Random Tree (RT), and REP Tree (RepT). The experiments results indicated that there is a significant improvement of 10 to 20% accuracy when using different feature selection sets.

2.3.4 Academic Prediction Systems

Early prediction or warning systems for predicting student performance is regarded as the improvement or next step in academic performance. It is referred to as prediction methods capable of discovering important and useful information about student learning patterns and risks of students such as retention, drop-out, and students' outcomes in an early stage. The purpose of using an educational early warning system is to give an earlier prediction of academic performance using features that influence students' success. Performance exhibited by students in their learning could be predicted in advance and possible failure can be prevented by the timely intervention [49].

Hu et al. [70] investigated students' interaction data in an online undergraduate course by using EDM techniques to develop an academic prediction system that could predict the students' learning outcomes exhibited by students in the course recorded in a learning management system (LMS). Various prediction models were used to predict the performance of the pass/fail of students. The optimal classification algorithm in the

prediction system is CART supplemented by AdaBoost. The experiment produced a classification accuracy of 90%. The study concluded that the early warning prediction system successfully predicted students' learning performance in an online course.

Akcapina et al. [51] proposed learning analytics to develop an early warning system for predicting at-risk students registered in an online course in a university. The study was carried out using a dataset of 76 second-year students registered in the Computer Hardware Course. The prediction model used in the system is a KNN. The experiment was examined regarding data obtained in Week 3, 6, 9, 12, and 14 to predict if at the end of the term student will be unsuccessful or successful. In the data of the first 3 weeks, the prediction rate of predicting unsuccessful student is 74%, while in the week 14th, the prediction rate increased to an accuracy of 89%.

Lee and Chung [52] developed a dropout early warning system based on machine learning to improve the performance of dropout prediction. The study dealt with the problem with the class imbalance between non-dropout and dropout groups of students. The two baseline prediction models used in this early warning system are RF and boosted DT. The RF and boosted DT were combined with the synthetic minority oversampling technique. The data used in this study is 165715 records of high school students taken from the National Education Information System (NEIS) of South Korea. The combination of a boosted DT produced the best results and improved the performance of the dropout early prediction system.

2.4 Summary Knowledge and State-of-the-art Methods

Different studies that discussed, highlighted, and investigated the use of EDM have been reviewed. Different research studies that employed EDM in analyzing educational data have been reviewed. EDM was defined through its key concepts, the researches, its purpose, used algorithms, and key findings, with an emphasis on predicting student

performance and is a focus point of this thesis. The state-of-the-art methods is stated in the Figure 2.2. Further reviews of studies that applied EDM to the analysis of the

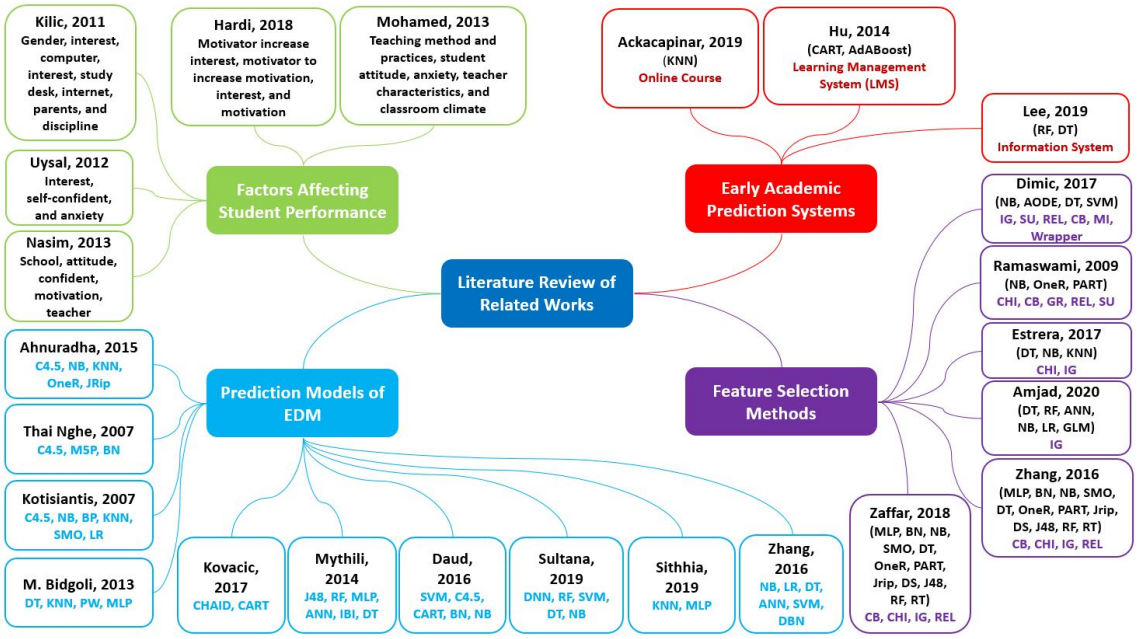


Figure 2.2: Summary of literature review of related works and its state-of-the-art methods

relationship between different factors of students' academic performance, have been taken into consideration throughout the chapter. We summarized the insight knowledge from the study of this literature reviews and listed as follows:

(i) The literature of previous research that is related to correlation and causal among features, and influencing features that affect academic performance. The knowledge from this part will be used as collection of predictive features, questionnaire design for data collection, and generation of synthetic datasets for addition use.

(ii) The related work to predict student performance is used to observe the potential algorithms which were used in predicting academic performance. The analysis investigate the optimal models and gaps of model performance for model

improvement.

(iii) The related works in feature selection methods is used to observed performance of combination of prediction models and the optimal feature selection methods that will be used for the merits of prediction models and mining the learning behaviors for intervention of learning improvement.

(iv) The related works in academic prediction systems presents the history of early warning systems or prediction systems that have been used in educational environment under the application of EDM methods.

The summary knowledge from this chapter is used as the background knowledge of the study. The investigation on various literature review is carried out for two main purposes. First, the study is to extract preliminary knowledge the state-of-the-art methods that have been studied so far to be the background of this study. The second consideration is to figure out the gap of previous works in order to improve the recent work. The existing works have done good work, however, it does not answer our proposed research questions yet. Hence, the following study in this dissertation will answer our research questions.

CHAPTER 3

EDUCATIONAL DATA MINING METHODS

3.1 Introduction

The analysis and extraction of information and patterns from the vast amount of data is called data mining. When the data comes from an educational environment we are dealing with a subdomain of data mining called Educational Data Mining (EDM). This field of research that applies data mining, statistics, and machine learning to data derived from educational environments. EDM has emerged in the past two decades due to the large volume of educational data that was made available. It is concerned with developing and applying data mining methods to detect patterns in large amounts of educational data, and to better understand students and their learning environments.

In the recent decade, innovation and information technology have proven its significance in many areas of applications. EDM has gained increasingly interest in exploring data in educational contexts has mentioned in Chapter 2. EDM combines several interdisciplinary fields of study such as machine learning, statistics, data mining, information retrieval, psycho-pedagogy, cognitive psychology, recommended system methods and techniques to various educational datasets to resolve educational issues. The past couple of years has seen a rapid rise of the number of research papers dedicated to EDM in its various forms [53]. This has been linked to the increase of available educational data and the widespread availability of cheap computing power and accessible digital

tools [54]. With such a wide availability of high-quality data and the potential to derive valuable educational insights, educational institutions, governments and researcher are increasingly looking for ways to put these techniques to good use. The data analyzed in EDM come from various sources like Learning Management Systems, administrative data from universities and schools and other structured or unstructured databases pertaining to education. Due to the habitually large size of these databases, they require a computerized approach to discern the patterns and relationships they contain [55]. In this case, an EDM approach is recommendable to extract the information it contains.

As illustrated in Chapter 2, EDM searches for effective methods to detect the patterns in a large collection of educational datasets which was previously hard or impossible to analyze. The aim of EDM is to develop techniques and methods that can manage these kinds of datasets. EDM includes a wide range of methods from statistical methods, machine learning, to deep learning. Intentionally, EDM proposes techniques to detect the structure and correlation of features, predict student performance or achievement of learning, and observe the features or behaviors affecting student performance.

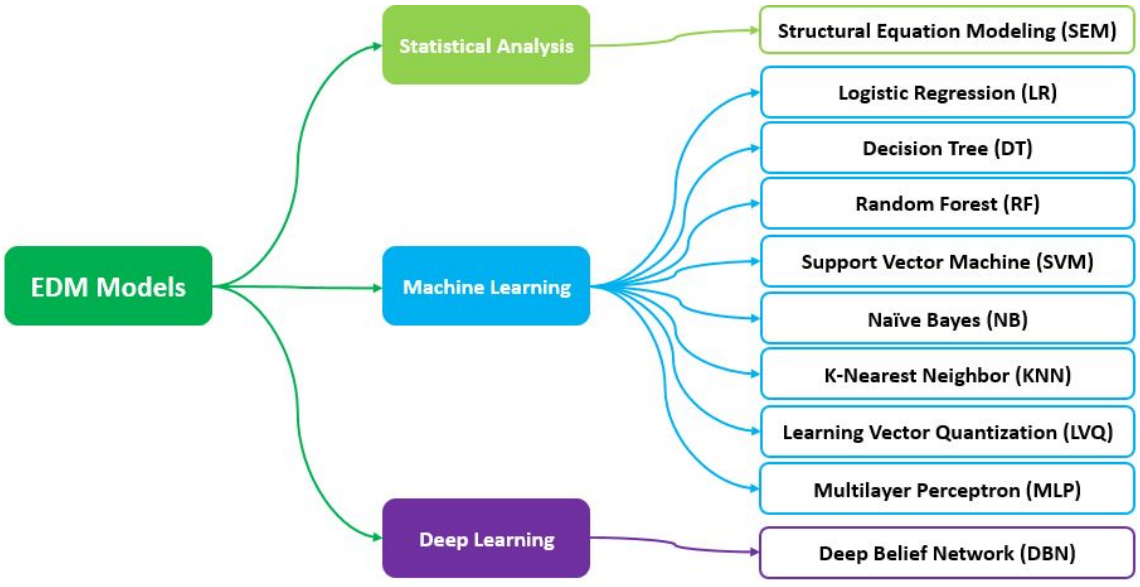


Figure 3.1: The Proposed baseline educational data mining techniques

There is no rule that a particular model is best model; this relies on experience. That is the reason we come up with mentioning several literature reviews and figure out the popular approaches and make a comparative study of these approaches. In our study, we implemented spot-checking algorithms on our dataset and see what works and drop what doesn't work. In this dissertation, we searched for high performance models/techniques with our data and categorized these techniques into three main classes with respect to the revolution, contribution, purpose, and merits as in Figure 3.1.

3.2 Statistical Analysis Techniques

Statistical analysis is an integral technique in the research area. In an education environment, research popularly applied the analysis of variance (ANOVA) test, factor analysis, structural equation modeling, path analysis, and multilevel regression to observed the correlation and structure of indicators in academic performance. We observed that among these techniques, structural equation modeling is a popular tool that was generated by the geneticist Sewall Wright [56]. In analyze factors affecting student performance, this technique is desired by numerous researchers [28][29][30][57][58] since it was used to decompose multiple and interrelated dependence on a single analysis.

3.2.1 Structural Equation Modeling

Structural equation modeling (SEM) comprise two components, a measurement model and a structural model [59]. The measurement model relates observed responses or indicators to latent variables and sometimes to observed covariates. The structural model then specifies relations among latent variables and regressions of latent variables on observed variables. When the indicators are categorical, we need to modify the conventional measurement model for continuous indicators. However, the structural model can remain essentially the same as in the continuous case.

3.2.2 Factor Analysis

Factor analysis (FA) is a technique used to extract the factors (component) from a group of indicators that are related to each factor [60]. This will be used instead of a group of indicators that have the same group. This is a technique that reduces the number of dimensions or manifest variable and considers the suitability of the extracted factors. By checking the statistics Kaiser-Meyer-Olkin: KMO ($KMO > 0.60$) factors obtained will only validate the considered values [61]. Able to explain the variability of all the factors together (total variance explained) with the inverse of each variable with no apparent extraction factor would greatly benefit this approach. If the value of a high percentage (cumulative explained variance) showed that the factors can represent a good indicator, this can be formulated as follows:

$$F_i = w_{i1}x_1 + w_{i2}x_2 + ... + w_{ip}x_p + e_i \tag{3.1}$$

where F is factor, w is coefficient of variable x, x is manifest variables, e is margin of error.

According to Mohamed [28] there are two types of FA; the first factor is the exploratory factor analysis, which attempts to discover the nature of the constructs influencing a set of responses and the second factor is the confirmatory factor analysis that tests whether a specified set of constructs is influencing responses in a predicted way.

Exploratory factor analysis (EFA)

EFA is the first step of FA occurred prior to the confirmatory factor analysis. EFA was intended to reduce the number of items in the instrument in order to maximize the explained variance and identify the appropriate number of items in each of the variables' factors.

Confirmatory factor analysis (CFA)

A CFA was performed to confirm the findings of the exploratory factor analysis. It is used to test whether measures of a construct are consistent with a researcher's understanding of the nature of that construct (or factor). As such, the objective of CFA is to test whether the data fit a hypothesized measurement model. The relationship of each variable to the underlying factor is expressed by the so-called factor loading. Factor loadings can be interpreted like standardized regression coefficients. The threshold of factor loading is determined to keep or remove variables in the factor.

3.2.3 Models of Structural Equation Modeling

SEM is constructed by two main models, which are called a structural model or inner model and a measurement model [62].

Structural model (inner model)

The structural model presents the relationships or paths among the constructs (or latent variables). The structure shows the relation between latent variables and regressions of latent variables on the observed variable. The structural model can take the same form regardless of response type. Letting j index units or subjects, Muthen [63] specifies the structural model for latent variables η_j as:

$$\eta_j = \alpha + B\eta_j + \psi x_j + \zeta_j \tag{3.2}$$

Here, α is an intercept vector, B a matrix of structural parameters governing the relations among the latent variables, ψ a regression parameter matrix for regressions of latent variables on observed explanatory variables x_j and ζ_j a vector of disturbances (typically multivariate normal with zero mean). Note that this model is defined conditional on the observed explanatory variables x_j. Unlike conventional SEM where all observed variables are treated as responses, we need not make any distributional assumptions regarding x_j.

Measurement model

The measurement model is a model built from the implicit or explicit models that relate the latent variable to its indicators. The distinguishing feature of the measurement model is that it is specified for latent continuous responses y_j^* in contrast to observed continuous responses y_j as in conventional SEM.

$$y_j^* = \nu + \Delta\eta_j + K\phi_j + \epsilon_j \tag{3.3}$$

Here ν is a vector of intercepts, Δ a factor loading matrix and ϵ_j a vector of unique factors or 'measurement errors'. Muthen [63] extend the measurement model by including the term $K\phi_j$ where K is a regression parameter matrix for the regression of y_j^* on observed explanatory variables ϕ_j. As in the structural model, we condition on ϕ_j.

Threshold model

The threshold model is introduced for the categorical response. Each observed categorical response y_{ij} is related to a latent continuous response y_{ij}^* via a threshold model:

$$y_{ij} = \begin{cases} 0 & \text{if} \quad -\infty < y_{ij}^* < t_1 \\ 1 & \text{if} \quad t_1 < y_{ij}^* < t_2 \\ . & . \quad \ldots \\ S & \text{if} \quad t_s < y_{ij}^* < \infty \end{cases} \tag{3.4}$$

This is illustrated for $S-1$ categories. Figure 3.2 shows the four categories ($S = 3$), where the area under the curve are the probabilities of the observed responses.

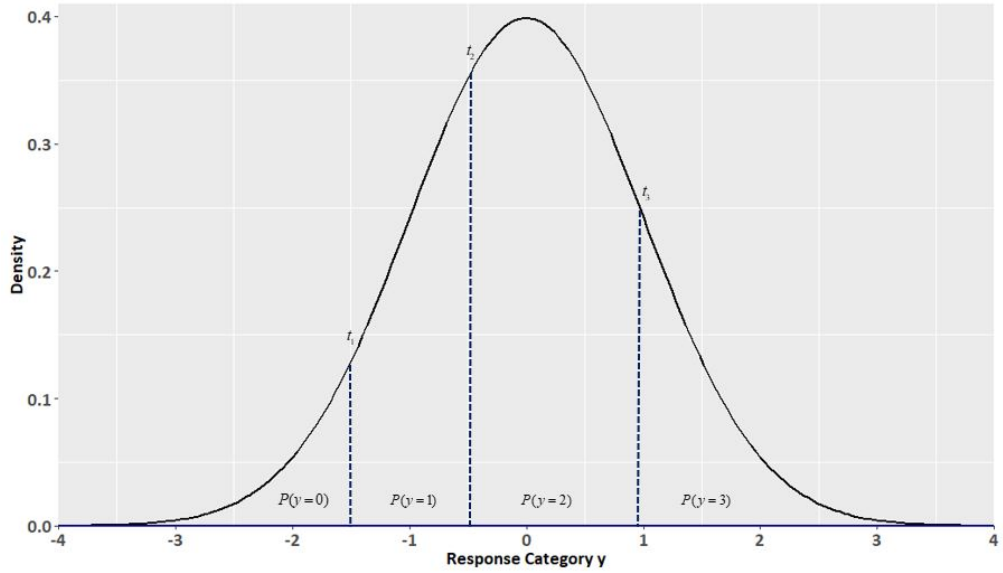


Figure 3.2: Threshold model of SEM

3.2.4 Modification Fit Indexes

In order to measure the goodness-off fitted of a causal model, a statistical measure of many different types of fit indexes has been used. The most frequency used indexes are model relative Chi-square χ^2/df index, goodness of fit index and adjust-goodness of fit index, normed fit index, comparative fit index, and the root mean square of error approximation index [122][65].

Model Relative Chi-square χ^2/df (CMINDF)

The statistically based measure of fit is the Likelihood Ratio Test, better known as the "Chi-square" (CMIN) statistic and its associated "probability" or p-value. The ratio of misfit, quantified by χ^2, to the model degrees of freedom, df, is an absolute fit index. In case of model fit the value of Chi-square (CMIN/DF or simply CMINDF [66]) is generally assumed to be between 1 and 5 with significant value $p = 0.05$.

Goodness-of-Fit Index (GFI) and the Adjusted Goodness-of-Fit Index (AGFI)

GFI is a measure of fit between the hypothesized model and the observed covariance matrix. AGFI corrects GFI, which is affected by the number of indicators of each latent variable. GFI and AGFI provide an absolute metric for how well the model describes the data [65]. GFI is conceptually related to the proportion of variance explained by the model, as in regression. Thus, it ranges between 0 (bad) and 1 (perfect). As with GFI, values for AGFI also range between 0 and 1 and it is generally accepted that values of 0.90 or greater indicate well fitting models.

Normed Fit Index (NFI)

One of the important index in evaluating SEM model is NFI [67]. This statistic assesses the model by comparing χ^2 value of the model to the χ^2 of the null model. The independence model is the worst case scenario as it specifies that all measured variables are uncorrelated. Values for this statistic range between 0 and 1 with Bentler and Bonnet [67] recommending values greater than 0.90 indicating a good fit.

Comparative Fit Index (CFI)

CFI is a revised form of NFI which takes into account sample size that performs well even when sample size is small [68][69]. As with NFI, values for this statistic range between 0.0 and 1.0 with values closer to 1.0 indicating good fit. A cut-off criterion of $CFI \geq 0.90$ was initially advanced however, recent studies have shown that a value greater than 0.90 is needed in order to ensure that misspecified models are not accepted [70].

Root Mean Square Error of Approximation (RMSEA)

RMSEA is an important fit statistic that was first developed by Steiger and Lind [71]. RMSEA tells us how well the model, with unknown but optimally chosen parameter estimates would fit the populations covariance matrix [72]. In recent years it has become regarded as one of the most informative fit indexes [73] due to its sensitivity to the number

of estimated parameters in the model. In other words, RMSEA favours parsimony in that it will choose the model with the lesser number of parameters. Recommendations for RMSEA cut-off points have been reduced considerably in the last fifteen years. Up until the early nineties, an RMSEA in the range of 0.05 to 0.10 was considered an indication of fair fit and values above 0.10 indicated poor fit [74]. It was then thought that an RMSEA of between 0.08 to 0.10 provides a mediocre fit and below 0.08 shows a good fit. However, more recently, a cut-off value is greater than 0.05 [75], close to 0.06 [70] or a stringent upper limit of 0.07 [71] seems to be the general consensus amongst authorities in this area. One of the greatest advantages of RMSEA is its ability for a confidence interval to be calculated around its value. This is possible due to the known distribution values of the statistic and subsequently allows for the null hypothesis (poor fit) to be tested more precisely [76]. It is generally reported in conjunction with RMSEA and in a well-fitting model the lower limit is close to 0 while the upper limit should be less than 0.08 [77].

The summary of thump values of modification fit indexes used in evaluating SEM model is presented in Table 3.1.

Table 3.1: Evaluation of Fit Indexes

Fit indexes	Rule of Thump	Source
CMINDF (χ^2/df)	$1 < \chi^2/df < 5$	Hair et al. [65] & Kleine [66]
GFI	> 0.90	Hair et al.[65] & Steiger [71]
AGFI	> 0.90	Hair et al.[65] & Steiger [71]
NFI	> 0.90	Hair et al.[65], Bentler et al. [67] & Steiger [71]
CFI	> 0.90	Bentler et al. [67], Hu et al. [70], & Byrne [72]
RMSEA	≤ 0.05	Joreskog et al. [75]
	≤ 0.06	Hu et al. [70]
	≤ 0.07	Steiger [71]
	≤ 0.08	Hooper et al. [77]

3.3 Machine Learning Algorithms

The application of machine learning (ML) covers a wide range of research areas. In the last decades, ML has been a paradigm shift toward the education field in evaluating higher education tasks and to extract the hidden useful information and desired knowledge to support the decision process for the progress of students. Large-scale and complexity of education databases required a suitable prediction and classification algorithms to discover meaningful patterns. ML is confirmed as an automated process that extracts these patterns from these undesirable data to be desirable data [78][79][80].

The four main tasks of machine learning are supervised learning, unsupervised learning, reinforcement, and recommender system. In that, supervised learning is a special class of machine learning tasks that create a rule for predicting and classifying unseen data by experience from a given set of instances. The algorithms used in this task is to determine the learning function L using input data E (experience or training instances) and output variables O, denoted as $O = L(E)$. Hence, given any arbitrary (test) data, we can predict the output variables. The aim is to get such a learning function L, which was named as a prediction model (algorithm). Distinct algorithm is going to handle a problem differently and there is no one algorithm for machine learning. The different tasks will require different algorithms. Hence, it is a good idea to observe different types of algorithms to see which is going to work best. We searched for optimal models for our problem by observing the vast machine learning classifiers [81][82], yet from our literature reviews of popularly used machine learning algorithms which are best used in predicting academic performance. The the background knowledge of the proposed machine learning algorithms used in this comparative study are presents in the following subsections.

3.3.1 Logistic Regression

Logit (logistic) Regression Model

In fact, Multinomial Logistic Regression (MLR) model is a fairly straightforward generalization of the binary model, and both models depend mainly on logit analysis or logistic regression [83]. Logit analysis in many ways is the natural complement of ordinary linear regression whenever the response is categorical variable. When such discrete variables occur among the explanatory variables they are dealt with by the introduction of one or several $(0,1)$ dummy variables, but when the response variable belongs to this type, the regression model breaks down. Logit analysis provides a ready alternative.

For a response variable Y with two measurement levels (dichotomous) and explanatory variable X, let: $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$, the logistic regression model has linear form for logit of this probability

$$Logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \qquad \text{where} \quad odds = \frac{\pi(x)}{1-\pi(x)} \tag{3.5}$$

The $odds = exp(\alpha + \beta x)$, and the logarithm of the odds is called $logit$.

The $logit$ has linear approximation relationship, and $logit$ = logarithm of $odds$. The parameter β is determined by the rate of increase or decrease of the S shaped curve of $\pi(x)$. The sign of β indicates whether curve ascends $(\beta > 0)$ or descends $(\beta < 0)$, and the rate of change increases as $|\beta|$ increases.

Multinomial Logistic Regressions

The logistic regression can be extending to models with multiple explanatory variables. Let n denotes number of predictors for a binary response Y by $x_1, x_2, ..., x_n$, the model for log odds is

$$Logit[P(Y = 1)] = \alpha + x_1\beta_1 + x_2\beta_2 + ... + x_n\beta_n, \tag{3.6}$$

and the alternative formula, directly specifying $\pi(x)$, is

$$\pi(x) = \frac{exp(\alpha + x_1\beta_1 + x_2\beta_2 + ... + x_n\beta_n)}{1 + exp(\alpha + x_1\beta_1 + x_2\beta_2 + ... + x_n\beta_n)}. \tag{3.7}$$

If we have n independent observations with p-explanatory variables, and the qualitative response variable has k categories, to construct the logits in the multinomial case, one of the categories must be considered the base level and all the logits are constructed relative to it. Any category can be taken as the base level, so we will take category k as the base level. Since there is no ordering, it is apparent that any category may be labeled k. Let π_j denote the multinomial probability of an observation falling in the j^{th} category, to find the relationship between this probability and the p explanatory variables, $X_1, X_2, ..., X_p$ multiple logistic regression model then is

$$\log\left(\frac{\pi_j(x_i)}{\pi_k(x_j)}\right) = \alpha_{0i} + \beta_{1j}x_{1i} + +\beta_{2j}x_{2i} + ... + +\beta_{pj}x_{pi}, \tag{3.8}$$

where $j = 1, 2, ..., (k-1)$, $i = 1, 2, ..., n$. Since all π's add to unity, this reduces to

$$\log\left(\pi_j(x_i)\right) = \frac{exp(\alpha_{0i} + \beta_{1j}x_{1i} + +\beta_{2j}x_{2i} + ... + +\beta_{pj}x_{pi})}{1 + exp(\alpha_{0i} + \beta_{1j}x_{1i} + +\beta_{2j}x_{2i} + ... + +\beta_{pj}x_{pi})} \tag{3.9}$$

For $j = 1, 2, ..., (k-1)$, the model parameters are estimated by the method of ML. MLR is often considered an attractive analysis because it does not assume normality, linearity, or homoscedasticity.

3.3.2 Decision Tree

Decision Tree (DT) is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision. DT is commonly used for the purpose of decision making. It is one of supervised machine learning algorithms which has been applied in various application. In EDM, DT is among the most popular prediction and classification models used for predicting academic performance.

C5.0 Algorithm

DT is "non-parametric white-box models" which is simple and effective for classification and regression task while C5.0 is one of the most famous algorithms of decision tree works by splitting the sample based on the field that provides the maximum information gain (IG) [85]. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned.

C5.0 algorithm is a successor of C4.5 algorithm, it builds decision trees from a set of training data using the concept of information entropy. C5.0 algorithm uses the concept of entropy to measure the purity of a subset, a term that comes from information theory

$$Entropy(S) = I(S) = -\sum_{i=1}^{c} P(c_i)\log_2 P(c_i) \tag{3.10}$$

In order to decide on which attribute the data segment should be split, the algorithm computes the change in entropy resulting from a split on each of the possible attributes, which is called IG.

$$IG(S) = I(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} I(S_v) \tag{3.11}$$

What complicates matters is that a split results in more than one segment, which is why we also need the so-called split information:

$$Splitinf_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{3.12}$$

After that we calculate IG, by which we can further calculate the information gain ratio. The information gain ratio can be calculated as follows:

$$GainRatio(A) = \frac{IG(A)}{SplitInf(A)} \tag{3.13}$$

C5.0 of DT is one among the popular ML algorithms that has been use in various applications.

CART Algorithm

CART stands for Classification and Regression Trees [86]. It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the towing criteria and the obtained tree is pruned by cost complexity Pruning. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. CART algorithm deploys the method of Gini index to originate binary splits. Gini index is determined by deducting the sum of squared of probabilities of each class from one, mathematically, Gini index can be expressed as:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{3.14}$$

where p_i denotes the probability of an element being classified for a distinct class. An important feature of CART is its ability to generate regression trees. Regression trees are trees where their leaves predict a real number and not a class. Usually the splitting algorithm will isolate outliers in individual node or nodes. An important practical property of CART is that the structure of its classification or regression trees is invariant with respect to monotone transformations of independent variables. One can replace any variable with its logarithm or square root value, the structure of the tree will not change. CART methodology consist of three parts:

(1) Construction of maximum tree

(2) Choice of the tree size

(3) Classification of new data using constructed tree

Table 3.2 illustrates the performance comparison of decision tree algorithms.

Table 3.2: Comparison of Decision Tree Classifiers

	ID3	C4.5	C5.0	CART
Type of data	Categorical	Continuous and Categorical	Continuous, Categorical, dates times, timestamps	Continuous and Nominal attributes data
Speed	Low	Faster than ID3	Highest	Average
Pruning	No	Pre-pruning	Pre-pruning	Post pruning
Boosting	Not supported	Not supported	Supported	Supported
Missing values	Can't deal with	Can't deal with	Can deal with	Can deal with
Formula	Use information entropy and information gain	Use split information and gain ratio	Same as C4.5	Use Gini diversity index

3.3.3 Random Forest

Random Forest (RF) developed by Leo Breiman [87] is a group of un-pruned classification or regression tree. As the name suggest, this algorithm creates the forest with a number of trees. A random forest algorithm is a tree-based tool that grows many classification trees. RF is an ensemble learning method that constructs multiple decision trees. Each decision tree can be implemented by CART procedure. CART recursively partitions on a nominal target category to reach a tree structure. The input of CART can be nominal or numerical. As the decision tree grows, a feature must be identified to split on it. So, all features are compared to each other to select the best feature. This comparison can be done by Gini index that measures pureness of feature separation. CART stopping rule occurs when the target feature in the last separations are insignificant.

RF samples randomly training data with replacement on constructing each decision tree that is called bagging. Each decision tree returns a class and then bagging combines them to reach a unique decision [86]. The configuration of RF algorithm is presented in

Figure 3.3.

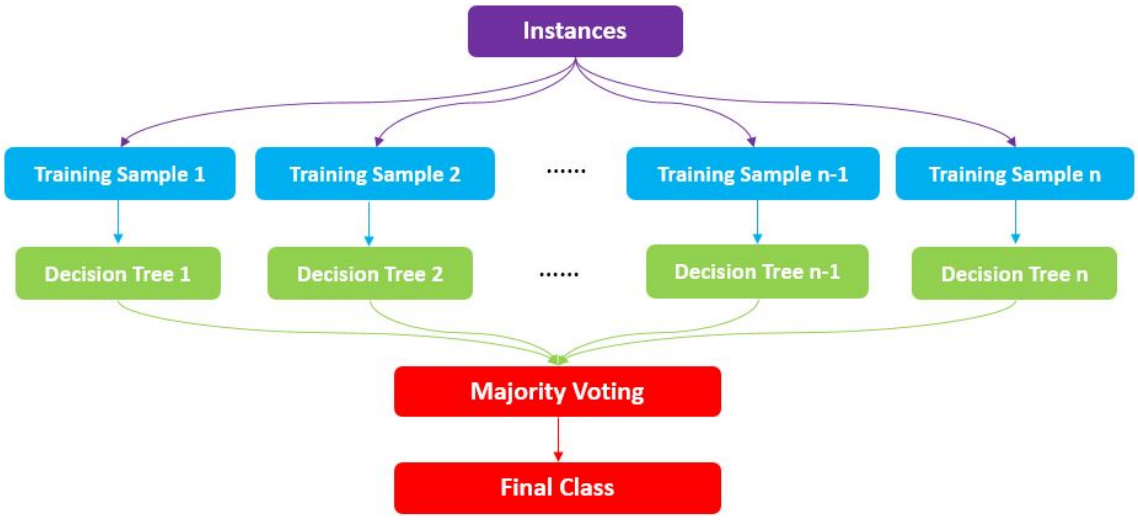


Figure 3.3: Random forest algorithm

3.3.4 Ensemble Methods

Bagging Algorithms

In ensemble methods, bootstrap aggregating (bagging) and boosting are the two popular paradigms. As indicates, bagging improves the accuracy of predictions by aggregating predictions from multiple learners. In bagging, the m base learners are trained in parallel on m bootstrap samples from original dataset, and then the m predictions from the m base learners are combined using majority voting or other synthesizing methods. The illustration of this method is shown in Figure 3.4. Because of its aggregating (or averaging) nature, bagging can reduce the variance of a model. RF is an example of ML algorithm that uses bagging paradigm. RF is the collection of m decision trees, and combines m predictions from the m decision trees trained on the m bootstrap samples to make a final prediction.

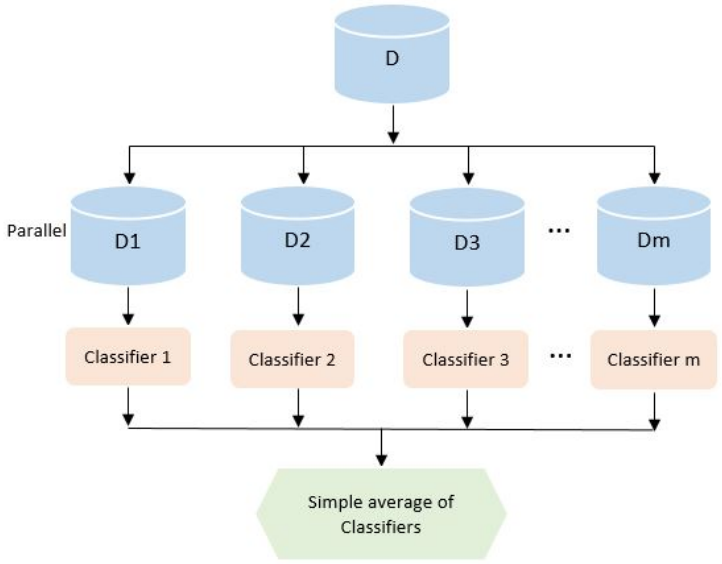


Figure 3.4: Ensemble methods with Bagging algorithms

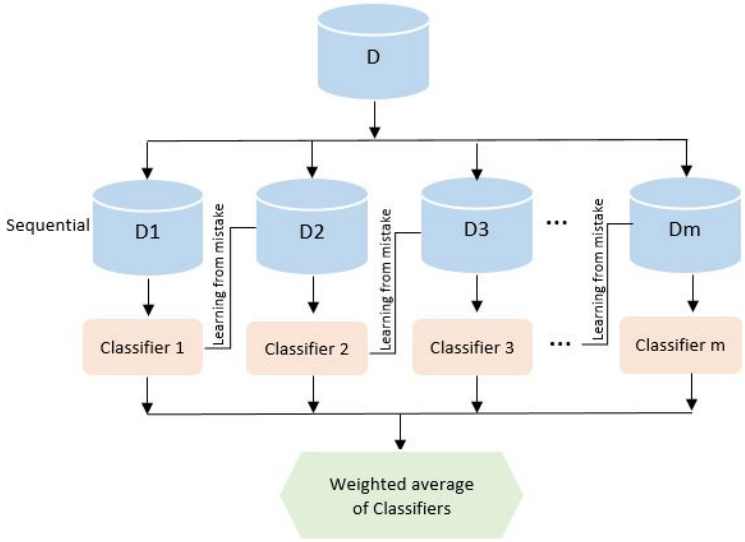


Figure 3.5: Ensemble methods with Boosting algorithms

Boosting Algorithm

Boosting is another ensemble paradigm. The idea behind is to produce a series of a weak learners to make a strong learner. Unlike bagging in which learners are trained in

parallel and the predictions are aggregated without preference to any learner, boosting sequentially trains learners, and the predictions are aggregated with heavier weights on better learners. The illustration of boosting algorithm is shown in Figure 3.4. The boosted C5.0 is introduced in this classification problem.

3.3.5 Support Vector Machine

Support Vector Machine (SVM) was invented by Vladimir Vapnik [88]. SVM can be briefly described as follows [89][90]. Let $(x_1, y_1), (x_2, y_2), ..., (x_m, y_m) \in \mathbb{X} \times \{-1, 1\}$ denote a set of training data, where $\mathbb{X}$ represents some nonempty set from which the pattern x_i; y_i are called target $y_i \in \{-1, 1\}$.

SVM attempts to identify a hyper-plane in a multidimensional space. This hyper-plane functions as a separating plane for data classification. Consider the class of hyper-planes in some dot product space H. The parameters w and b are described as follows:

$$(< w, x > +b) = 0 \qquad \text{where} \quad w \in W, b \in \mathbb{R} \tag{3.15}$$

If a hyper-plane that satisfies equation (3.15) exists, then linear separation is obtained. In this case, w and b can be rewritten as follows: equation (3.15) becomes

$$\max_{w \in H, b \in R} \min_{1 \le i \le N} \{\|x - x_i\| \, | x \in H, < w, x > +b = 0, i = 1, 2, ..., m\} \tag{3.16}$$

Let the distance from the data point to the hyper-plane be $\frac{1}{\|w\|}$.

The separating hyper-planes include one optimal separating hyper-plane (OSH), which has the largest distance between two support vector points on its two sides. Because the distance between two support vector points is $\frac{1}{\|w\|^2}$, the minimal distance to OSH, $\|w\|^2$, can be derived from equation (3.16).

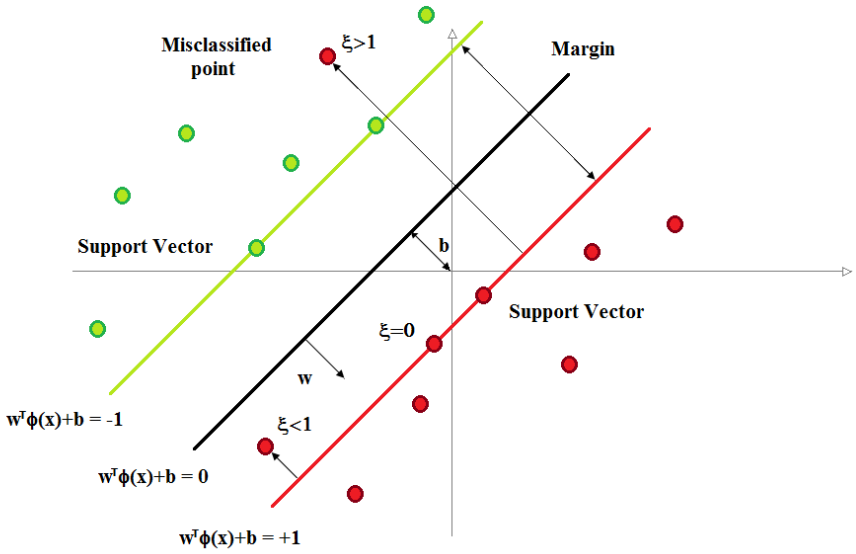


Figure 3.6: Support vector machine classification

The margin $\frac{2}{\|w\|^2}$ of a separating hyper-plane can be regarded as the hyper-plane's generalization ability, and OSH has the maximal margin among separating hyper-planes. $\|w\|^2$ can be minimized with equation (3.16) and Lagrange's polynomial. Let a denote $(a_1, ..., a_m)$. Combining Lagrange's polynomial (in the order of m) with equation (3.16) produces the following maximization equations:

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle \tag{3.17}$$

$$\text{subject to } 0 \leq \alpha_i \leq C \text{ for all } i = 1, 2, ..., m \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3.18}$$

Quadratic programming can be applied to solve this maximization problem.

Given a vector satisfies equation (3.16) in maximization, OSH expressed in terms of (w, b) can be written as follows:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i, \tag{3.19}$$

where the support vector points must satisfy $\alpha_i \geq 0$ and equation (3.16). When considering expanding in constraint equation (3.18), the determinant function of

hyper-plane is written as follows:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \langle x, x_i \rangle\right) = 0. \tag{3.20}$$

In most cases, the data is not linearly separable, and is thus mapped to higher dimensional feature space. This means that if the data cannot be classified explicitly in the current dimensional space, then SVM algorithm maps the data to a higher dimensional space for classification.

The input data are mapped to a higher dimensional feature space by plotting a nonlinear curve. OSH is built in the feature space. The feature space vectors x, x' are constructed in terms of the kernel k evaluated on input patterns $x, x', k(x, x') = \langle x, x' \rangle$.

The kernel function can be applied because all feature vectors occur only in dot products. The weight vector then becomes an expansion in feature space, and therefore typically no longer corresponds to the ϕ-image of a single input space vector.

The decision function is expressed as follows:

$$\begin{aligned} f(x) &= \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \langle \phi(x), \phi(x') \rangle + b\right) \\ &= \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i k(x, x') + b\right). \end{aligned} \tag{3.21}$$

The quadratic program (QP) or dual optimization problem is written as

$$\max_{\alpha \in \mathbb{R}^n} \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \tag{3.22}$$

$$\text{subject to } 0 \leq \alpha_i \leq C \text{ for all } i = 1, 2, ..., m \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0. \tag{3.23}$$

Several kernel functions help SVM in finding the optimal solution. The most frequently used such functions are the polynomial kernel, sigmoid kernel and radial basis kernel

function (RBF) [91][92]. RBF is used most often in general cases, since it can classify multi-dimensional data, unlike a linear kernel function. Furthermore, RBF has fewer parameters than a polynomial kernel. Overall, performance of an RBF and other kernel functions is not significantly different. Consequently, RBF is an effective choice for the kernel function. Therefore, this study employs an RBF kernel function in the SVM to discover the optimal solution.

Sequential Minimal Optimization (SMO)

Sequential Minimal Optimization (SMO) algorithm is derived by taking the idea of the decomposition method to its extreme and optimizing a minimal subset of just two points at each iteration. SMO is the optimized SVM which breaks the large quadratic programming (QP) in SVM into series of smallest possible QP problems.

SMO algorithm is aimed to solve the following QP problem of equation (3.22) and equation (3.23). SMO decomposes overall QP problem into QP sub-problems. At every step in QP problems, SMO chooses to solve the smallest possible optimization problem, two Lagrange multipliers to jointly optimize, find the optimal values for these multipliers, and updates SVM to reflect the new optimal values .

3.3.6 Naive Bayes

Naive Bayes (NB) is a simple, yet effective and commonly-used ML classifier. The algorithm has become an attractive method for academic performance evaluation [93]. In educational domain, an assumption of conditional independency is often ignored and disturbed. Considering that variables are inter-connected, NB can tolerate strong supervising dependence between independent variables. It is a probabilistic classifier that makes classifications using a posteriori decision rule in a Bayesian setting. It can also be represented using a very simple Bayesian network. NB classifier has been especially popular for EDM. Suppose D is the dataset of n dimensional vector $X = (x_1, x_2, \cdots, x_n)$:

describing attributes of each student and there be m classes $C_1, C_2, \cdots, C_m$. NB algorithm predicts X belong to class C_i if and only if $P(C_i|X) > P(C_j|X)$ for all $1 \leq j \leq m, i \neq j$. NB classifier is found on conditional Bayes probability as in equation (3.24).

$$P(C_k|X) = \frac{P(C_K) \times P(X|C_k)}{P(X)}. \tag{3.24}$$

The probability $P(X)$ is normalizing constant or total probability and $X = (x_1, x_2, ..., x_p)$ is the set of p predictive features with strong assumption of independence predictors, then equation (3.24) can be written as:

$$P(C_k|X) \propto P(C_k) \times P(X|C_k) = P(C_k) \times \prod_{i=1}^{n} P(x_i|C_k). \tag{3.25}$$

The naïve Bayes classifier holds many advantages such as it is a very simple algorithm, not contain any parameter to optimize, efficient for classification, and easy interpret.

3.3.7 k-Nearest Neighbors

k-Nearest Neighbors (KNN) is considered as a more flexible technique and the simplest machine learning algorithms because it is able to classify test data into label classes by searching for train data which is relatively the same as test data [94][95]. KNN is known to be one of the best state of the art classifiers and could achieve best accuracy. It is an algorithm that falls under the category of supervised learning algorithms [96]. The classification as per this algorithm is done based on Euclidean distance between the training data and the testing data. Based on the similarity between the training and the testing data the nearest k neighbors are selected. Here k is a positive integer and the configuration of KNN classification is shown in Figure 3.7. The label associated with these neighbors is taken as the reference. The testing data is associated to the class which has majority of the votes amongst [97]. When dealing with continuous attributes

the difference between the attributes is calculated using the Euclidean distance. If the first instance is $x_1 = (x_{11}, x_{12}, ..., x_{1n})$ and the second instance is $x_2 = (x_{21}, x_{22}, ..., x_{2n})$ the distance between them is calculated by the following equation (3.26):

$$d(x_1, x_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + ... + (x_{1n} - x_{2n})^2} \tag{3.26}$$

KNN usually deals with continuous attributes however it can also deal with discrete attributes. When dealing with discrete attributes if the attribute values for the two instances x_1, x_2 are different so the difference between them is equal to one otherwise it is equal to zero.

The k-nearest neighbor classification can be summarized as:

(1) Determine the value of k.

(2) Compute the distance between each record of the training set and the testing record.

(3) Sort the neighbors in the increasing order of the distances.

(4) Select the first k neighbors from the sorted list.

(5) Check for the class, that majority of the neighbors belongs to and assign that class to the training data.

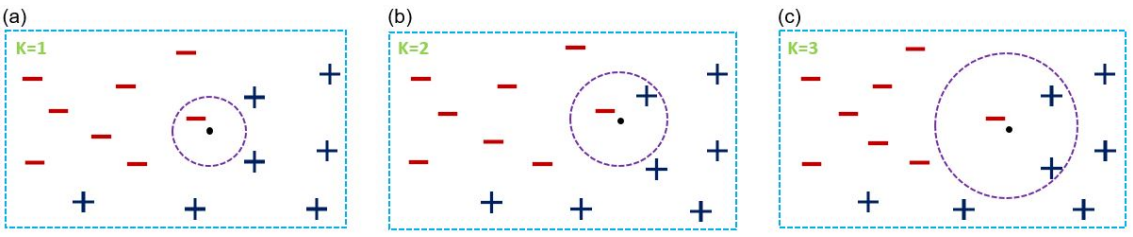


Figure 3.7: k-nearest neighbor(s): (a) 1- nearest neighbor, (b) 2- nearest neighbor, and (c) 3- nearest neighbor

3.3.8 Learning Vector Quantization

The Learning Vector Quantization algorithm (or LVQ for short) is an artificial neural network algorithm that allows you to choose how many training instances to hang onto and learns exactly what those instances should look like [98]. The representation for LVQ is a collection of codebook vectors. LVQ was developed and is best understood as a classification algorithm. It supports both binary (two-class) and multiclass classification problems. A codebook vector is a list of numbers that have the same input and output attributes as your training data.

The model representation is a fixed pool of codebook vectors, learned from the training data. They look like training instances, but the values of each attribute have been adapted based on the learning procedure. In the language of neural networks, each codebook vector may be called a neuron, each attribute on a codebook vector is called a weight and the collection of codebook vectors is called a network.

Predictions are made using the LVQ codebook vectors in the same way as k-Nearest Neighbors. Predictions are made for a new instance by searching through all codebook vectors for the k most similar instances and summarizing the output variable for those k instances. For classification this is the mode (or most common) class value. Typically predictions are made with $k = 1$, and the codebook vector that matches is called the Best Matching Unit (BMU). To determine which of the k instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance. Euclidean distance is calculated as the square root of the sum of the squared differences between a point a and point b across all input attributes i.

LVQ network are based on the supervised competitive learning. LVQ networks attempt to define decision boundaries in the input space, given a set of exemplary decisions (the training data). Topologically, the network contains an input layer, a competitive layer

and an output layer. The output layer has the neurons equal to the number of classes. In the competitive layer, each competitive unit corresponds to a cluster, the center of which is called a codebook vector. The euclidean distance of an input vector is computed with each codebook vector and nearest codebook vector is declared winner. Unlike perceptron, LVQ network can classify any set of input vectors, not just linearly separable sets of input vectors. The only requirement is that the competitive layer must have enough neurons and each class must be assigned enough to competitive neurons.

There are a number of somewhat different LVQ algorithms appearing in the literature, they are all based on the following algorithms:

(1) A learning sample consisting of input vector x_i together with its correct class label c_i is presented to the network.

(2) A suitable number of codebook vectors are selected for every class label c_i.

(3) Using distance measure between codebook vectors and input vector d_i, the winner is determined. In some cases, the second best winner is also determined

LVQ1 [99] selects a single set of best machine codebook vectors is selected and moved closer or further away from each data vector, per iteration if the classification decision is correct or wrong, respectively. This paper used the optimized learning vector quantization (OLVQ1) [100] that is the same to LVQ1, except that each codebook vector has its own learning rate.

3.3.9 Multilayer Perceptron

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements

(neurons) working in unison to solve specific problems. ANN accepts the values of inputs by input nodes which is also called input layer. These input values are then multiplied by a set of numbers which are called weights and stored in the links. After multiplication, these values are added together to become inputs to the set of nodes that are to the right of the input nodes. Finally, the values from the hidden layer are fed into an output node as in Figure 3.8, where a special mapping or thresholding function is applied and the resulting number is mapped to the prediction class.

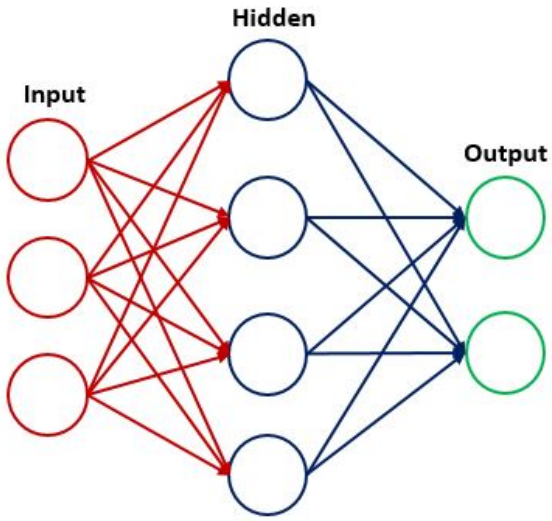


Figure 3.8: The structure of neural network

ANN has been used successfully in numerous business application for pattern recognition, prediction and classification [101]. These properties make them particularly suit for domain like EDM where large quantities of data are available and where the data is often noisy and not linearly separable owing to the fact that it is based on human behavior.

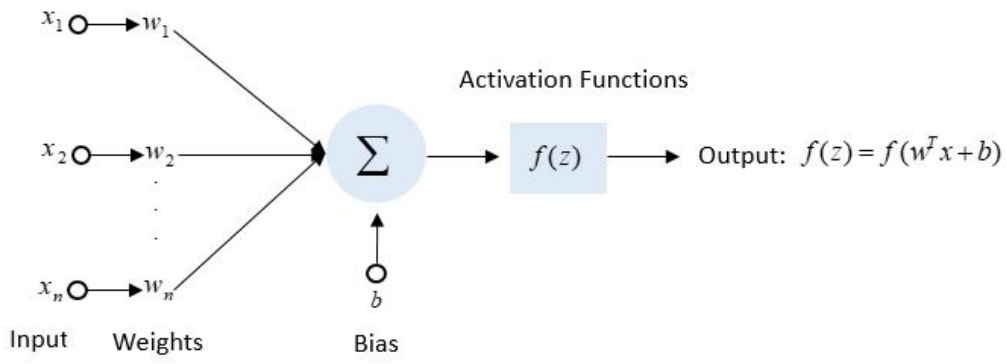


Figure 3.9: Perceptron (single layer perceptron)

Multilayer Perceptron (MLP) is a classical type of ANN. They are comprised of one or more layers of neurons. Data is fed to the input layer, there may be one or more hidden layers providing levels of abstraction, and predictions are made on the output layer, also called the visible layer. MLP are suitable for classification prediction problems where inputs are assigned a class or label.

MLP is feed forward ANN model that maps sets of input data onto a set of appropriate output with more hidden layer. It is made up of interconnected process units that make use of learning algorithms to create models of knowledge, which are saved as weighted connections similar to the way the human brain functions [102]. An ANN uses the feed-forward approach where information is only allowed to move in one direction from the input to the output, and training is achieved using the back propagation algorithm [103]. It is a type of supervised learning algorithm that requires independent attributes and the target classes. MLP is so named because they contain at least one hidden layer in addition to the input and output layers [104].

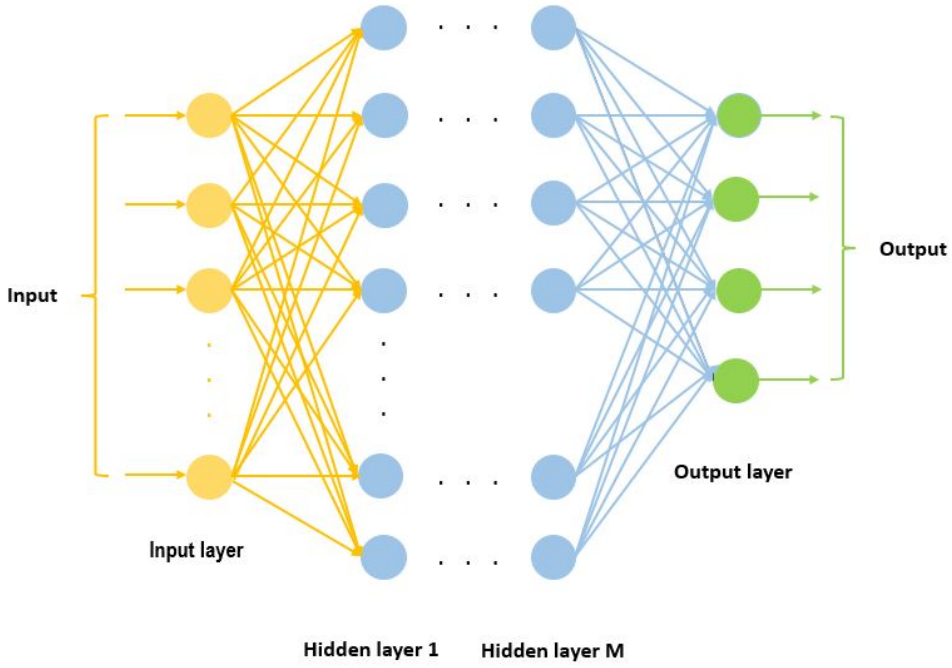


Figure 3.10: Scheme graph of a MLP network

A typical MLP is created by interconnecting artificial neurons. A typical neural network with one hidden layer is illustrated in Figure 3.10. The perceptron is made up of inputs $x = (x_1, x_2, ..., x_n)$ their corresponding weights $w = (w_1, w_2, ..., w_n)$. In MLP, the weighted sum of the inputs and bias term are passed to activation level through a transfer function to produce the output, and the units are arranged in a layered feed-forward topology called Feed Forward Neural Network [105].

MLP networks consist of an input layer, one or more hidden layers and an output layer. Each layer has a number of processing units and each unit is fully interconnected with weighted connections to units in the subsequent layer. MLP transforms n inputs to l outputs through some nonlinear functions. The output of the network is determined by the activation of the units in the output layer as follows:

$$y = f\left(\sum_h w_{ho} x_h + b\right) = f(w^T x + b), \tag{3.27}$$

where f is activation function, x_h: activation of h^{th} hidden layer node and w_{ho}: is the interconnection between h^{th} hidden layer node and o^{th} output layer node. The most used activation functions for MLP to used are the sigmoid and ReLu functions, and it is given as follows:

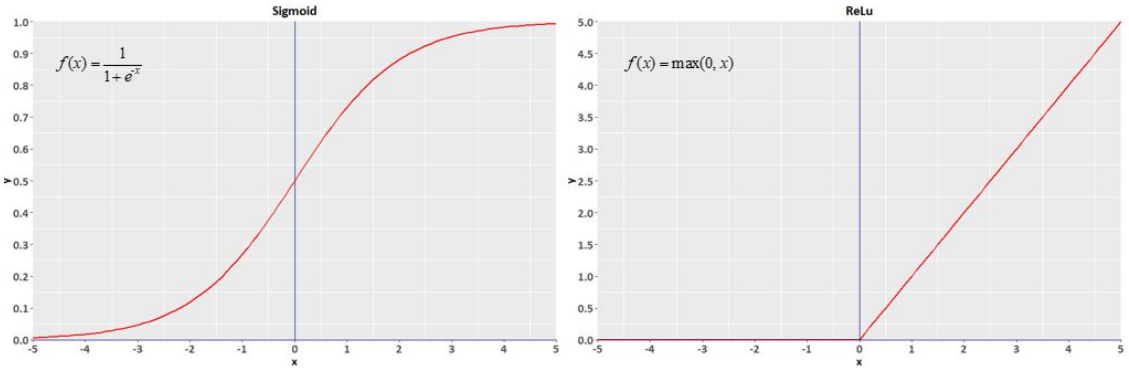


Figure 3.11: Signmoid and ReLu functions

3.4 Deep Learning Algorithms

In education, teachers, administrators, management boards, and associated individuals are always wish to predict and classify their students based on their performance levels, motivation, characteristics, and attitude toward academic performance. The International Educational Data Mining Society [106] defines EDM as "an emerging discipline, concerned with developing methods for exploring the unique types of data come from educational settings, and using those methods to better understand students, and the setting which they learn in". It searches for effective methods to detect the patterns in large collection of educational datasets which was previously hard or impossible to analyze. EDM aims at developing techniques and methods to deal with such kinds of dataset. EDM includes a wide range method from statistical methods, machine learning, to deep learning. Purposely, EDM proposes techniques to detect structure and correlation of features, predict student performance or achievement of learning, and observe features

or behaviors affecting student performance. E-learning platforms such as Massive Open Online Course (MOOC), Learning Management System (LMS), and Intelligent Tutoring System (ITS) are analyzed by EDM to obtain multiform and massive information from students' activities in educational environment. The knowledge from these complex and large data are effectively discovered by architecture of deep learning [107].

Deep Learning (DL), an extension framework of artificial neural network, contributed a wide range of applications and gained so much attention by researchers in the last decade. DL is a recent hot trend in ML and artificial intelligent (AI) that provides an effective result in different fields such as healthcare, natural language processing (NLP), image processing, object detection, finance, and many more. Educational intuitions presently store a large collections of data, such as students' geographical information, students' behaviors such as dropout and attendance records, and their examination score. DL provides the effective techniques to discovers the information from these datasets.

DL contains many architectures such as classical MLP, Recurrent Neural Network (RNN), Long Short Time Memory (LSTM), Convolutional Neural Network (CNN); and unsupervised pertained networks such as Autoencoders, Generative Adversarial Network (GAN), and Deep Belief Network (DBN). These frameworks have role importantly in solving various rea-world complicated problems. The theoretical foundation of DBN was found Geoffrey Hinton [108]. The special architecture of DBN effective solved many problems in healthcare, finance, and natural langue processing. The various application of DBN has been applied to a number of recent studies ranging from image recognition [109][110][111], audio, speech and natural language [112], health care [113][114][115]. In this papers, we proposed a novel trend of DBN to education field. We optimized DBN model for predicting and classifying high school student performance in mathematics.

In this study, we first introduce Restricted Boltzmann Machine (RBM) and DBN which are preliminaries to our proposed method. Then we propose reconstructive learning and discrimative learning approaches for learning and evaluating model for slecting the

highly influenced faftors and students' performance prediction.

3.4.1 Restricted Boltzmann Machine

A restricted Boltzmann machine (RBM) is a generative stochastic artificial neural network that can learn a probability distribution over its set of inputs, with the restriction that their visible units and hidden units must form a fully-connected bipartite graph [108]. It is two-layer stochastic network, a kind of Boltzmann Machine (BM) with terminology "restricted" because no two nodes in the same layer share a connection. That is, RBM is a restricted version of BM. RBM is a method that can automatically find patterns in our data by reconstructing the input.

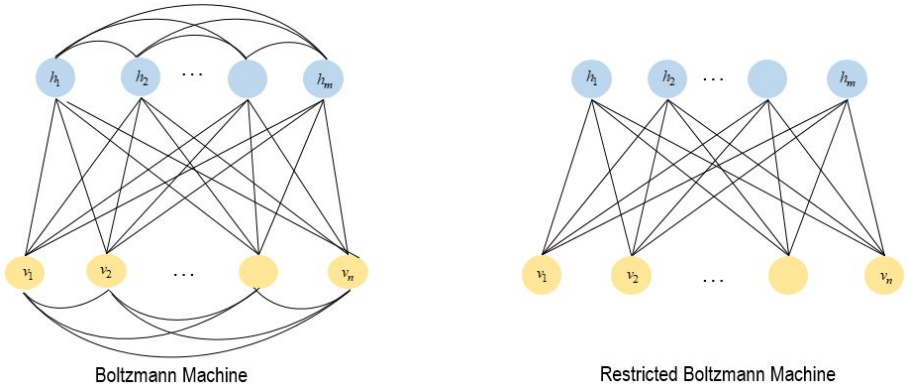


Figure 3.12: Illustration of BM and RBM

The standard type of RBM has binary-valued (Boolean/Bernoulli) hidden and visible units, and consists of a matrix of weights $W = (w_{ij})_{m \times n}$ (size $m \times n$) associated with the connection between hidden unit h_j and visible unit v_i, as well as bias weights (offsets) a_i for the visible units and b_j for the hidden units. Given these, the energy function of a configuration (pair of boolean vectors) (v, h) is defined as

$$E(h, v) = -\sum_{i}^{m} a_i v_i - \sum_{j}^{n} b_j h_j - \sum_{i}^{m} \sum_{j}^{n} v_i w_{ij} h_j, \tag{3.28}$$

or written in matrix notation

$$E(h, v) = -a^T v - b^T h - v^T w h. \tag{3.29}$$

Here, $a = (a_1, a_2, ..., a_n)$ is the offset (bias vector) of visible layer $v = (v_1, v_2, ..., v_n)$, and $b = (b_1, b_2, ..., b_m)$ is the bias vector of hidden layer $h = (h_1, h_2, ..., h_m)$. Denote $\theta = (a, b, w)$ is the model parameters, n and m are the number of visible units and hidden units, respectively. The joint probability configuration over the hidden layer and/or visible layer is defined in term of energy function as:

$$P(h, v) = \frac{1}{Z} e^{-E(h,v)}, \tag{3.30}$$

where $Z = \sum_v \sum_h e^{-E(h,v)}$ is called the normalization factor, or partition function.

The probability that the network assigns to a visible unit is achieved by the summation of all hidden units:

$$P(v) = \frac{1}{Z} \sum_h e^{-E(h,v)}. \tag{3.31}$$

From equation (3.31) we construct the log likelihood probability as follows

$$\ln P(v) = \ln\left(\sum_h e^{-E(h,v)}\right) - \ln\left(\sum_v \sum_h e^{-E(h,v)}\right). \tag{3.32}$$

The model parameters $\theta = (a, b, w)$ are possible to solve respectively by the maximum likelihood function methods:

$$\frac{\partial \ln P(v)}{\partial a} = E_{Pd}[v] - E_{Pm}[v], \tag{3.33}$$

$$\frac{\partial \ln P(v)}{\partial b} = E_{Pd}[h] - E_{Pm}[h], \tag{3.34}$$

$$\frac{\partial \ln P(v)}{\partial w} = E_{Pd}[v h^T] - E_{Pm}[v h^T], \tag{3.35}$$

where E_{Pd} denotes the expectation of the input conditional probability distribution of training data and E_{Pm} denotes the expectation of the joint probability of the model.

Since there is no connection between neuron in the same layer in an RBM, then it is easy to calculate the expectation of data probability distribution.

Given randomly selected visible unit v, the binary state h_j, of each hidden unit, j, is set to 1 with probability

$$P(h_j = 1|v) = \sigma\left(b_j + \sum_{i=1}^{n} w_{ij} v_i\right), \tag{3.36}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic sigmoid function. Because there are no direct connections between visible units in an RBM, it is also very easy to get an unbiased sample of the state of a visible unit, given a hidden vector:

$$P(v_i = 1|h) = \sigma\left(a_i + \sum_{j=1}^{m} w_{ij} h_j\right). \tag{3.37}$$

In order to obtain updated weights and biases, it is required to perform gradient ascent in the log-likelihood that can be calculated from the rule:

$$\Delta w = \eta(E_{Pd}[vh^T] - E_{Pm}[vh^T]), \tag{3.38}$$

where η is called the learning rate. However, the expectation E_{Pm} cannot be derived analytically. To avoid computing E_{Pm} we compute reconstructed expectation E_r of distribution of samples using "Contrastive Divergence" (CD) [116] via Gibbs sampling method. Instead of using equation (3.38), we use:

$$\Delta w = \eta(E_{Pd}[vh^T] - E_r[vh^T]). \tag{3.39}$$

The Gibbs sampling normally can be repeated many times to converges closer to model distribution. Howerver, in RBM training process, the Minimizing Contrastive Divergence (MCD) training rule for an RBM replace the computationally expensive relaxation search of the Boltzmann Machine with a single step of Gibbs sampling [116][117]. In each iteration, the weight w can be updated based on the MCD rule, as in Figure 3.13.

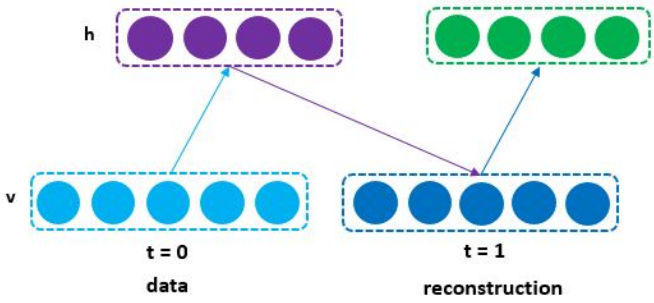


Figure 3.13: Illustration of single step of CD or MCD algorithm

The purpose of training RBM is to find out the best parameter set $\theta = (a, b, w)$, that is why the maximum likelihood function is maximized approximately by gradient descent.

3.4.2 Deep Belief Network Architecture

Deep Belief Network (DBN) is class of deep neural network, which were conceived by Geoffrey Hilton [118]. It is multi-layer belief network, which each layer is RBM and they are stacked each others to construct DBN as in Figure 3.14. DBN is a probabilistic generative model, that is composed of multi-layer of stochastic, latent variables. The latent variables typically have binary values and are often called hidden units or feature detectors.

The top two layers form a RBM which can be viewed as an associative memory. The states of the units in the lowest layer represent a data vector. Because of the specific structure of RBMs, visible and hidden layers $h^{(k)}$ $(k = 1, ..., l, l \geq 2)$ are conditionally independent given one-another. Using this property, we can write:

$$\begin{aligned} P(h^{(1)}, h^{(2)}, .., h^{(l)}|v) &= P(h^{(l)}|h^{(l-1)})P(h^{(l-1)}|h^{(l-2)}) \cdots P(h^{(1)}|v) \\ &= \left(\prod_{k=2}^{l} P(h^{(k)}|h^{(k-1)})\right) \times P(h^{(1)}|v) \end{aligned} \tag{3.40}$$

The probability of the bottom-up inference from observed variables V to hidden layer

$h^{(k)}$, and so on is defined as

$$P(h_j^{(k)}|h^{(k-1)}) = \sigma\left(c_j^k + \sum_i^n w_{ij}^k h_i^{k-1}\right), \tag{3.41}$$

where c^k is denoted as the bias for layer k.

Similarly, the top-down inference is a symmetric version of the bottom-up inference, which can be written as

$$P(h_i^{(k-1)}|h^{(k)}) = \sigma\left(c_i^{k-1} + \sum_{j=1}^m w_{ij}^{k-1} h_j^k\right), \tag{3.42}$$

where c^{k-1} is denoted as the bias for layer $k-1$.

The training of DBN can be divided into two steps: pre-training and fine-tuning using backpropagation. Pre-training followed by fine-tuning is a powerful mechanism for training a deep belief network.

The supervised DBN is conducted by adding a final layer in unsupervised DBN architecture that represent the desired output then performing discriminative fine-tuning phase using back-propagation. Using back-propagation to fine-tune feature detectors of hidden layers that are initially learned in the stacked trained RBM. The merits of this process is much better than using back-propagation with random initial weights that used in the traditional neural network.

A DBN training is divided into two main stages. The first stage is unsupervised pre-training called training section that learns features only with the input values without a label. The second stage is tuning with backpropagation algorithm using a label with supervised fine-tuning, called tuning section.

- **(1). Pretraining Section:** The training section is implemented by using layer by layer training approach called "Greedy Layer" training algorithm as shown in Figure 3.14. This section learning the parameters in each layer of RBM. The first RBM is training using MCD algorithms to reconstructs its inputs to be accurate as

possible. The hidden layer of the first RBM is treated as visible layer for the second RBM, and the output of second layer is inserted to be visible layer of the next RBM, and so on. This procedure is repeated until every predefined hidden layer in the network is trained. This pretraining is implemented to obtain the pre-trained parameters before the supervised training work.

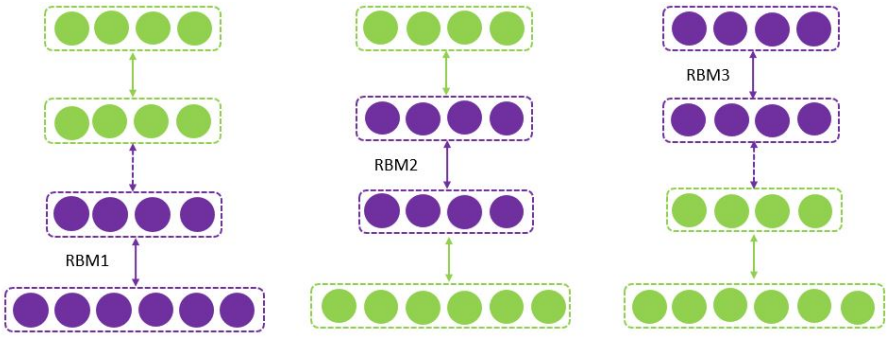


Figure 3.14: Layer by layer training of Deep Belief Network

- **(2). Fine-tuning Section:** This section is supervised training stage. It uses labeled samples to supervise DBN model in the top-down phase. The trained weights and biases of RBM layers from unsupervised RBM learning are utilized in this section. The algorithms use parameters obtained in the training section as it will avoid falling into local optimum and overfitting which caused by random initialization of the traditional network. The training is stopped after reaching a predefined number of epochs, or is finished successfully after reaching the target error rate. The supervised learning process further reduces the training error and improves the classification accuracy of DBN classifier model. The procedure of DBN is illustrated in Figure 3.15.

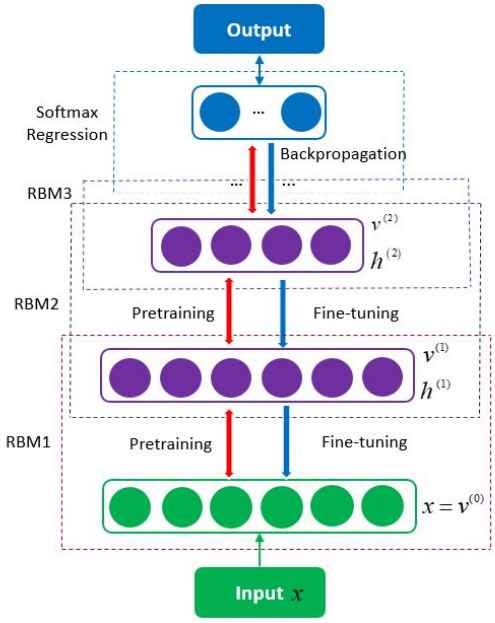


Figure 3.15: Framework of supervised DBN Model

3.5 Feature Selection and Feature Extraction Methods

Dimensionality reduction plays an important role in classification performance. In data mining, to catch useful indicators and obtain a more accurate result, we tend to add as many features as possible at first. However, after a certain point, the performance of the model will decrease with the increasing number of elements. This phenomenon is often referred to as *The Curse of Dimensionality*. In statistics, machine learning, and information theory, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. There are two components of dimensional reduction: Feature Selection (FS), and Feature Extraction (FE) [121]. FS is a technique of filtering optimal set of lower dimension that contain only informative features, while FE is a technique of getting lower dimension by generating new features based on transformation or combination of original feature set. The configuration of these two approach is illustrated

in Figure 3.16.

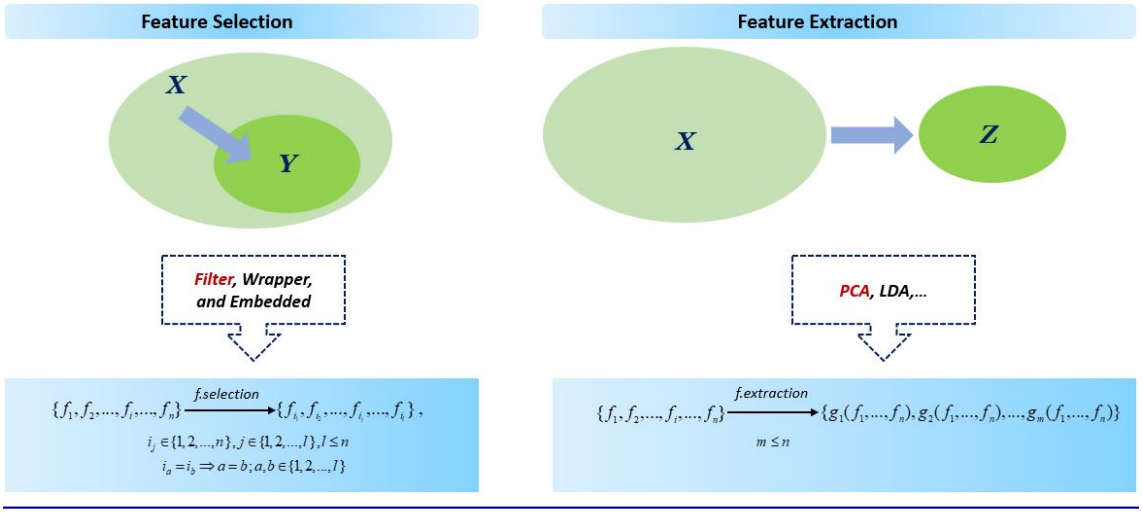


Figure 3.16: Feature engineering: Feature selection and feature extraction

This section present a literature on application of FS to educational data mining. The presented feature selection was used for two main purposes. The first use of the feature selection will be used in the preprocessing step. The second purpose is to select the important feature in order to investigate the factor and learning behavior which is highly influenced on the academic performance.

The first task of FS using in this thesis to improve the performance of proposed models by enhancing the quality of data, filtering important and relevant features. In dealing with data, the quality and information of data are very important. FS is an integral technique in statistics, pattern recognition, machine learning, and data mining. Overfitting is a problem in supervised learning that lowers the accuracy of algorithms due to redundant and irrelevant features. Irrelevant or less informative variables not only provide no more information but also disturb the performance of the model, which is called noisy features. To solve this problem, we qualify the data by choosing the optimal subset of data containing the most relevant features using FS technique. It is defined as a process of reducing the size of a dataset by eliminating some of the irrelevant features;

it improves classifier performance and enhances clarity of results [122]. FS has also been defined as as a process of selecting the smallest size subset of features that achieved the best possible classifier performance [123]. Therefore, the objective of feature selection is to determine a subset of features, also called attributes, or independent variables, that will enable a classifier to perform optimally [124]. Findings have shown that classifier performance could be increased by eliminating irrelevant or redundant features in the list [125]. These early studies establish the fact that feature selection is an importance step in EDM.

FS method, therefore, determines the most predictive minimum number of attributes that will be used to predict the target attribute. The process improves the effectiveness of classifier training and enhances the prediction performance, in addition to eliminating difficulties of understanding the results. For example, an improved classifier accuracy and simplification of results was achieved when a small dataset was used after feature selection [126]. Further, the study reported improved effectiveness in terms of computer resource saving. Other identified advantages of feature selection include reducing the classifier training time, enhancing the ability for the classifier to generalize, and to facilitate a greater comprehension of the area of study [127]. Therefore, it is important to use only the features that are indicative of the target classes [128].

The second task is to rank the most important features. A crucial concept in the education field is extracting useful information and learning the behaviors of students in the learning process . For educational institutions, whose aim is to contribute to improvement of quality of education, the success of creating human capital is the subject of a constant analysis whose effects are long-term investment in an educational process that will result in other necessary products. Therefore, the quality of education at the level of primary and secondary schools represents one of the most important factors of forming a future full member of the community. Regarding, the quality of education is one of the most important carriers of continuity and intergenerational exchange of

basic principles and social values, but also actuator of necessary review and adaptation of those principles and values to the new circumstances. Because of this, relevant data and information are gathered regularly, discussed on the certain bodies and certain measures are taken in order to maintain quality. The quality of educational institutions involves providing services, which mostly meets the needs of students, teaching stuff and other participants of educational system.

FE is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. FE is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set. FE method is used to remove the high noises from samples from the collected dataset in the preprocessing step. Additionally, it can reduce the high dimensional data to low dimensional one which remains the important characteristics of data. More importantly, applications of EF methods enhance the equality of importance of features that effectively improve the performance of classification.

3.5.1 Filter-Based Feature Selection

Feature selection (FS) techniques aim to select an optimal subset of features. FS can be classified into three categories: embedded approaches, wrapper approaches, and filter approaches [129].

Filters are a technique that performs subset selection independent of the classifiers; they rate the predictive ability of the feature subset using the characteristics of the data [130]. Filter algorithms rank the features in order of their predictive ability. Filters are preferred because they have the following advantages: they take a much shorter time to determine subsets as they utilize less computer resources; they can be used with increased volumes of data; and they are not attached to any classifier [131]. Their

shortcoming is the assumption that features are separate entities and the fact that they operate independently of the classifiers. Filters have been identified as having an ability to operate using two approaches: features can be ranked using some predictive scale - *ranking method* or, they can be determined by the process of maximizing a determined cost function *space search method* [132]. Ranking is the more common approach in which, the optimal subset is achieved by picking the high ranked features and eliminating the ones that are low in the rank. The ranking approach is in four steps: 1) determining a scoring function for ranking the features from the most predictive to the least predictive; 2) approximate the *statistical significance* scores; 3) picking out the highly predictive features; and 4) validate the determined feature subset [133].

The popularly used techniques of filter feature selection methods are correlation based, Chi-square, gain ratio, information gain, relief, and symmetric uncertainty are mentioned in the studies [44][45][46][47][48][132][134]. The dataset used consisted of higher secondary school students in India. The initial set had 43 features. The features were ranked using each of the six techniques, in order of their predictive ability, from the most predictive to the least predictive. Different techniques achieved different ranking.

Information Gain Feature Evaluator

Information Gain (IG) is an important measurement used for ranking informative features. It explains how well a given attribute separated the instances with respect to their target classes [134][135]. IG uses the concept based on Shannon's informative entropy to split the instances. The Shannon informative entropy determines the impurity of example dataset D with different class c_i, express as in this equation:

$$I(D) = -\sum_{i=1}^{c} P(c_i)\log_2 P(c_i), \tag{3.43}$$

where c is the number of target classes, $P(c_i)$ is the proportion of examples in D with respect to class c_i. For $D = \{D_1, D_2, ..., D_m\}$, the m partitions of D, the information of

attribute A in example dataset D can be calculated in summation rule below.

$$I(D|A) = \sum_{i=1}^{m} \frac{|D_i|}{|D|} \times I(D_i). \tag{3.44}$$

The information of attribute A in dataset D is denoted as

$$IG(D, A) = I(D) - I(D|A) \tag{3.45}$$

Symmetrical Uncertainty Feature Evaluator

Symmetrical Uncertainty (SU) is one of the leading feature selection techniques [136][137]. SU determines the correlation between a feature and target variable using entropy and information gain theory. The calculating can be done via equation

$$SU(A, D) = \frac{IG(D, A)}{I(D) + I(A)}, \tag{3.46}$$

where $I(D)$ and $I(A)$ are entropies of based on probability of class associated with the example set D and attribute A , respectively; $IG(D, A)$is the information gain as shown in equation (3.45).

Mutual Information Feature Evaluator

Mutual information (MI) of two random variables or features is a measure quantifies the dependence measurement between those variables [138]. It is a symmetric measurement such that $I(X, Y) = I(Y, X)$ that can recognize non-linear relationships between variable. MI of two discrete variables and can be described as:

$$I(X, Y) = \sum_{x,y} P_{x,y}(X, Y) \times \log \left(\frac{P_{xy}(X, Y)}{P_x(X) \times P_y(Y)} \right), \tag{3.47}$$

where $P_x(X)$ and $P_y(Y)$ are marginal probability such that

$$P_x(X) = \sum_{y} P_{xy}(X, Y) \tag{3.48}$$

$$P_y(Y) = \sum_x P_{xy}(X, Y). \tag{3.49}$$

If the variables X and Y are independent, then the joint probability is separable; that is $P_{xy}(X, Y) = P_x(X) \times P_y(Y)$ and $I(X, Y) = 0$.

Chi-Square Based Feature Evaluator

Chi-Square (CHI) is a statistical technique used to determine the dependency of each input variable to the output variable. CHI algorithm can implement the comparison tests of independence [134][135][139]. For feature selection, CHI was used to assess the worth of a feature by calculating the chi-squared score of the classes, to obtain the ranking list of all features. Discretization was employed for the numeric attributes (making them discrete), in order to use the chi-squared statistic to find inconsistencies in the data [140]. The test used the idea of chi-square scores of the classes to get the ranking list of all attributes. The list of order informative features can be obtained through the following equation:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(n_{ij} - \varepsilon_{ij})^2}{\varepsilon_{ij}}, \tag{3.50}$$

where l is the number of class or determined intervals of a particular feature; c is the number of classes in target variable; n_{ij} is the observed (actual) frequency of sample of i^{th} interval and j^{th} class; ε_{ij} indicate the expected frequency of n_{ij}.

John et al. [141] has presented the definitions of *Strong Relevance* and *Weak Relevance* by considering the correlations among feature samples. Hall and Smith [142] formulated a measure of *Goodness of feature* as follows: "*Good feature subsets contain features highly correlated (predictive of) with the class, yet uncorrelated with (not predictive of) each other.*"

3.5.2 Feature Extraction Using Principal Component Analysis

The process of feature extraction is useful when you need to reduce the number of resources needed for processing without losing important or relevant information. Feature extraction can reduce the amount of redundant data for noisy data for a given analysis and improve the performance of proposed algorithm. Feature extraction can be either a linear or non-linear transformation of the original feature space. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are the most commonly used techniques for feature extraction. PCA is an unsupervised technique which is intended for feature extraction, while LDA is supervised. In overall conclusion is that in most applications, PCA outperforms LDA and, also, that PCA is less sensitive to different training data sets [143].

PCA is one of the top dimensionality reduction algorithms [144]. In PCA procedure, the data is firstly transformed to standardized data with zero mean. The idea behind getting the principle components is the covariance matrix is computed, and the result is used to calculate eigenvector and eigenvalues. The eigenvector which produce the highest eigenvalue is chosen as the principle component of new data which show the most significant relationship between data set attributes. This subsection, we discuss the powerful statistical methods of principle component analysis as an effective dimensional reduction method using the idea of linear algebra.

The idea of PCA is to preserve the maximum variance after transformation of the original features into new features. The new features are also referred to as "principal components" or "factors." Some factors carry more variance than other, but if we limit the total variance preserved after such a transformation to some portion of the original variance, we can generally keep a smaller number of features. PCA performs this reduction of dimensionality by determining the covariance matrix. After PCA transformation in a d-dimensional feature space the m $(m < d)$ largest eigenvalues of the $d \times d$ covariance

matrix are preserved. That is, the uncorrelated m projections in the original feature space with the largest variances are selected as the new features, thus the dimensionality reduces from d to m [145]. As explained in PCA, m uncorrelated linear projections are selected as the extracted features.

These derived variables, called principal component, as linear combination of observed variables. The purpose is to transform original dataset $X = (x_{ij})_{n\times l}$ of l dimension consisting of possibly correlated features to a new dataset Z of lower dimension m $(l \geq m)$, where Z is the principal component of X i.e $Z = PC(X)$, consisting linear uncorrelated features Suppose we have a set of observation $X_1, X_2, \cdots, X_n$, the principal component i which weighted combination of the n observed variables is denoted as

$$PC_i = a_{in}X_1 + a_{i2}X_2 + ... + a_{in}X_n, \quad (i = 1, 2, .., m), \tag{3.51}$$

where $a_i = (a_{i1}, a_{i2}, ..., a_{in})$ is the coefficients that need to be found. For any $i \neq j$, the coefficient vector a_i is perpendicular to vector a_j. If $Var(PC_1)$ is the largest then it is called the first principal component. In general, PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the coordinate, and so on. The procedure of calculating for principal components is summarized as follow:

(i) *Find Covariance Matrix*: After computing for mean, variance, and covariance, we can obtain covariance matrix S, which the entry s_{ij} , $i \neq j$, is the covariance between the i^{th} and j^{th} variables, and diagonal s_{ii} is the variance of i^{th} variables.

(ii) *Compute Eigenvalues and Eigenvectors*: The core of a PCA is the eigenvector and eigenvalues of the covariance matrix. The eigenvectors will determine the directions (coefficients) of the new feature space while the eigenvalues determine the magnitude.

(iii) *Obtain orthonormal eigenvectors*: By mean of linear algebra above, we can obtain the nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq ...\lambda_m > 0$ with corresponding orthonormal eigenvectors $v_1, v_2, ..., v_m$. The eigenvectors are called the principal components of the dataset.

PCA is less sensitive to different datasets than other holistic methods, so is mostly widely used technique as best feature reduction. The goal of PCA is to replace a large number of correlated variables with a smaller number of uncorrelated variables while capturing as much information in the original variables as possible. The merit of PCA is to improve performance of classification models and reduce computational cost.

CHAPTER 4

METHODOLOGY FOR ACADEMIC PERFORMANCE PREDICTION MODELING

4.1 Introduction

The proposition of this research is that Educational Data Mining (EDM) models could be developed for academic performance prediction to poor-performing students in the purpose of intervention and improvement. In order to get a better academic performance prediction models, effective and useful applicability, the data mining process called Cross-Industry Standard Process for Data Mining (CRISP-DM) is observed [146]. CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. CRISP-DM is a data mining process that has been used in the field of EDM [147][148].

CRISP-DM comprised of six main steps is illustrated in Figure 4.1. The six main steps are: (i) Domain Understanding, (ii) Data Understanding, (iii) Data Preparation, (iv) Data Mining: Modeling, (v) Evaluation, and (vi) Deployment: Using Discovered Knowledge.

Domain understanding: This initial phase focuses on understanding the research objectives and requirements from a domain perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. Only understanding the problem clearly, then we can convert it into a

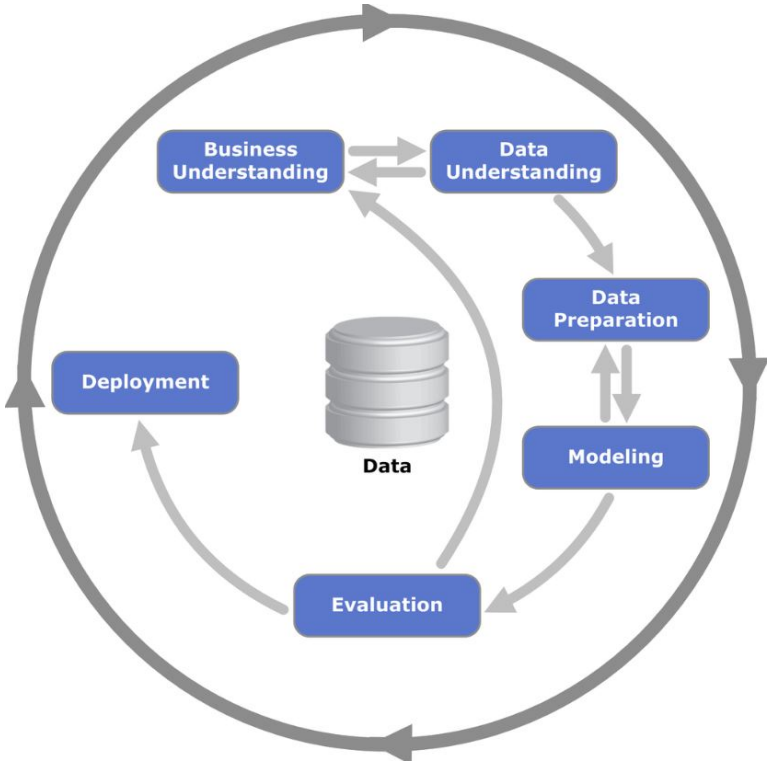


Figure 4.1: Outline of CRISP-DM process (adapted from [146])

well-defined analytics problem and can lay out a brilliant strategy to solve it.

Data Understanding: Data understanding phase focuses on collecting the data, describing and exploring the data. This step is to know what can be expected and achieved from the data. It checks the quality of the data, in several terms, such as data completeness, values distributions, data governance compliance. At this point data usefulness in terms of meeting the desired goal is also confirmed. In the context of this step in this research, we observed all variables related and influence to academic performance.

Data Preparation: The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modeling tool(s)] from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as cleaning and transformation of data to be in the right format for modeling tools.

Data Mining or Modeling: Selection of modeling technique is the very first step to take. This step entails selecting the methods (classification, regression, or clustering) to be used for knowledge generation and applying those methods to the data. All the models are then assessed to make sure that they fall in line with the business initiatives.

Model Evaluation: At this stage of CRISP-DM, we have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached. In the concept of data mining techniques, evaluation metrics are used to measure the performance of the used models.

Model Deployment: This stage entails putting the discovered knowledge to use by incorporating it into a performance system. It could also involve just documenting the knowledge and passing it to the interested stakeholders. The aforementioned steps have been adopted in our study to generate a general research framework as discussed next.

4.2 Domain Understanding

Understanding the domain, namely poor academic performance, was the initial stage of the EDM process. The domain understanding as been discuss in Chapter 1. In the study, the background, history, and problems in educational settings are studied. At this point the goal of the study was put into focus: to develop an intervention-level classifier model that predicts whether a student will require what levels of intervention to achieve passing marks in a high secondary school exit examination. Clearly this is a classification problem; classification techniques of EDM were employed.

4.3 Data Understanding

Improvement of student performance and enhancement in quality of education are of utmost importance for all educational institutions. Performance of students in high schools is believed to be the basic background for higher education. The statement of problem, some of related factors or features that affect students performance is briefly stated in Chapter 1. Investigation on those factors and other related factors (as shown in Figure 4.2) are treated as information (data) used for predicting high school student performance.

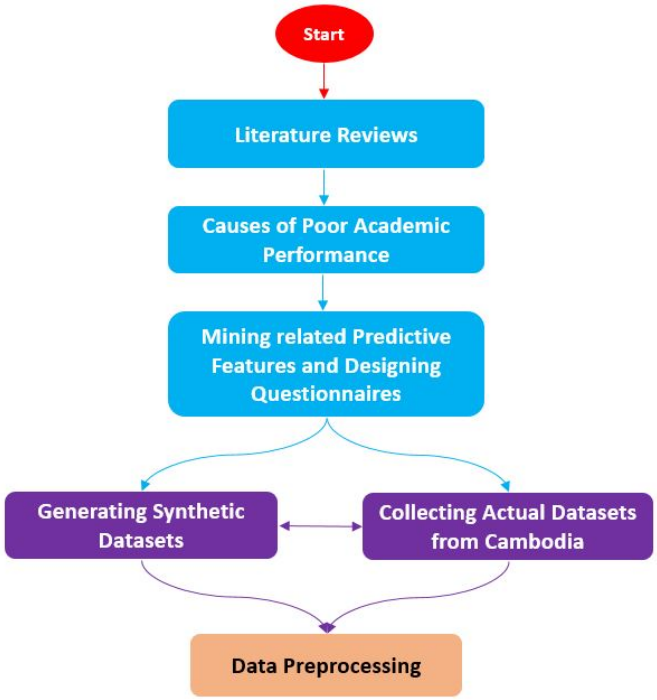


Figure 4.2: Systematic steps in data collection and preprocessing

An effective way to assist the success of students and boost the quality of education is to follow up on student performance, identify factors affecting students' achievements and then try to rectify any weakness in learning behaviors. There is an ever growing need to improve the performance of students by various institutions, but students' academic

performance hinges on diverse factors such as socio-economic, personal, psychological, family background, and other environmental variables. Diverse factors influenced a student's achievement in high school education, but are mainly considered to be school factors, domestics (or home) factors, and individual (student) factors.

4.3.1 The Related Factors Affecting Student Performance

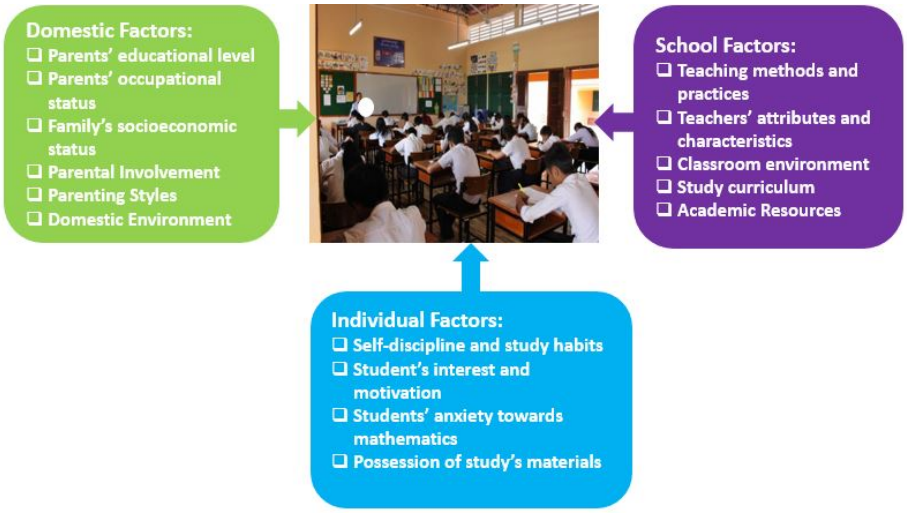


Figure 4.3: Factors affecting student performance

Since for the case of adolescence, the cognitive and non-cognitive factors are very important that affecting their daily life and performance, that's why we try to observe several external factors that affect their academic performance. From literature reviews in Chapter 1 and Chapter 2 plus this Chapter, there are various external and internal factors that affect academic performance and we categorized it into three main factors: Domestic or Home Factors, Individual or Student Factors, and School Factors, as illustrated in Figure 4.3. Moreover, most of the used features existing in the educational databases cannot describe enough of factors that affect the student performance. The adolescence performance are easily affected from external factors. Our purpose in this study is to observe all the internal and external that affect student performance. Hence, we try to

observe all the factors that affect their achievement in school. Once, we can effectively predict their future performance or grade levels, then we can give recipe to improve the performance. We collect a dataset of descriptive features obtained from three main factors, domestic features, individual features, and schooling features, which include home background, academic, and attitudinal information describing each student.

Domestic Factors

It is now well recognized that student and family background characteristics consistently explain a large proportion of the variance student outcomes [149][150][151][152][153]. These characteristics include student gender, socio-economic background, and aspiration and belief about learning. Parenting styles play a central role on the better and effective academic performance of adolescents irrespective of the level of education of education of adolescents [154]. These factors are important because students spent most of their life with their parents, the study [155] demonstrated that parent's educational inspiration for their children probed to be strongly related to students' academic growth.

Family, especially parents, has direct role in the social and academic life of adolescents and young children [156]. Adolescents, who gain family support from parents, are successful throughout in their social and academic life [157]. It is also evident from past research that home and communities have influencing role on the academic performance of adolescents regardless of the level of education of adolescents [158]. Strage and Brandt found that parenting supportiveness, demandingness and autonomy granting have positive relationship with the academic performance of students [159]. The domestic factors consist of several factors that influence the success of students' performance. These characteristics include parents' educational level, parents' occupations, socio-economics status, parental involvement, and parenting styles.

Individual or Student Factors

There are several factors of student attitudes that influence students' achievement in mathematics [160]. Student personal attitudes affect toward mathematics achievement are such as self-regulated, student's interest and motivation, mathematics anxiety, and students possession. Extensive research has been conducted that established a causal relationship between behavioral issues and academic achievement. There are two aspects of children's functioning, social skills and problem behaviors, that mediate the relationship between preschool self-regulation and literacy and math achievement [161]. Self-regulation or self-discipline was found to be an important factors that relates directly to academic achievement [162]. Student's interest and motivation have singnificant impace on the achievement of students [163][164][165][166].

As found from the experiment, mathematics' anxiety is another factor that appears to affect students' mathematics achievement. According to Ashcraft [167] mathematics anxiety is all about having the feeling of tension and apprehension, or fear that interferes with one's performance in mathematics. Many studies done on mathematics anxiety consider that there is relationship between mathematics anxiety and mathematics achievement of students, as shown, for example, [168][169]. These studies have also outlines clearly the importance of the possible effects that mathematics anxiety have on mathematics performance.

School Factors

The main instructions related to characteristics of the classroom are the interactions that take place between the teacher, students, and the curriculum. Hiebert and Grouws [170] have noted that there are some features of teaching that can be more effectively adopted to enhance the efficiency of skills for the purpose of boosting students' conceptual understanding. The findings from the studies have revealed that a review of the instructions that stresses on the development of concepts can facilitate both skill learning

and theoretical understanding, besides, when education to promote the struggle with mathematics, the students' understanding will increases a similar observation has been noted by Stevenson [171], that when students are exposed to rigorous mathematical content, their achievement will tend to improve.

In the recent decades, some extensive literature has examined the effective consequences of education in many countries [18][172][173]. Researchers in mathematics education have come up with many factors which demonstrate some significant relationships with mathematics achievement. These factors, among which are the teaching method and teaching practices [174][175], teachers' attribution and their accompanying characteristics [176], and the classroom climate [176][177]. Moreover, according to Adeogun and Osifila [178], there has been a significant impact of inadequate educational resources on students' academic performance. They have emphasised that teachers' attributions, teaching methods and teaching practices are the most important factors that leave an impact on student learning. However, current discussions concern whether or not, and the degree to which the teachers can make a difference in student learning relative to other factors that is expected to influence student learning [179]. In addition, the present talks also dwell into the specific elements of the teaching that can be logically and causally correlated to student achievement [180]. Teachers' attributes and their demonstration of behaviors in the classroom, contribute to the learning environment of the students, which in turn have an effect on the achievement of the students. There is a common hypothesis establishing the association between the experience of teachers and student achievement, whereby students who are taught by the most experienced teachers tend to achieve better [181]. According to [165][175][176][178], there are significant impact of inadequate educational resource on student performance. Several studies have indicated school-related factors as the most significant factors affecting the performance of students. These factors include classroom's environment, teaching methods and practices, teachers' attitudes and characteristics, and academic resources.

Table 4.1: Factors Affecting Student Performance

Factors	Sub Factors	Feature Code	Description	Type
Domestics Factors	Parents' education	PEDU1	Father's educational level	Nominal
		PEDU2	Mother's educational level	Nominal
	Parents' occupation	POCC1	Father's occupational status	Nominal
		POCC2	Mother's occupational status	Nominal
	Family's income	PSES	family's socioeconomic level	Ordinal
	Parent's involvement	PI1	Parents' attention to students' attitude	Ordinal
		PI2	Parents' time and money spending	Ordinal
		PI3	Parents' involvement as educator	Ordinal
	Parenting styles	PS1	Parents' feeling responsive and need	Ordinal
		PS2	Parents' respond to children's attributes	Ordinal
		PS3	Parents' encouragement	Ordinal
		PS4	Parents' compliment	Ordinal
	Domestic environment	DE1	Domestic environment for study	Ordinal
		DE2	Distance from home to school	Ordinal
Individual Factors	Self discipline	SELD1	Number of hours for self-study	Nominal
		SELD2	Number of hours for private study	Nominal
		SELD3	Frequency of doing math's homework	Ordinal
		SELD4	Frequency of absence in math class	Ordinal
		SELD5	Frequency of preparing for exams	Ordinal
	Students' interest and Motivation	SIM1	students' interest in study	Ordinal
		SIM2	students' enjoyment in class	Ordinal
		SIM3	students' attention in math class	Ordinal
		SIM4	students' motivation to succeed in study	Ordinal
	Student's anxiety toward class	ANXI1	Students' anxiety in math class	Ordinal
		ANXI2	Students' nervousness in exam	Ordinal
		ANXI3	Students' feeling helpless in math	Ordinal
	Students' possession	POSS1	Internet's use at home	Binary
		POSS2	Possession of computer	Binary
		POSS3	Student's study desk at home	Binary
School Factors	Class environment	CENV	Classroom environment	Ordinal
	Curriculum	CU1	Contents' language of math	Nominal
		CU2	Class session	Nominal
	Teaching Methods and Practice	TMP1	Teachers mastering in math class	Ordinal
		TMP2	Teachers' absence in math class	Ordinal
		TMP3	Teaching methods in math class	Ordinal
		TMP4	Teacher' involving in education's content	Ordinal
	Teachers' Attribute and Characteristic	TAC1	Math teachers' ability	Ordinal
		TAC2	Teachers' encouragement to students	Ordinal
		TAC3	Math teachers' connection with student	Ordinal
		TAC4	Math teachers' help	Ordinal
	Academic Resource	ARES1	Adequate use of math teachers	Nominal
		ARES2	Adequate use of classroom	Nominal
		ARES3	Adequate use of math handout	Nominal

4.3.2 The Target Variable: Student Learning Performance Levels

It is hard to measure the real ability of students. Grade does not always reflect the learning outcomes that the student have or have not achieved. However, for measuring and managerial purposes, it is a must to predefined grades as A, B, C, and D or performance levels as $Excellent$, $Good$, $Average$, and $Poor$.

In measuring the performance levels of students in mathematics, we discretized the performance of student based on their obtained score in Semester I. Base on the generalization of score category, the excellent learners are the group of students who obtain the score with equal to 90 % and above of the total score. The second group means good learners who obtain 75% to less than 90%. The moderate learners are those who get from 60% to 75% of the total score, while the poor (slow) learners are the at-risk group that get the score below 60%. The target variable that is the final score of the each student in Semester I as shown in Table 4.2.

Table 4.2: The Target Variables

N	Performance Levels	Score-based Category	Level of Intervention Needs
1	Excellent learner	90% and above	No risk-no need intervention
2	Good learner	75% to less than 90%	Low risk
3	Moderate learner	60% to less than 75%	Medium risk
4	Poor learner	less than 60%	High risk-need high intervention

4.3.3 Synthetic Datasets

Secondary education is an compulsory degree to build sufficient background of knowledge for higher education. Provision of strong background in high schools reply to any policies in the development of educational institutions. Up to this present, the results of students' learning are still not satisfactory in various countries. Many researches hold on the task of studying the dominant factors and learning behaviors of students towards the achievement of students in secondary schools. Mohamed et al. [18][28][57] and Caro et

al. [150] studied the relationships and causal structure among effective variables, such as home background, socioeconomic status, self-confidence, attitude towards learning performance, mathematics anxiety, and academic achievement for students. Hardi [31] investigated the dominant factors of the teacher's role as a motivator of a student's interest and motivation. Studies decompose the structure and inter-relation between explanatory factors and students' achievement.

For the merits of testing for the effective models, generalization of datasets, and limitation of cost of collecting actual datasets with big samples size, we generated additional synthetic datasets to use in this study. The knowledge to generated the generalized, useful, and effective synthetic datasets was adopted via the knowledge of many literature in causal and correlation literature which is based on the actual datasets. The traditional statistical techniques for causal structure is combined to generate the synthetic datasets.

Data Generating

To reach some specific needs and more generalized characteristics of the data, which are hard to find in the existing or original datasets, we generate the synthetic data with description shown in Table 4.1 and Table 4.2. The data were generated based on causal structures in the Related Work to Factors Affecting Student Performance [18][28][29][31] and other observed information concerning factors affect academic performance mentioned in Chapter 1, Chapter 2, and this Chapter (in the earlier section). We use Data Synthetsizers and simulateData in *synthpop* and *lavaan* packages in R to generate our synthetic data. In investigating the performance of prediction models, datasets with the same characteristics, but different sizes were used for implementation. Here, the data was generated based on structure of relation show in the Figure 4.4.

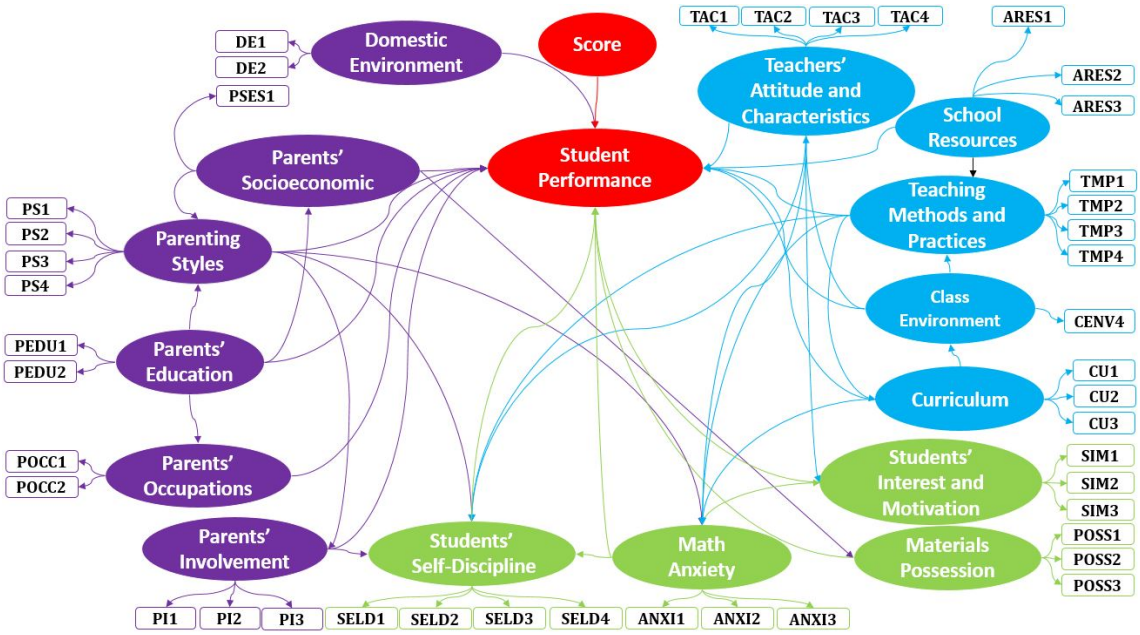


Figure 4.4: Structural model for generating synthetic datasets

The Merits of Synthetic Datasets

There are two main reasons in using synthetic datasets in this study. First the synthetic datasets are used in the comparative study to observed the performance of algorithms of educational data mining that are used in this study. To reach some specific needs and more generalized characteristics of the data, which are hard to find in the actual or original datasets, the generated datasets were proposed.

Secondly, the synthetic datasets will be used in addition to the actual datasets to confirm the robustness and effectiveness of the proposed models. The actual data which is collected from various high schools in Cambodia. In our dissertation, we proposed the hybrid machine models and optimized deep belief network more as optimal predictive model in predicting student performance in secondary schools. Here, we implement our models using the actual datasets and synthetic datasets to confirm the effectiveness and robustness of the proposed prediction models.

4.3.4 Actual Dataset

Our target dataset are from information of high school students in Cambodia. The scope of this study is to focus on external and internal factors that affect the student performance, we then exclude the information in demographic background which is invariant and unchangeable. As shown in Table 4.1, the total of 43 predictive features are mainly from three main factors contain the various sub-factors with shorthand variables notation listed on its left-hand side. The questionnaires were prepared with references, assistance, and guidance from:

(i) Hundreds of review literature,

(ii) Teachers and school directors from diverse schools,

(iii) Staff from the department of research (MOEYS: Ministry of Education Youth and Sport, Cambodia),

(iv) Senior researchers in the education field.

The data type used in questionnaire is categorical data types including binary, ordinal, and nominal data. The nominal types are for features that describes about parents' educational level, parents' occupation, curriculum, and adequate use of academic resource. The binary types using {Yes, No} responds regarding to existence of possession material for their study. And the rest but the most is ordinal type that use 5-point Likert scale as measurement.

Why Using 5-point Likert Scales? (Questionnaires)

Over the years, numerous methods have been used to measure character and personality traits Likert [182]. A variety of methods are available to assist evaluators in gathering data. One of those methods involves the use of a scale. The Likert-type scale involves a series of statements that respondents may choose from in order to rate their responses

to evaluative questions [183]. The difficulty of measuring attitudes, character, and personality traits lies in the procedure for transferring these qualities into a quantitative measure for data analysis purposes. Likert [182] responded by developing a procedure for measuring attitudinal scales. The original Likert scale used a series of questions with five response alternatives. He combined the responses from the series of questions to create an attitudinal measurement scale. His data analysis was based on the composite score from the series of questions that represented the attitudinal scale. With Likert data, agreement with a statements is measured at an ordinal level.

Table 4.3: The 5-point Likert Scale Measurement

Respond Set	1	2	3	4	5
Agreement	Strongly disagree	Disagree	Neural	Agree	Strongly agree
Frequency	Never	Rarely	Sometime	Often	Always
Quality	Very poor	Poor	Fair	Good	Very good
Satisfaction	Not all all satis.	Slightly satis.	Fairly satis.	Very satis.	Extremely satis.

The questionnaire (see APPENDIX) was mostly made with 5-point Likert scale answers. The Likert scale is a popular measurement scale using social study and human behaviors. There are several use of Likert scale such as 4-point Likert scale, 5-point Likert scale, 7-point Likert scale, and 10-point Likert scale. However the variables of ordinal types in our research was made using 5-point Likert scale. The Likert scale from 1 to 5 have a middle-point. Respondents also find it difficult to define their views on scales with values greater than seven. This means that if you offer more than 7 answer choices, respondents are likely to choose their answers at random and thus, their data will be insignificant. It is recommended using five-point scales for single scales (the presence or absence of a selected attribute is referred) and seven-point bipolar scales (Includes two sensations or different attributes related to each other). David Moller conducted a study on using Likert scale with Children with better understanding with 5-point Likert scale [184]. In our questionnaire, we encode the respondent of 5-point Liket scale with number

$\{1, 2, 3, 4, 5\}$ represent the answers described in Table 4.3.

Data Collection Process

The data collection was done by giving questionnaires for filling. For being more convenient in getting questionnaire to both contributors and receivers, we deliver questionnaire for students to fill via google form. However, the students can either filling the questionnaire using paper-based or via Google Forms of the given links through their phones or computers.

Once the questionnaire design was done, we then confirm our target respondents. Our target data is from high school students in Cambodia. The response was obtained from:

(i) Students from various high schools in Cambodia,

(ii) Teachers in corresponding high schools,

(iii) Study records (final semester score) from administrative offices.

Finally, after the data collection process, we obtain a dataset of 1204 records. The samples of obtained dataset is shown in the Figure 4.5.

Sort fields: Data source order — Show aliases — Show hidden fields — 1,000 rows

Pedu1	Pedu2	Pocc1	Pocc2	Pses1	PI1	PI2	PI3	PS1	PS2	PS3	PS4	DE1	DE2	Seld1	Seld2
Lower_Secondary	Elementary	Government_O...	Housewife	2	3	4	4	4	4	4	3	4	Moderate	2 to 4	< 2
Lower_Secondary	Elementary	Business	Business	2	4	4	4	5	5	3	5	4	Moderate	< 2	> 4
Upper_Secondary	Lower_Secondary	Business	Business	2	3	5	5	3	3	2	3	5	Short	2 to 4	> 4
Graduate	Undergraduate	Government_O...	Government_O...	3	3	4	5	3	5	3	4	4	Moderate	< 2	> 4
Lower_Secondary	Elementary	Farmer	Housewife	2	4	5	4	4	4	3	3	3	Short	< 2	Never
Upper_Secondary	Lower_Secondary	Government_O...	Farmer	3	4	5	4	3	3	2	3	3	Moderate	Never	> 4
Elementary	Elementary	Others	Others	2	4	5	5	2	5	2	5	5	Moderate	< 2	< 2
Elementary	Elementary	Business	Business	2	5	5	5	5	5	5	5	5	Moderate	< 2	> 4
Upper_Secondary	Upper_Secondary	Business	Business	3	5	5	5	4	4	4	5	4	Moderate	2 to 4	2 to 4
Undergraduate	Undergraduate	Government_O...	Government_O...	3	5	5	5	4	4	5	5	5	Moderate	2 to 4	> 4
Lower_Secondary	Elementary	Others	Housewife	2	5	5	5	4	3	4	4	3	Short	< 2	> 4
Undergraduate	Undergraduate	Government_O...	Government_O...	3	4	4	5	4	3	2	3	3	Far	2 to 4	> 4
Undergraduate	Upper_Secondary	Government_O...	Farmer	3	5	4	5	4	4	5	5	4	Moderate	2 to 4	> 4
Lower_Secondary	Lower_Secondary	Elders	Government_O...	2	3	5	5	4	4	4	4	4	Moderate	< 2	> 4

Figure 4.5: Samples of educational dataset consist of predictive features that is used in this academic prediction problem

4.3.5 Evaluation on Utility of Synthetic Data

Synthetic or artificial dataset is generally used for studying problems of diverse applications. Generation of synthetic data is adopted for many reasons such as the privacy of original datasets, generalization of original datasets or specific needs that doesn't exist in the original data, and larger size of original datasets [185]. Synthetically generated datasets provide a way to retain the utility of original dataset, which tries to keep or preserve the overall properties and characteristics of the real-world datasets without revealing information about actual individual data samples [186]. The experiment in study utilized the artificial datasets as original/real dataset. Hence, this section provide the evaluation on the utility of our artificial datasets.

The utility of generated dataset needs to be evaluated at two levels (common methods). Firstly, we need to evaluate differences between the distribution of original attribute values and generated attributes values (Statistical Comparison), the descriptive statistics (distributions) of both datasets. Secondly, we need to evaluate the difference between the quality of estimation of a certain estimand for synthetic data and generated data (Overlap).

Statistical Comparison of Synthetic Data to Original Data

Good quality synthetic data would allow us to draw the same analytic conclusions as we can draw from real data [186][187]. Hence, the first step of our evaluation on validity and utility of the synthetic dataset is a statistical comparison of the synthesized and the original data. For each of the datasets, we performed the following steps.

Step 1: For each attribute, we generated a histogram/barplot visualising both the distribution of the real and the synthetic data.

Step 2: We computed the correlation coefficients and generated a heat map to visualise dependencies between attributes.

Figure 4.6: Distibution of all features of synthetic and real datasets

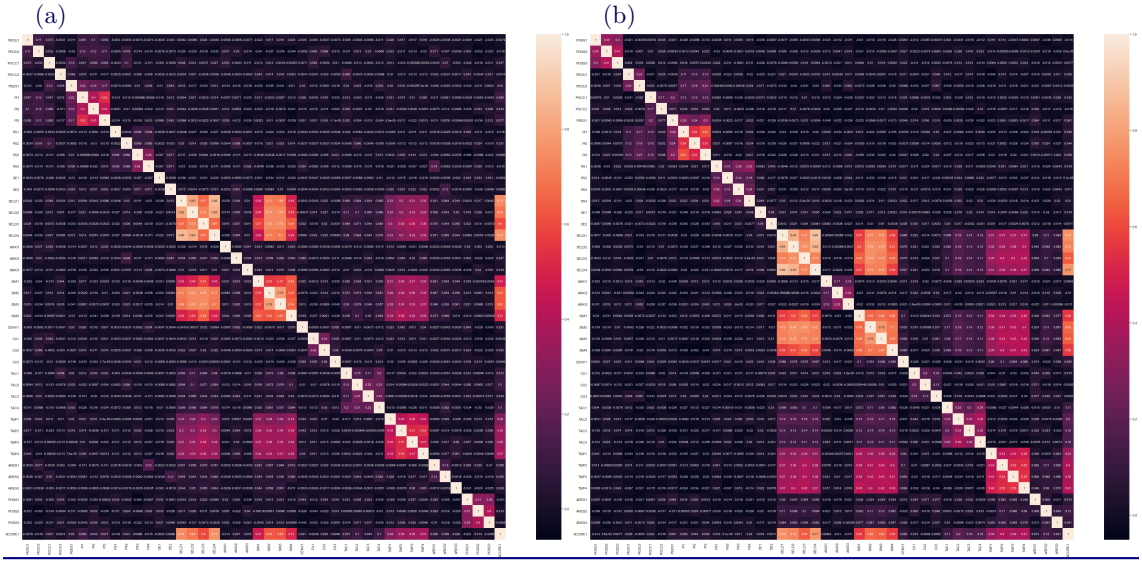


Figure 4.7: Heatmap of the (a) Synthetic Data, and (b) Original Data

Utility of synthetic data is deserved when it presents the characteristics of real-world data (the original data). Therefore, we require some kind of distance measure between datasets. We want to identify whether there are any individuals (i.e. rows) in the original dataset that also appear in the synthesised datasets. According to Figure 4.6, even the normalized frequency of index of each features are not exactly matched but the trend of synthetic data is generated identically to the original dataset. The heatmap in Figure 4.7 show the significant difference in the dependencies between features, compare the heatmap of both dataset. The figure indicates the similarity of the synthetic and original data imply that the synthetic data maintains the overall characteristics of original data.

Overlap

Karr [188] develop a mechanism based on **overlap** between confident interval to evaluate the utility and validity of synthetic data. We constructs a 95% confident interval. Let (L_s, U_s) be confidence interval for synthetically generate y and (L_o, U_o) be interval from original data. We compute intersection of these intervals denoted as (L_i, U_i). The overlap

utility measure is calculated as given in equation (4.1).

$$I = \frac{(U_i - L_i)}{2(U_o - L_o)} + \frac{(U_i - L_i)}{2(U_s - L_s)} \tag{4.1}$$

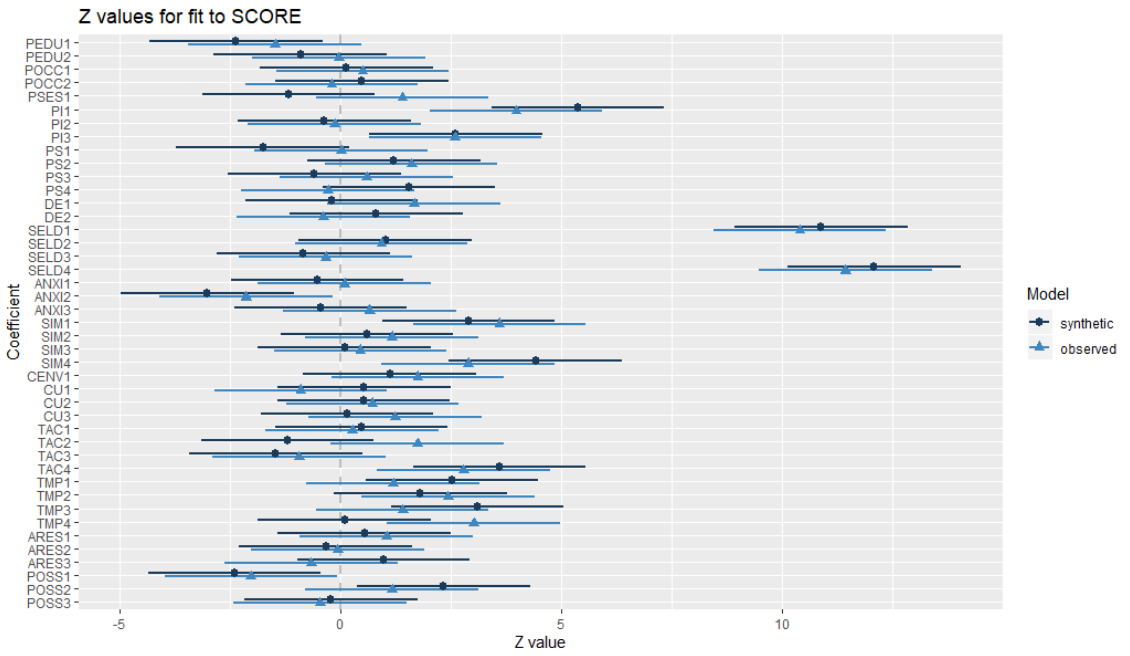


Figure 4.8: Comparison of confident interval (C.I) overlap indicate the utility of synthetic dataset.

The value of I is close to 1 confirms the utility and validity of synthetic data and $I = 0$ refers to the dissimilar confidence intervals. The utility evaluation methods was adopted in many studies to evaluate and validate the utility of synthetic datasets [189][190][191]. When we compare our models for generating dataset, we can see for that there's no significant difference between our synthetic and observed (original) data for the interaction ($p = 0.194$ with a mean C.I overlap of 0.79). This conclude the utility and effectiveness of the synthetic dataset.

4.4 Data Preprocessing

The best algorithm is useless unless your data is ready for the implementation. Coming up with features is difficult, time-consuming, and requires expert knowledge. Data preprocessing is an integral step data mining as the quality of data and the useful information that can be derived from it directly affects the ability of our model. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. The contribution of this step is not only to get data ready for executing but also improving the accuracy of algorithms.

4.4.1 Data Cleaning

Even though data collecting was done with careful concentration, there are still many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with bio-metrics and much more. Ignoring this is not a good choice since it may play an important role or when data get large it will lead to an inaccurate result. Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

It is important to remove or correct records with corrupted or invalid values from raw data, as well as remove records that are missing a large number of columns. Data cleansing or data cleaning deals with the task of identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or deleting the dirty or coarse data. In our dataset, the number of missing or invalid values is low, then we use the imputation method. The students questionnaire completion was done with missing some questions or input invalid number. In case of missing value in categorical variables such as in parents' education, and parents occupation, we replace the missing values with the cardinal value "Others". In the ordinal and binary variable, the modes of variables or

the most frequency-category value were replaced. For the output variable (score in final semester I), there were problem with missing value or outliers, we similarly ignore it and replace by the mean value.

4.4.2 Feature Encoding

Data transformation is the process of transforming or converting data from one format or structure into an executable form and not miss important information. In some stage of data type, the attributes need to be transformed or changed to numeric form for many reasons. In our proposed learning algorithms (PSEM, KNN, LVQ, LR, PCA) require the numeric type of attributes. More importantly, in the neural network models (MLP and DBN), the attributes have to be in the numeric format and normalized to achieved the best classification results. Hence, we apply the feature encoding techniques to obtain this purpose. The techniques concern with replacing values, label encoding, dummy encoding, and one-hot encoding.

4.4.3 Data Normalization

Normalization is considered one of the recommended preprocessing task that should be preprocessed the dataset to some kind of classification and prediction algorithms, such as neural network models (MLP, KNN and DBN). Those algorithms required to make the instances values within the specific range either [0,1] or [-1,1], since scaling to this ranges ten to give better results. In this case of these algorithms (MLP, KNN and DBN), Min-Max scaler was used as the normalization method.

4.4.4 Data Discretization

Data discretization is an important concept dealing with a large range of numerical value, when we want to classify it into an ordinal classes for the purpose of comfortable form in computation. In our study, for the purpose of predefining the class of student performance

levels, we converted the final semester score of students into ordinal classes to categorize the distinct levels of students (slow or poor learners, average learners, good learners, and excellent learners).

4.4.5 Dimensional Reduction

Dimensional reduction plays an important role in classification performance. Dimensional reduction is one of the effective techniques used in a large dataset in the preprocessing step. Two main categories of dimensional reduction approach are feature selection and feature extraction. In this research, the features selection and features extraction approaches were introduced to combined with the proposed models.

4.5 Data Mining or Modeling: Methods of EDM

In Chapter 2, we have observed or investigated a wide range of literature reviews, give a review of EDM researches, its purpose, used algorithms, and key findings. The popular and effective baseline methods are mentioned in Chapter 3. The modeling of academic performance prediction is studied in Chapter 3, with the comparative study. The optimal models are further developed in Chapter 6, Chapter 7, and Chapter 8.

4.6 Model Evaluation Metrics

The performance of each proposed models in analyzing and predicting student performance in high school can be evaluated from an analysis of a graphical confusion matrix. Without loss of generality, our output variable can be categorized into four ordinal categories as mention in Table 4.2. Table 4.4 shows the graphical confusion matrix that represents four classes of student performance levels in high school. The groups of poor-performing students are classified into four levels: high risk (HR), medium risk (MR), low risk (LR), and no risk (NR). The parameters below were calculated. In

understanding these metrics, we are denoting some terms like CP (Correct Prediction), E_{ij} (Error or Incorrect Prediction), G^a (Actual Grade), and G^p (Predicted Grade).

Table 4.4: Confusion Matrix

		Predicted Classes			
		HR	**MR**	**LR**	**NR**
Actual Classes	**HR**	CP_1	E_{12}	E_{13}	E_{14}
	MR	E_{21}	CP_2	E_{23}	E_{24}
	LR	E_{31}	E_{32}	CP_3	E_{34}
	NR	E_{41}	E_{42}	E_{43}	CP_4

4.6.1 Classification Accuracy

Classification Accuracy (ACC) is a common evaluation metric used in prediction and classification problem [192]. Accuracy was used to quantify the percentage of correct predictions which is not appropriate with imbalance classes. However, from our data collection, our predefined target classes are nearly balanced, that accuracy is the best classification metrics to use. Here, we want to evaluate the potential of our prediction model by measuring the percentage of correctly predicted levels of student performance. It is also known as the overall accuracy, or the proportion of correctly predicted high and low intervention to all the students. Hence, accuracy can be calculated by the fraction of total correctly predicted values to the total number of predicted values via equation below:

$$ACC = \frac{\sum CP_i}{\sum CP_i + \sum E_{ij}} \times 100\%. \tag{4.2}$$

4.6.2 Root Mean Square Error

Root mean square error (RMSE) is used to evaluate the goodness of our proposed models by measuring the variations in predicting correct values. It rates the performance of

prediction models in term of the size of errors that the model make [193]. RMSE has been found to be a more a appropriate metric for model performance compared to the popular mean absolute error metric [194]. It is sometime hard to reflect the actual ability of student performance. Therefore, we aim not only to predict the performance levels of students but also to estimate how much our prediction is close to their actual grades. Without loss of generality, we encoded these ordinal performance levels {HR, MR, LR, NR} as {1,2,3,4}, respectively. RMSE can be obtained from equation below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(G_i^a - G_i^p)^2}, \tag{4.3}$$

where $G^a \in \{1,2,3,4\}$ is the actual outcome and $G^p \in \{1,2,3,4\}$ is the predicted outcome, which {1,2,3,4} are represented for HR, MR, LR, and NR. The better RMSE is obtained from an almost diagonal matrix or a tri-diagonal confusion matrix. Unlike ACC, the smaller the RMSE, the better model. RMSE is a nonnegative metric, which RMSE equal to 0 shows the prediction model is perfect [195].

These two standard evaluation metrics are used to measure and rate the performance of the prediction models in determining the best prediction models and best dominant sets for student academic performance.

4.7 Model Deployment

The implication or applicability contribution of this work is the prediction system for academic institutions. By using the developed models proposed in this study, the optimal model are then subsequently integrated in to web-based application, termed as Academic Performance Prediction System (APPS). The development and implementation of APPS will be described in Chapter 9.

CHAPTER 5

COMPARATIVE STUDY OF EDUCATIONAL DATA MINING MODELS FOR STUDENT PERFORMANCE PREDICTION

5.1 Introduction

This chapter gives a comparative study of prediction models of educational data mining (EDM) to observe the optimal classifiers. The experiment is divided into two phases. Prior to the study of actual dataset, the first phase in the experiment is the comparative study of prediction models using synthetic datasets to generalize the performance of the prediction models. The second phase is the experiment using actual datasets, which were collected from high schools in Cambodia. In the context of data science or data analysis, when come to any problems of prediction modeling, there is no best model; there is only the right model for a particular problem. Therefore, the choice of a suitable model with the specific data is very important and it is an challenging task.

A single model may perform well in a particular problem, yet poor in another problem. That is the reason we come up with mentioning several literature reviews and figure out the popular approaches, and then make a comparative study of these approaches. In order to achieve the right model for our comparative study in this academic prediction problem, we propose three main consideration for selecting the models. The first consideration is

based on the state-of-art methods, which we have obtained in Chapter 2. The second consideration is based on investigation of the characteristics of those models, advantages, and disadvantages of those models in our problem. Finally, we apply the preliminary checking on those algorithms. This process is to give primary execution on our dataset and see what works and drop what does not work. This process is called spot-checking algorithms. Owing to these three main consideration, we obtained a set of EDM models and categorized it into three main categories or disciplines. Those are statistical analysis techniques, machine learning models, and deep learning models. The brief introduction on these three main EDM models are presented as follows.

5.2 The Comparative Prediction Models

5.2.1 Statistical Analysis Techniques

Statistical analysis is a component of data analytic, which is an integral tool in all research area. Structural Equation Modeling (SEM) is a popular statistical analysis technique used in education research. This technique is a combination of factor analysis and multiple regression analysis. It is preferred by many researchers since it was used to estimate the multiple and interrelated dependence in a single analysis. The traditional SEM was a causal model used to analyze the structural relationship between measured variables and latent constructs [62]. As mentioned in preliminary methods in Chapter 3, SEM comprises two models: a structural model and a measurement model [59]. The structural model is to show the relationships or paths among the constructs (or latent variables). The structural model presents the relations between latent variables and regressions of latent variables on the observed variable. The measurement model is built from the implicit or explicit models that relate a latent variable to its indicators. In addition, another term called the threshold model is introduced for the categorical response. A novel concept in this work is to modify the classic causal model SEM to be a predictive model, termed

predictive SEM (PSEM). By combining factor analysis and multiple regression analysis, we adapted our causal model SEM to be prediction model, PSEM.

5.2.2 Machine Learning Algorithms

Machine learning (ML) is defined as an automated process that extracts patterns from data [79][80]. To build the models used in predictive data analytic applications, we use supervised ML. Supervised ML techniques automatically learn a model of the relationship between descriptive features and the target variable. It is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of supervised ML is to build algorithms that can receive input data, and to use statistical analysis and mathematical concepts to predict an output while updating the output as new data become available. In the last decades, ML has been a paradigm shift for the education field in evaluating higher education tasks and extracting the hidden useful information and desired knowledge to support a decision process for the progress of students.

Several ML classifiers were summarized in Chapter 2 and the introductory knowledge of those models are mentioned in Chapter 3. The popularly used models are Multilayer Perceptrons (MLP), Sequential Minimal Optimization (SMO) of Support Vector Machine (SVM), Decision Tree, Logistic Regression (LR), and Random Forest (RF), k-Nearest Neighbor (KNN), and Ensemble methods etc. Those popular methods are introduced in this comparative study.

5.2.3 Deep Learning Algorithms

Deep learning (DL) is a booming subfield of machine learning that has solved increasingly complicated applications with increasing accuracy over time in artificial intelligence (AI) regime [196]. The Deep Belief Network (DBN) is a class of DL frameworks that were conceived by Geoffrey Hilton [118]. It is a multi-layer Belief Network that each layer is

Restricted Boltzmann Machine (RBM), and they are stacked with each others to construct DBN. A three-hidden-layer DBN contains pretraining that follows a greedy layer-wise learning procedure is used in this study. One layer is added on top of the network at each step, and every top layer is trained as an RBM using a Contrastive Divergence (CD) strategy [108]. After each RBM has been trained, the weights are clamped and a new layer is added to repeat the same procedure.

5.3 Comparative Study Using Synthetic Datasets

This section utilizes synthetic data for our experiment in order to observe the performance our chosen models. The first comparative study was carried out using three synthetic datasets.

5.3.1 Synthetic Datasets

To reach some specific needs and more generalized characteristics of the data, which are hard to find in the existing or original datasets, we generate three synthetic datasets to execute the performance of prediction models in this experiment. We generalized the characteristics of datasets based on various interrelation and coefficient of each variables. Three synthetic datasets are presented as DS1 (2000 samples) for average population of high school students in one town, DS2 (4000 samples) for average population of high school students in one district, and DS3 (10000 samples) as average high school student population in one province in Cambodia.

5.3.2 Experimental Results

We conduced a comparative analysis of three main categories of educational data mining models. These models are PSEM of statistical analysis methods, five superior classes of machine learning models: MLP, SMO, C5.0, LR, and RF, and with one of DL models,

that is DBN. We execute the performance of all the seven learning algorithms on the three datasets DS1, DS2 and DS3. Table 5.1 presents the accuracy (ACC) and root mean square error (RMSE) of the mentioned models with respect to the three datasets.

Table 5.1: Performance of the Proposed Prediction Models

		ACC			RMSE		
Classes	Models	DS1	DS2	DS3	DS1	DS2	DS3
Statistical Anaysis	PSEM	56.40	60.90	62.66	0.681	0.651	0.637
Machine Learning	MLP	78.81	86.92	88.28	0.458	0.357	0.341
	SMO	75.28	82.21	91.40	0.524	0.424	0.336
	C5.0	**79.84**	**88.37**	**93.42**	**0.449**	**0.345**	**0.254**
	LR	77.42	86.32	89.27	0.484	0.368	0.349
	RF	**80.58**	**90.10**	**93.52**	**0.435**	**0.314**	**0.252**
Deep Learning	DBN	75.48	87.20	92.01	0.494	0.356	0.301

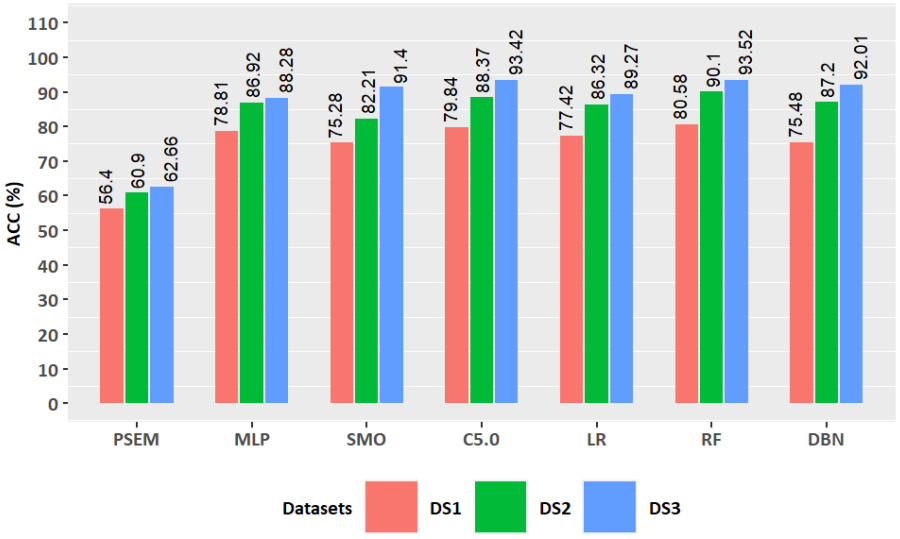


Figure 5.1: Accuracies corresponding to the three datasets

The performance with respect to accuracy of the proposed models can be presented more perfectly in Figure 5.1. Analysis of variance (ANOVA) with significance level $p = 0.05$ was used to differentiate the difference between all classifiers in term of accuracy.

The degree of comparison are Statistically Different (SD) and Not Statistically Different (NSD).

Table 5.2: Accuracy Comparison of Comparative Models on dataset DS1

	PSEM	MLP	SMO	C5.0	LR	RF	DBN
PSEM	-	SD	SD	SD	SD	SD	SD
MLP	-	-	SD	NSD	NSD	SD	SD
SMO	-	-	-	SD	SD	SD	NSD
C5.0	-	-	-	-	SD	NSD	SD
LR	-	-	-	-	-	SD	NSD
RF	-	-	-	-	-	-	SD
DBN	-	-	-	-	-	-	-

Table 5.3: Accuracy Comparison of Comparative Models on dataset DS2

	PSEM	MLP	SMO	C5.0	LR	RF	DBN
PSEM	-	SD	SD	SD	SD	SD	SD
MLP	-	-	SD	SD	NSD	SD	NSD
SMO	-	-	-	SD	SD	SD	SD
C5.0	-	-	-	-	SD	SD	SD
LR	-	-	-	-	-	SD	NSD
RF	-	-	-	-	-	-	SD
DBN	-	-	-	-	-	-	-

Table 5.4: Accuracy Comparison of Comparative Models on dataset DS3

	PSEM	MLP	SMO	C5.0	LR	RF	DBN
PSEM	-	SD	SD	SD	SD	SD	SD
MLP	-	-	SD	SD	SD	SD	NSD
SMO	-	-	-	SD	SD	SD	SD
C5.0	-	-	-	-	SD	NSD	SD
LR	-	-	-	-	-	SD	SD
RF	-	-	-	-	-	-	SD
DBN	-	-	-	-	-	-	-

From Table 5.1, the accuracies of all proposed models executed on dataset DS1 ranges from 56.40% to 80.58%. The accuracy of MLP to C5.0 and LR are not so much different.

The performance of SMO in classifying the student performance is similar to the DBN. Using ANOVA, we compare the person of the models on dataset DS1 in Table 5.2. The pairwise comparison of ANOVA concluded that the pairs of C5.0-MLP, LR-MLP and DBN-SMO, C5.0-RF, and DBN-LR are not statistically different. Simultaneously, C5.0 and RF algorithms are statistically different to other models and outperform the rest algorithms. We can conclude that RF algorithm is a superior classier and followed by the tree based C5.0.

Table 5.3 gives performance comparision base on ACC of models for dataset DS2, the pairs of DBN-MPL, LR-MLP, and DBN-LR are found to be not statistically different, while the rest are found to be statistically different. Howver, the ACC of DBN is comparatively better than MLP and LR. RF model generate the highest ACC of 90.10% followed by C5.0 and DBN.

Table 5.4 presents performance of the models on dataset DS3. In this case, the pairs of LR-MPL, and C5.0-RF are found to be not statistically different and outperform the rest models. DBN, C5.0, and RF were the top three models with accuracy of 92.01%, 93.42%, and 93.40%, respectively. DBN is the third-best model follow the two outperformed tree-based models. The RF model remain outperforms other models, which generates the highest ACC.

We aim not only to check how much we correctly predict the student performance, but also to estimate how much our prediction is close to their real ability. That is RMSE, is used to evaluate the goodness of fit of our proposed models. Without loss of generality, our output variable can be categorized in to four ordinal categories in which RMSE is a suitable evaluation metric. The performance based on RMSE of the proposed models is illustrated in Figure 5.2. In contrast to accuracy, the smaller value of RMSE, the better model is. The tree-based RF algorithm generates the smallest value of RMSE in all three datasets. With respect to this metric, RF produces the smallest error and followed by tree-based C5.0.

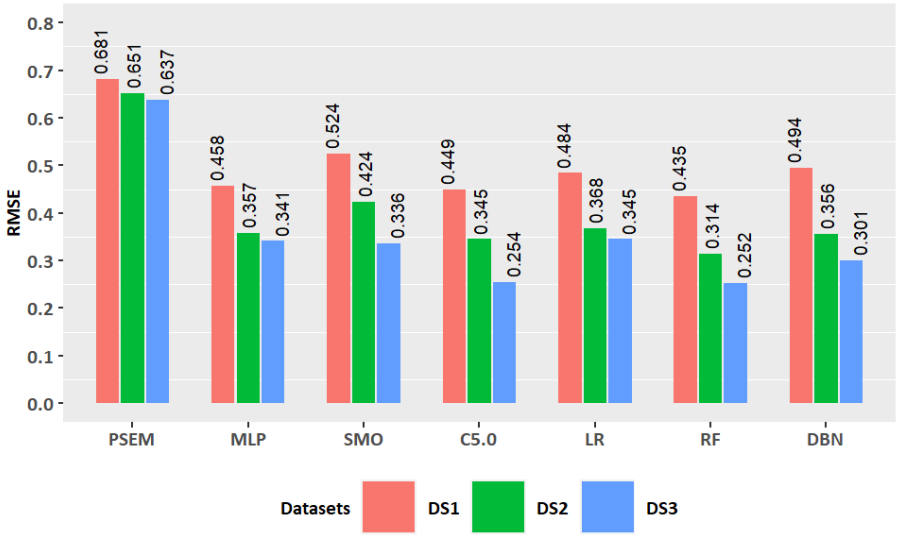


Figure 5.2: RMSE corresponding to the three datasets

5.4 Comparative Study Using Actual Datasets

The target of the study is to improve the poor performance of students in high schools. The data used in this study was obtained from 22 high schools in Cambodia. The dataset ADS consist of 1204 sampels is used in this experiment. The second experiment was carried out with the actual dataset ADS. ML model has been successfully applied for predictive modeling in various applications. Especially, Ensemble method has gained considerable attention in recent years because this method can substantially improve the accuracy of predictions by combining multiple machine learning algorithms [197]. From the experiment in the comparative study using synthetic datasets, we found the tree-based classifiers performance better in this prediction In this section, three models bagging CART, RF, and boosting C5.0 are used in addition to the models used in the experiment of previous section.

5.4.1 Experimental Results

Table 5.5 and Table 5.6 summarize ACC and RMSE of the proposed model, respectively. The experiment conducted using three main disciplines of EDM models. Those are PSEM of statistical analysis methods, seven superior classes (ANN, SMO, LR, KNN, Boosted C5.0, Bagged CART, and RF) of machine learning models, and DBN of deep learning. For instance-based KNN classifier, we implemented the algorithm with various values (from 1 to 30 where $30 = [\sqrt{\#\text{trainingset}}]$). However, the best performance of this model was obtained at $k = 1$ as shown in Figure 5.3.

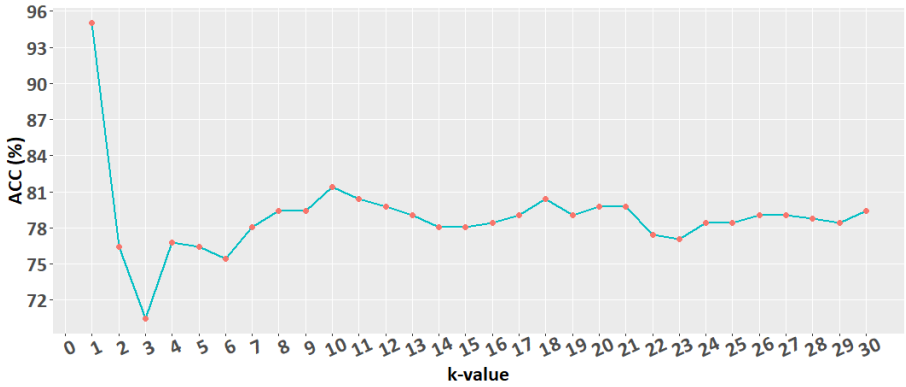


Figure 5.3: Performance of KNN on different values of parameter k

Table 5.5: The Model Performance base on ACC

Classes	Models	Lowest ACC	Highest ACC	Average ACC	Std.
Statistical Analysis	PSEM	31.50	54.20	38.20	5.01
Machine Learning	MLP	68.79	72.56	70.53	1.30
	SMO	85.52	94.14	89.66	3.01
	LR	52.75	60.34	56.37	3.01
	KNN	94.01	96.31	94.55	0.80
	Boosted C5.0	**95.27**	**96.28**	**95.67**	**0.30**
	Bagged CART	95.22	96.05	95.60	0.30
	RF	**96.04**	**97.10**	**96.69**	**0.30**
Deep Learning	DBN	72.60	79.78	75.76	3.01

From Table 5.5, ACC of the nine proposed models ranges from 38.20% to 96.69%. The four leading algorithms are KNN (94.95%), and the three ensemble tree-based algorithms: Boosted C5.0 (95.67%), Bagged CART (95.60%), and RF (96.69%). Using ANOVA test with significant level $p = 0.05$, we found KNN is significantly different from the other models. There is no difference between the Boosted C5.0 and Bagged CART algorithms. RF algorithm is statistically different from the other with the highest accuracy, which shows the robustness of RF and implied that it is a superior predictive algorithm in this predicting problem. As shown in Figure 5.4, KNN, Boosted C5.0, Bagged CART, and RF are the top four classifiers.

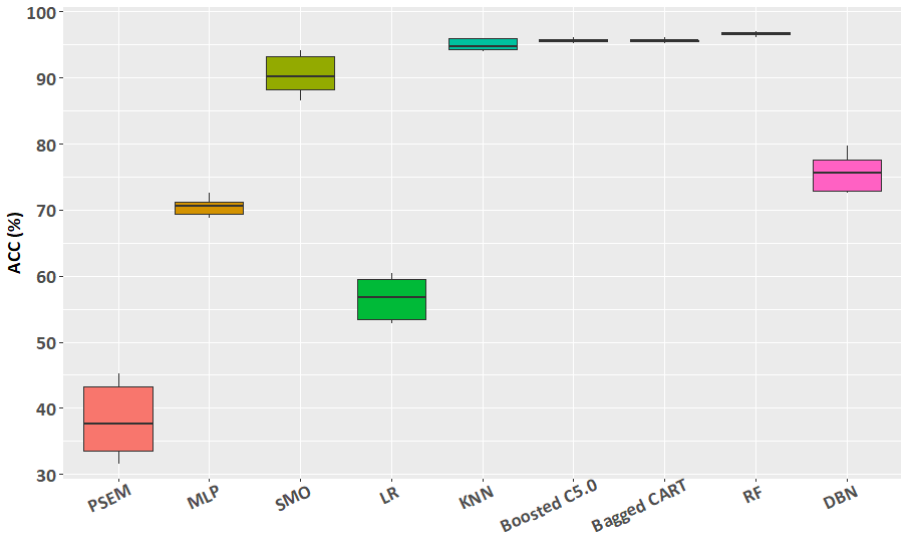


Figure 5.4: Boxplot indicate the accuracies of the proposed models

The summary value of RMSE of the proposed models is represented in Table 5.6. Figure 5.5 can be more perfectly visualizes the performance of the proposed models via RMSE. Recall that contrasting to accuracy, RMSE measures the average of the squared distance from the predicted output to the actual output. The smaller RMSE, the better model is. The zeros value of RMSE shows the model is perfect. From the figure, KNN,

Boosted C5.0, Bagged CART, and RF generate the smallest RMSE comparing to the the other models. It figured that RF produced the smallest RMSE, which shows that our prediction model is the best classifier and produces very little error in our prediction problem.

Table 5.6: The Model Performance base on RMSE

Classes	Models	Lowest RMSE	Highest RMSE	Average RMSE	Std.
Statistical Analysis	PSEM	1.146	1.211	1.176	0.032
Machine Learning	MLP	0.782	0.998	0.881	0.082
	SMO	0.654	0.758	0.708	0.048
	LR	0.958	1.052	1.007	0.033
	KNN	0.242	0.281	0.261	0.015
	Boosted C5.0	**0.089**	**0.143**	**0.122**	**0.011**
	Bagged CART	0.095	0.147	0.125	0.012
	RF	**0.089**	**0.134**	**0.114**	**0.011**
Deep Learning	DBN	0.813	0.881	0.844	0.025

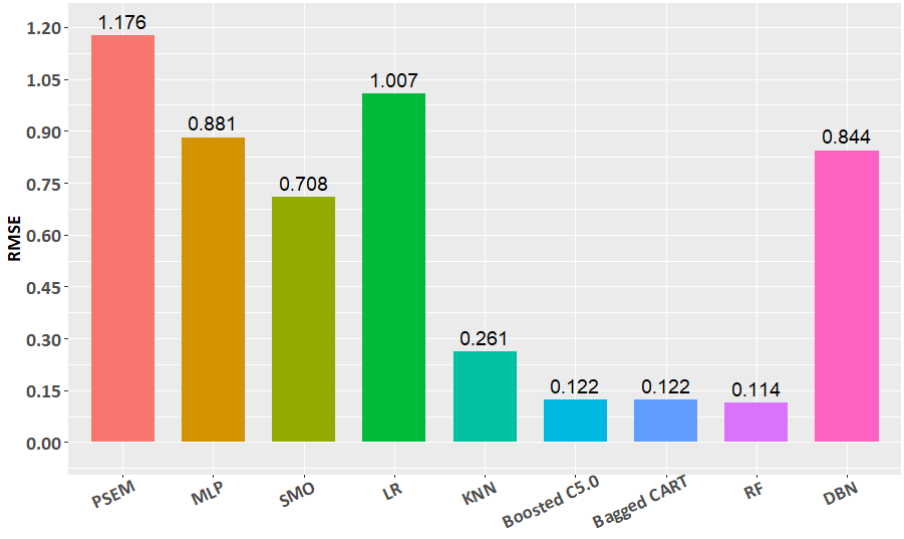


Figure 5.5: Boxplot indicate RMSE of the proposed models

5.5 Chapter Summary: Summary of Experimental Results

In this chapter, we wish to predict high school student performance with the analyze from various algorithms in educational data mining. To maximize the information and benefits for education institutes, we introduce multi-models in this study. We produce multiple classifiers to investigate the model that can best classify the performance levels of students in high schools. Hence, this research proposes three different main categories of research methods, statistical analysis method, machine learning, and deep learning as initial models in our comparative study.

The experiments were conducted with two phases. The first phase was conducted using three synthetics datasets of increasing sizes which stand for the average population of high school students in one town, one district, and one province in Cambodia. The second phase was computed using the actual dataset collected from Cambodia.From comparative study of prediction models using synthetic datasets, the treeb-based C5.0 and RF out perform the other models. DBN is found to perform better with the larger dataset. In the comparative study using the actual dataset, the experimental results reported that KNN, Boosted C5.0, Bagged CART, and RF are the four superior models standout the rest EDM models. In overall, RF model generate the highest accuracy with the lowest RMSE, which shows its effectiveness as the best algorithm. To sum up, two popular machine learning models RF and C5.0 are the leading models that will be used for model development or improvement. In addition, in context of larger datasets, the deep learning model DBN will be another candidate that can be improved and optimized for our academic prediction problem.

CHAPTER 6

THE PROPOSED HYBRID MACHINE MODELS FOR PREDICTING ACADEMIC PERFORMANCE

6.1 Introduction

In the age of the information and technology revolution, analysis of the database in education environments such as learning analytic, predictive analytic, educational data mining, and machine learning techniques has become a hot area of research. The supervised learning was used to predict, classify the students' performance and analyze their learning behaviors to follow up on the progress in their academic performance. However, the challenging task is to find the optimal prediction models which could produce satisfying results.

The preliminary experiment of comparative study of educational data mining (EDM) techniques was conducted in Chapter 5. The study picked up the state-of-art methods from literature review in Chapter 2, give investigation on model characteristics with advantages and disadvantages of the models in Chapter 3, and then apply the algorithm spot-checking to pick up to optimal models for comparative study. The preliminary experiment of comparative study of EDM techniques was conducted in Chapter 5. The study introduce three classes of EDM techniques consist of statistical analysis

techniques (predictive structural equation modeling), machine learning techniques (k-nearest neighbor, logistic regression, decision tree, sequential minimal optimization, and multilayer perceptron, and ensemble methods), and a deep learning model called deep belief network. The study utilized multiple of actual and synthetic educational datasets to extract the leading models for academic prediction problem. The experimental results demonstrate the tree-based models (random forest and C5.0) as the best models. This chapter aim at giving the study of model development for improvement of the leading models obtained from comparative study in the previous chapter.

6.2 The Concept for Model Improvement

From the results of comparative study of EDM methods, we found that the tree-based prediction models, C5.0 and random forest (RF) are the optimal algorithms that produce the highest classification accuracy and the smallest RMSE. However, the goal of this dissertation is to develop EDM classifiers that can produce the superior prediction and classification results on predicting academic performance. Hence, in this chapter, we propose the development of tree-base models (C5.0 and RF), with two potential machine learning algorithms: support vector machine (SVM) and naive Bayes (NB). We improve the classification performance of the four baseline models by solving the misclassification problem. The concept for improvement is to figure out the problem behind the misclassifcation of the models and propose the right methods to solve it.

In the tree-based algorithms C5.0 and RF, the major problem in the splitting process of the decision tree is overfitting. Overfitting caused by noisy data and irrelevant features and creates missclassifications resulting in overfitting and data imbalance. In return, overfitting reduces the accuracy of the decision tree in the classification model. In order to reduce high dimensional data which, contain noisy and irrelevant data, a common approach is to use feature reduction to obtain a lower-dimensional data which

depend on the feature criteria of problem sets. This study introduce principal component analysis and resampling method are applied to solve the misclassificaiton problem in these tree-based models. In addtion, support vector machine and naive Bayes are introduced to confirm the effectiveness of the proposed method.

SVM is known as one of potential classifier in prediction problem that produce the high classification performance. However, for high dimensional in SVM, if it is adopted without feature selection, then the dimension of input space is large and non-clean, lowering the performance of SVM. Thus, SVM requires an efficient and robust feature extraction methods that discard noisy, irrelevant and redundant data, while still retaining the discriminating power of the data. Removal of such features can increase the search speed and accuracy rate.

NB is a classifier that hold many advantage, however, the greatest weakness of the NB classifier is that it relies on the often-faulty assumption of equally important and independent features. Within this formula, if there are any features that are irrelevant to some class then the whole probability goes to zeros for that class because of this production, which leads to misclassification. In order to solve this problem, feature extraction will be the best tool to reduce irrelevant features, noisy data, and improve the classification performance.

6.3 The Proposed Machine Learning Techniques

In the classification problem, the misclassification that declines the performance of the model is from the quality of data that lowering the classification performance of the algorithms. We apply principal component analysis as a feature reduction to transform the original data into a new dataset of high quality that can reduce the overfitting in the tree-base classification rule. In addition to principal component analysis, we also introduced resampling method to enhance the performance of prediction model on new

or unseen dataset. The developed or improved model is called hybrid model which is the combination of the baseline machine learning model with the principal component analysis and resampling method.

6.3.1 The Principal Compoment Analysis

The majority task in supervised machine learning is classification. The classification problem is a hot issue in data mining and machine learning. However, the major problem in classification is overfitting and noisy data which leads to misclassification and reduce the accuracy of classification. To overcome this matter, we try to reduce irrelevant feature and non-correlated features which disturb in the classification process. In data analysis, it requires more computational resources and consumes much time when that data of huge volume. Hence, the feature extraction approach is used to remove the irrelevant or noisy feature from the data in order to reduce time and resource usage. The dimensional reduction could improve accuracy and boost up the performance by combining it with classification techniques. Using more high-quality data and feature reduction is one of the effective approaches to improve the performance of machine learning models. The four proposed models: SVM, NB, C5.0, and RF are the effective algorithms for classification problem, yet there is no perfect algorithm in machine learning.

In this proposed research, we proposed the hybrid models by a conjunction of baseline models with a resampling method (k-fold cross-validation) and principal component analysis (PCA). The introductory background of PCA was introduced in Chapter 3. This subsection give a summary of advantages of this unsupervised learning method in improving our prediction models. The merit of PCA is as follow:

(i) Removing the high noises from samples from the collected dataset in the preprocessing step.

(ii) Reducing the high dimensional data to low dimensional one which remains the

important characteristics of data.

(iii) Enhance the equality of importance of features that effectively improve the performance of classification.

The PCA is choosen among other well-known methods since it is less sensitive to different datasets than other holistic methods. That makes it become the most widely used unsupervised machine learning technique, so is considered as the best proposed method for this model improvement.

6.3.2 Resampling Method

Resampling is an important step that can significantly improve both the model's performance and the confidence in EDM model [198]. Resampling methods are an indispensable tool in modern statistics. They involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model. This allows us to gain more information that could not be available from fitting the model only once. Usually, the objective of a supervised learning is to create a model using training data, and have it make predictions on new data. Hence, the resampling methods allow us to see how the model would perform on data it has not been trained on, without collecting new data.

In this study, cross-validation (CV) is addressed as it is the most powerful preventative measure of resmapling method against overfitting [199][203]. The motivation to use CV techniques in this study is that the overfitting in the tree-based models can be reduced by generalization of this techniques on unseen or testing data. Without cross validation we only have information on how does our model perform to our in-sample data. Ideally we would like to see how does the model perform when we have a new data in terms of accuracy of its predictions. In science, theories are judged by its predictive performance. The CV, sometimes called rotation estimation, is a model validation technique for

assessing how the results of a statistical analysis will generalize to new dataset.

The k-fold cross-validation (k-FCV) is a standard cross-validation that resample the data into k subsets, called folds. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods. The process of resampling method by k-FCV can be summarized as the following algorithm:

Algorithm 1 : The Procedure of k-fold Cross Validation algorithm

1. Step 1: Shuffle the dataset randomly
2. Step 2: Split the dataset into k sub folds
3. Step 3: For each unique fold:
   - Take 1 fold as a hold out or test dataset
   - Take the remaining $k-1$ folds as training dataset
   - Retain the evaluation score and discard the model
4. Step 4: Repeat the iteration until every single fold was treated as test set. Finally, compute the average score of the recorded scores.

Each CV split dataset into a training set and a validation set. The training set is use to evaluate performance. The k-FCV is a model-assessment method system that evaluates the execution performance of machine learning prediction on a novel datasets on which is has been trained. It assists us in solving the problem of overfitting. To make the propose classifier more reliable and generalize it to independent datasets, the resampling method k-FCV was employed. The choice of k does affect the performance and validation of the model. In our study, we chose the 10-fold cross validation (will be shortly called 10_FCV)

to access our proposed algorithms since this split produce the highest performance of models.

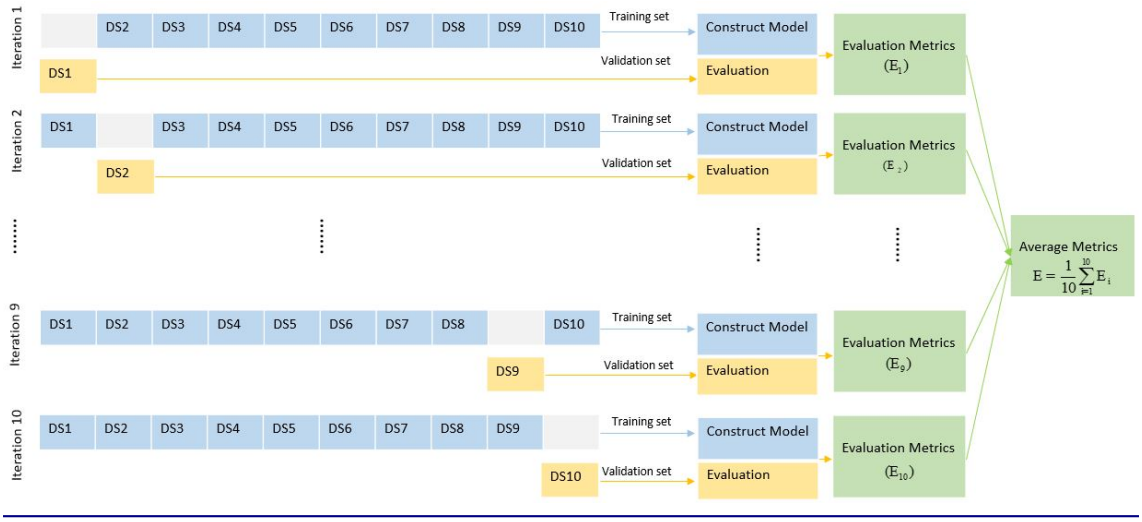


Figure 6.1: The resampling algorithm of 10-fold cross validation

6.4 Experimental Results

This section give the experimental results on the four models: RF, C5.0 , SVM with radial basis function kernel, and NB of the Bayesian network. The experiment was executed with two main phases. The first experiment or Phase 1 is the experiment on the baseline models to check the model performance and the overfitting. The second experiment or Phase 2 is the experiment on the proposed method. To confirm the effectiveness and robustness of our proposed algorithms, we use three datasets. The three datasets used in the experiment are an actual dataset ADS (1204 samples), and the two datasets are the synthetic datasets namely GDS2 and GDS3 in which GDS2 (2000 samples) is represented for the average population of high school students in one town, GDS3 (4000 samples) for the average population of high school students in one district in Cambodia. The experimental results are structured and shown in the following subsections.

6.4.1 Result of Baseline Models

The tree-based models RF and C5.0 is believed to be the potential model that oftentimes generate the high classification and prediction performance. The lower performance cause by the overfitting occure in these tree-based models. This section gives illustration of performance of baseline models and the difference between the training and testing set (the ovefitting and classification performance). The experimental results two standard evaluation metrics, accuracy (ACC) and root mean square error (RMSE) of the four baseline classifiers are shown in Table 6.1, Table 6.2, and Table 6.3.

Table 6.1: Performance Comparison on Dataset ADS

Model	Test ACC	Train ACC	Test ACC_std	Train ACC_std	Test RMSE	Train RMSE	Test RMSE_std	Train RMSE_std
SVM	78.33	98.35	2.991	0.144	0.888	0.095	0.041	0.016
NB	56.83	61.46	4.153	0.853	1.312	1.266	0.058	0.057
C5.0	82.47	97.45	3.612	0.312	0.732	0.245	0.034	0.017
RF	90.97	98.64	2.499	0.066	0.507	0.147	0.058	0.017

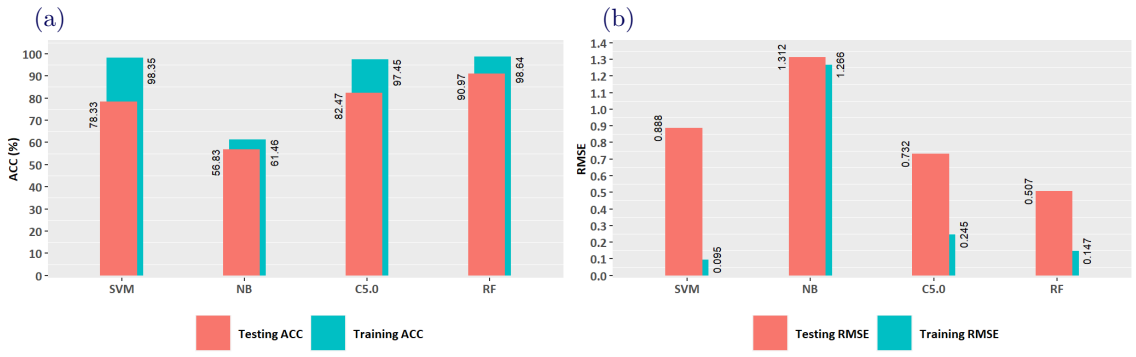


Figure 6.2: Performance of baseline models on data set ADS: (a) ACC; (b) RMSE

The details statistics of the performance of the four classifiers on dataset ADS are illustrated in Table 6.1. The second and third columns show the values of ACC on the testing and training set, where the seventh and the eighth columns demonstrate its

RMSE on both sets. The comparison of ACC and RMSE on testing set to the training set is shown in Figure 6.2. The ACC of SVM, NB, C5.0, and RF on the testting set are 78.33%, 56.58%, 82.47%, and 90.97% whereas the performance on the training set are 98.35%, 61.46%, 97.45%, and 98.64%. Simultaneously, RMSE of SVM, NB, C5.0, and RF on the testting set are 0.888, 1.312, 0.732, and 0.507 whereas the performance on the training set are 0.095, 1.266, 0.245, and 0.147. The standard deviation (std) of ACC and RMSE on the training set tends to be much more stable than on testing set. This results illustrated the difference performance of ACC and RMSE on the training and testing sets. The tree-based models RF and C5.0 produce the highest performance comparing to the other two models. However, the performance of the two models on training section are much higher than on the testing section resulting in the overfitting, which need to be solved or improved.

Table 6.2: Performance Comparison on Dataset GDS2

Model	Test ACC	Train ACC	Test ACC_std	Train ACC_std	Test RMSE	Train RMSE	Test RMSE_std	Train RMSE_std
SVM	80.64	96.78	2.991	0.044	0.731	0.343	0.085	0.011
NB	50.10	53.44	3.157	0.841	1.356	1.257	0.056	0.058
C5.0	83.56	97.69	3.612	0.312	0.655	0.219	0.054	0.027
RF	91.16	98.73	2.167	0.032	0.466	0.136	0.055	0.021

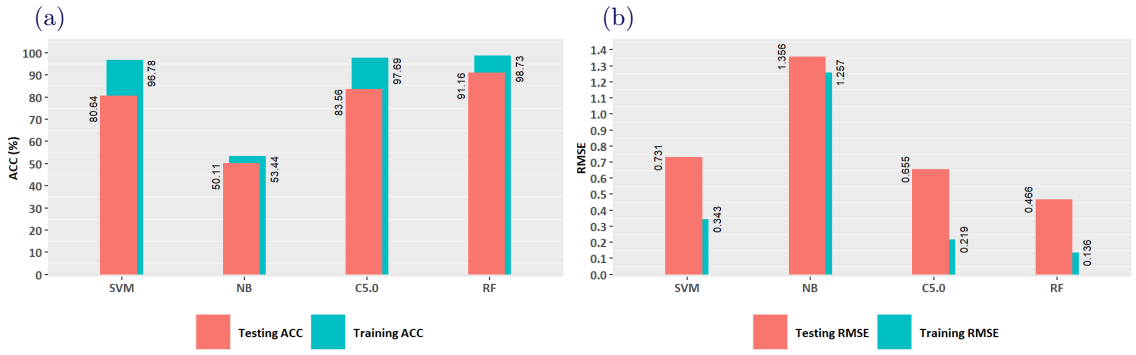


Figure 6.3: The ACC and RMSE one the four baseline models on GDS2

Table 6.2 shows the performance of the models on GDS2 dataset. The table summarize the comparative results of ACC and RMSE on the testing and training set. The performance of both metrics on both sets is shown in Figure 6.3. The ACC of SVM, NB, C5.0, and RF on the testing versus training sets are 80.64% : 96.70%, 50.10% : 53.44%, 83.56% : 97.69%, and 91.16% : 98.73%. In Figure 6.3 (b), RMSE of SVM, NB, C5.0, and RF on the testing set are 0.731, 1.356, 0.655, and 0.466 whereas the performance on the training set are 0.343, 1.257, 0.219, and 0.466. In this case, the models perform much better on the training set, yet perform poorer on the testing set.

Table 6.3: Performance Comparison on Dataset GDS3

Model	Test ACC	Train ACC	Test ACC_std	Train ACC_std	Test RMSE	Train RMSE	Test RMSE_std	Train RMSE_std
SVM	82.64	96.56	2.651	0.049	0.675	0.327	0.079	0.015
NB	65.68	69.12	2.118	0.857	0.917	0.775	0.063	0.042
C5.0	85.78	98.27	2.715	0.309	0.552	0.157	0.041	0.025
RF	93.25	98.95	2.213	0.051	0.378	0.108	0.043	0.025

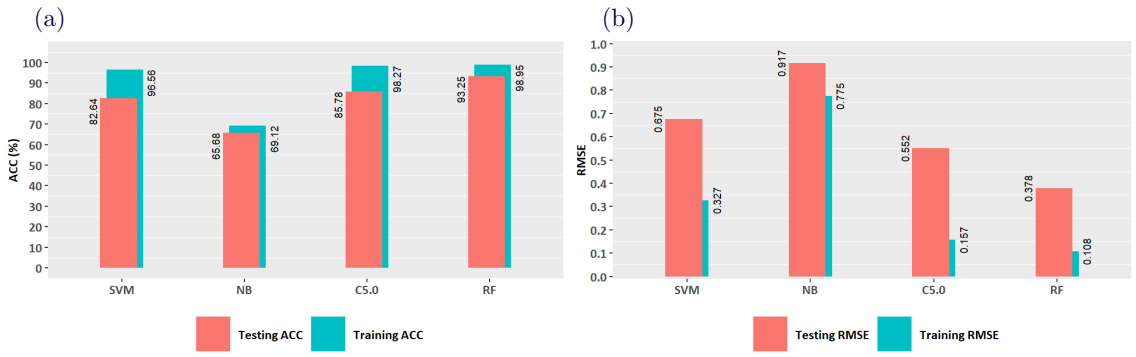


Figure 6.4: The ACC and RMSE one the four baseline models on GDS3

In addition to the previous experiment on dataset ADS and GDS2, another execution was carried out with dataset GDS3. The experimental results is shown in Table 6.3. The table records all the important statistics and visualize these results in Figure 6.4. The overlap barcharts demonstrate ACC and RMSE of four models on the testing and training

sets. ACC on the testing set are 82.64%, 65.68%, 85.78%, and 93.25% whereas ACC on the training set are 96.56%, 69.12%, 98.27%, and 98.95%. Simultaneously, RMSE of SVM, NB, C5.0, and RF on the testting set are 0.675, 0.917, 0.552, and 0.378 whereas the performance on the training set are 0.327, 0.775, 0.157, and 0.108. The figure perfectly show the different performance of the four models on both testing and training set.

In the overall performance of these baseline models, NB was found to be the poorest model in both training and testing set that infer the quality of data on classification rule of NB, which show the underfitting problem of this model. RF generate the highest performance outperform C5.0 and SVM on both training and testing set; however, the three models perform well on training data but poor on the testing data which show the overfitting problem that cause by noise of data. The next section introduce the proposed method to solve this problem and enhance the classification performance of the proposed models.

6.4.2 Experimental Results of the Proposed Models

Our proposed hybrid model was constructed by combing the supervised learning models with two powerful unsupervised learning methods, dimensional reduction PCA and the sampling method k-FCV. Feature reduction is one of the powerful unsupervised learning methods in classification models that are used for the purpose of removing irrelevant or non-related features. Dimensionality reduction by orthogonal vector projection via PCA can definitely serve as regularization in order to prevent overfitting and improve the model performance.

Many algorithms have been used to overcome the misclassification and overfitting problem, which splitting attribute is the major process in classification. In addition to PCA, we applied the resampling method to our dataset to gain information from data during splitting procedure. The sampling method used in our experiment is the k-FCV with the choice of $k = 10$ (denoted as 10_FCV) since our tree-based models performs the

best at this split. The 10_FCV was carried out with tuning the hyperparameters of the model with manual and grid search methods.

The experiment was executed as follows: baseline models with 10_FCV, baseline models with PCA, and lastly is our proposed hybrid model which is the combination of baseline model with 10_FCV and PCA. Table 6.5, 6.6, and 6.4 record the experimental results of the these models to the three datasets, GDS2, GDS3, and ADS, respectively. This section aims to illustrate the improvement of ACC and RMSE on the proposed methods by solving the overfitting problem and improve the classification performance.

Table 6.4: Performance Comparison of Proposed Models on Dataset ADS

Model	Test ACC	Train ACC	Test ACC_std	Train ACC_std	Test RMSE	Train RMSE	Test RMSE_std	Train RMSE_std
SVM+10_FCV	90.66	98.36	2.707	0.145	0.578	0.092	0.033	0.016
SVM+PCA	96.12	98.75	1.112	0.123	0.257	0.081	0.055	0.012
Hybrid SVM	**97.01**	**99.17**	**0.987**	**0.112**	**0.178**	**0.069**	**0.042**	**0.011**
NB+10_FCV	91.75	93.66	1.923	0.144	0.601	0.577	0.044	0.043
NB+PCA	97.23	99.37	0.457	0.143	0.218	0.149	0.051	0.032
Hybrid NB	**98.94**	**99.68**	**0.363**	**0.142**	**0.145**	**0.124**	**0.049**	**0.031**
C5.0+10_FCV	94.82	98.36	1.165	0.288	0.361	0.203	0.032	0.015
C5.0+PCA	98.13	100	1.081	0.000	0.148	0.000	0.034	0.000
Hybrid C5.0	**99.15**	**100**	**0.406**	**0.000**	**0.073**	**0.000**	**0.035**	**0.000**
RF+10_FCV	97.22	99.34	1.353	0.048	0.213	0.043	0.034	0.012
RF+PCA	98.77	100	0.981	0.000	0.139	0.000	0.033	0.000
Hybrid RF	**99.27**	**100**	**0.357**	**0.000**	**0.045**	**0.000**	**0.031**	**0.000**

Table 6.4 illustrates the performance of the proposed models on ADS dataset. The ACC and RMSE on the proposed Hybrid models generate the most successful results on both testing and training sets. The ACC of Hybrid SVM, Hybrid NB, Hybrid C5.0, and Hybrid RF on the test sets are 97.01%, 98.94%, 99.15%, and 99.27%. Simultaneously, RMSE of this four hybrid models on th test set are 0.178, 0.145, 0.073, 0.045.

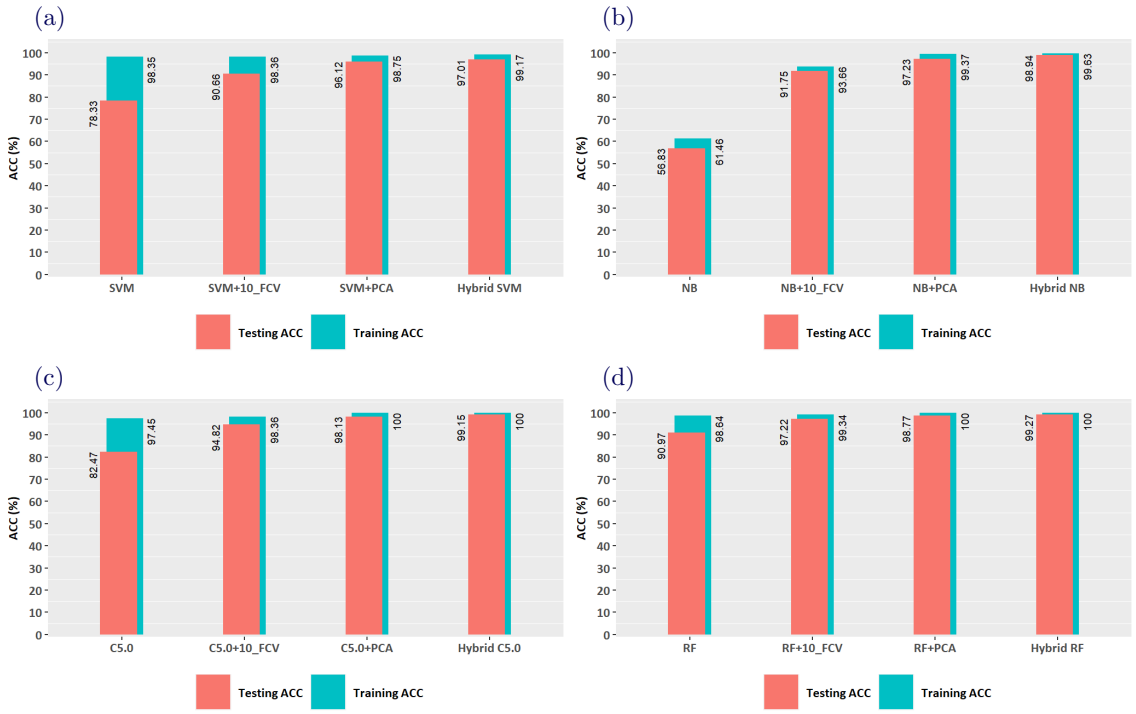


Figure 6.5: The ACC on ADS dataset (a) SVM; (b) NB; (c) C5.0; (d) RF

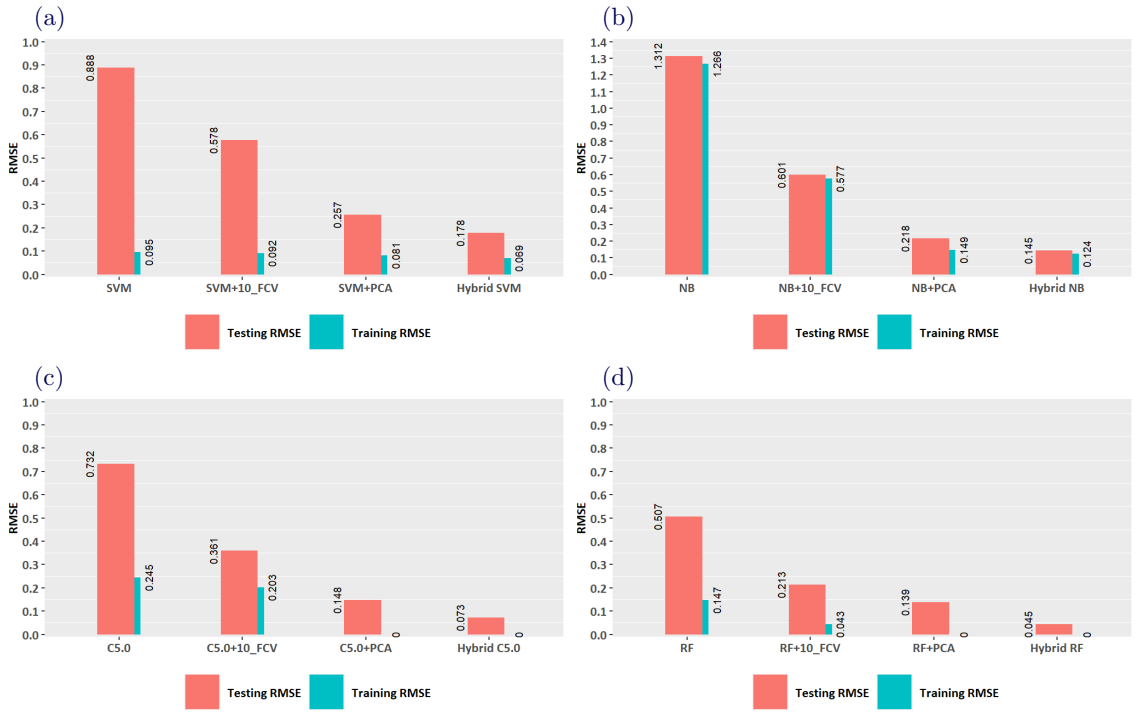


Figure 6.6: The RMSE on ADS dataset (a) SVM; (b) NB; (c) C5.0; (d) RF

To demonstrate the merit and effectiveness of the proposed hybrid models, we visual ACC and RMSE as in Figure 6.5 and Figure 6.6. The figures perfectly show the improvement of the model performance and reducing the overfitting in the models. The gap of ACC (between training ACC and testing ACC) of SVM, SVM+10_FCV, SVM+PCA, Hybrid SVM are 20.02%, 7.7%, 2.63%, and 2.16%. The gap of ACC of NB, NB+10_FCV, NB+PCA, Hybrid NB are 4.63%, 1.91%, 2.04%, and 0.74%. The gap of ACC of C5.0, C5.0+10_FCV, C5.0+PCA, Hybrid C5.0 are 14.98%, 3.56%, 1.87%, and 0.85%. The gap of ACC of RF, RF+10_FCV, RF+PCA, Hybrid RF are 7.67%, 2.12%, 1.25%, and 0.75%. That is the performance of the proposed Hybrid models can be significantly improved. The proposed RF generates the superior classification results.

Table 6.5: Performance Comparison of Proposed Models on Dataset GDS2

Model	Test ACC	Train ACC	Test ACC_std	Train ACC_std	Test RMSE	Train RMSE	Test RMSE_std	Train RMSE_std
SVM+10_FCV	87.08	97.52	1.124	0.051	0.496	0.272	0.025	0.015
SVM+PCA	92.27	98.03	1.119	0.017	0.316	0.229	0.048	0.013
Hybrid SVM	**93.88**	**98.11**	**1.123**	**0.013**	**0.274**	**0.165**	**0.016**	**0.012**
NB+10_FCV	68.33	72.77	1.763	0.122	0.854	0.658	0.070	0.033
NB+PCA	90.15	94.87	1.457	0.143	0.488	0.317	0.042	0.037
Hybrid NB	**91.42**	**95.12**	**1.593**	**0.154**	**0.393**	**0.221**	**0.033**	**0.035**
C5.0+10_FCV	95.24	98.66	1.105	0.167	0.295	0.117	0.056	0.021
C5.0+PCA	97.83	100	0.989	0.000	0.188	0.000	0.051	0.000
Hybrid C5.0	**98.12**	**100**	**0.979**	**0.000**	**0.167**	**0.000**	**0.027**	**0.000**
RF+10_FCV	95.48	99.44	1.064	0.055	0.293	0.083	0.029	0.017
RF+PCA	97.89	100	0.957	0.000	0.159	0.000	0.031	0.000
Hybrid RF	**98.25**	**100**	**0.918**	**0.000**	**0.156**	**0.000**	**0.031**	**0.000**

Table 6.5 illustrates the performance of the proposed models on GDS2 dataset. The ACC and RMSE on the proposed Hybrid models generate the most successful results on both testing and training sets. The ACC of Hybrid SVM, Hybrid NB, Hybrid C5.0, and Hybrid RF on the test sets are 93.88%, 91.42%, 98.12%, and 98.25%. Simultaneously, RMSE of this four hybrid models on th test set are 0.178, 0.145, 0.073, 0.045.

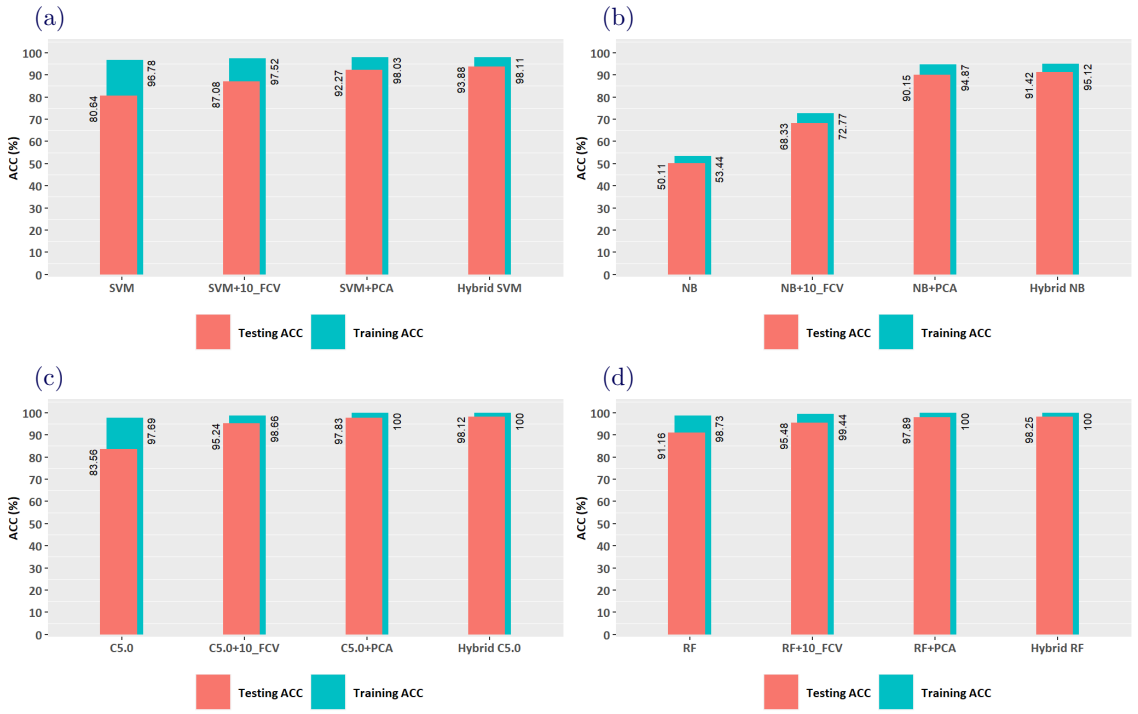


Figure 6.7: The ACC on GDS2 dataset (a) SVM; (b) NB; (c) C5.0; (d) RF

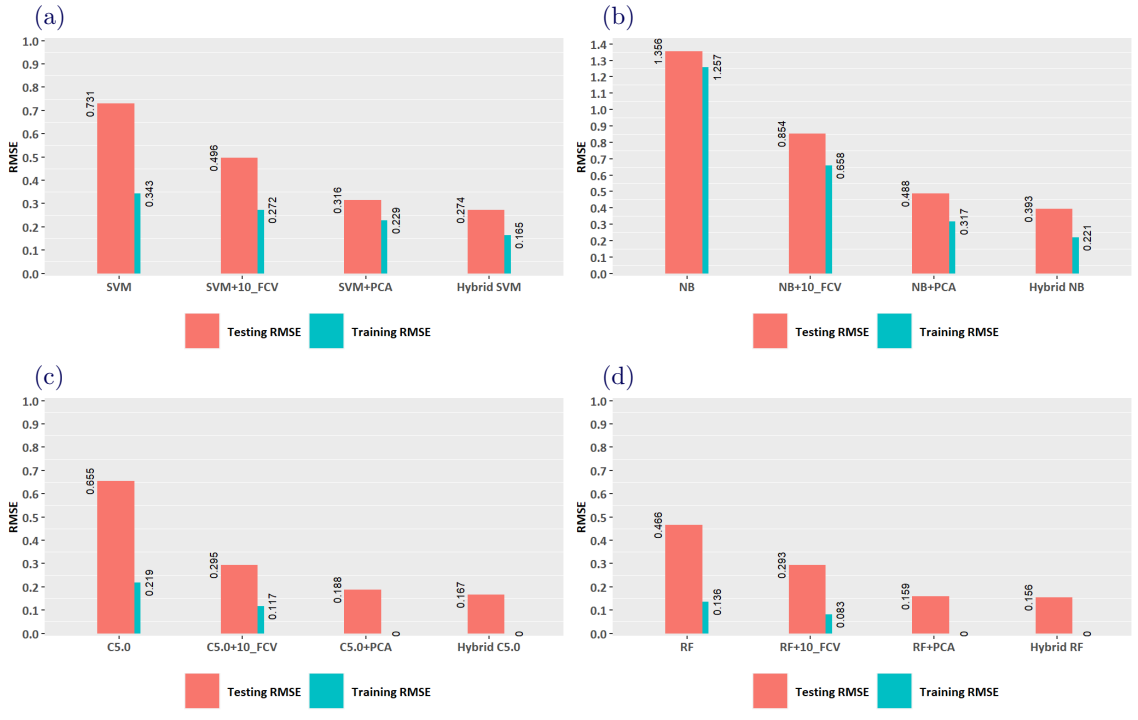


Figure 6.8: The RMSE on GDS2 dataset (a) SVM; (b) NB; (c) C5.0; (d) RF

We illustrate ACC and RMSE as in Figure 6.7 and Figure 6.8. The figures perfectly show the improvement of the model performance and reducing the overfitting in the models. The gap of ACC (between training ACC and testing ACC) of SVM on baseline, SVM+10_FCV, SVM+PCA, Hybrid SVM are 16.14%, 10.44%, 5.76%, and 4.23%. The gap of ACC of NB on baseline, NB+10_FCV, NB+PCA, Hybrid NB are 3.34%, 4.44%, 4.72%, and 3.7%. The gap of ACC of C5.0 on baseline, C5.0+10_FCV, C5.0+PCA, Hybrid C5.0 are 14.13%, 3.42%, 2.17%, and 1.88%. The gap of ACC of RF on baseline, RF+10_FCV, RF+PCA, Hybrid RF are 7.57%, 3.96%, 2.11%, and 1.75%. In this case, ACC and RMSE are improved both on testing and training sections.

Table 6.6: Performance Comparison of Proposed Models on Dataset GDS3

Model	Test ACC	Train ACC	Test ACC_std	Train ACC_std	Test RMSE	Train RMSE	Test RMSE_std	Train RMSE_std
SVM+10_FCV	93.15	98.12	0.909	0.042	0.374	0.189	0.045	0.016
SVM+PCA	95.69	98.55	0.811	0.029	0.305	0.149	0.053	0.011
Hybrid SVM	**96.52**	**99.33**	**0.591**	**0.015**	**0.212**	**0.088**	**0.046**	**0.013**
NB+10_FCV	76.47	79.55	1.263	0.122	0.598	0.473	0.069	0.037
NB+PCA	93.21	95.85	1.317	0.125	0.385	0.309	0.031	0.026
Hybrid NB	**95.37**	**97.11**	**1.263**	**0.137**	**0.289**	**0.193**	**0.047**	**0.029**
C5.0+10_FCV	95.69	98.79	1.015	0.213	0.324	0.103	0.047	0.022
C5.0+PCA	97.75	100	0.874	0.000	0.178	0.000	0.045	0.000
Hybrid C5.0	**98.62**	**100**	**0.775**	**0.000**	**0.165**	**0.000**	**0.028**	**0.000**
RF+10_FCV	96.58	99.32	1.033	0.049	0.273	0.044	0.028	0.015
RF+PCA	98.19	100	0.889	0.000	0.159	0.000	0.033	0.000
Hybrid RF	**98.71**	**100**	**0.732**	**0.000**	**0.155**	**0.000**	**0.027**	**0.000**

Table 6.6 presents the performance of the proposed models on GDS3 dataset. The ACC and RMSE on the proposed Hybrid models generate the most successful results on both testing and training sets. The ACC of Hybrid SVM, Hybrid NB, Hybrid C5.0, and Hybrid RF on the test sets are 96.52%, 95.37%, 98.62%, and 98.71%. Simultaneously, RMSE of this four hybrid models on th test set are 0.178, 0.145, 0.073, 0.045.

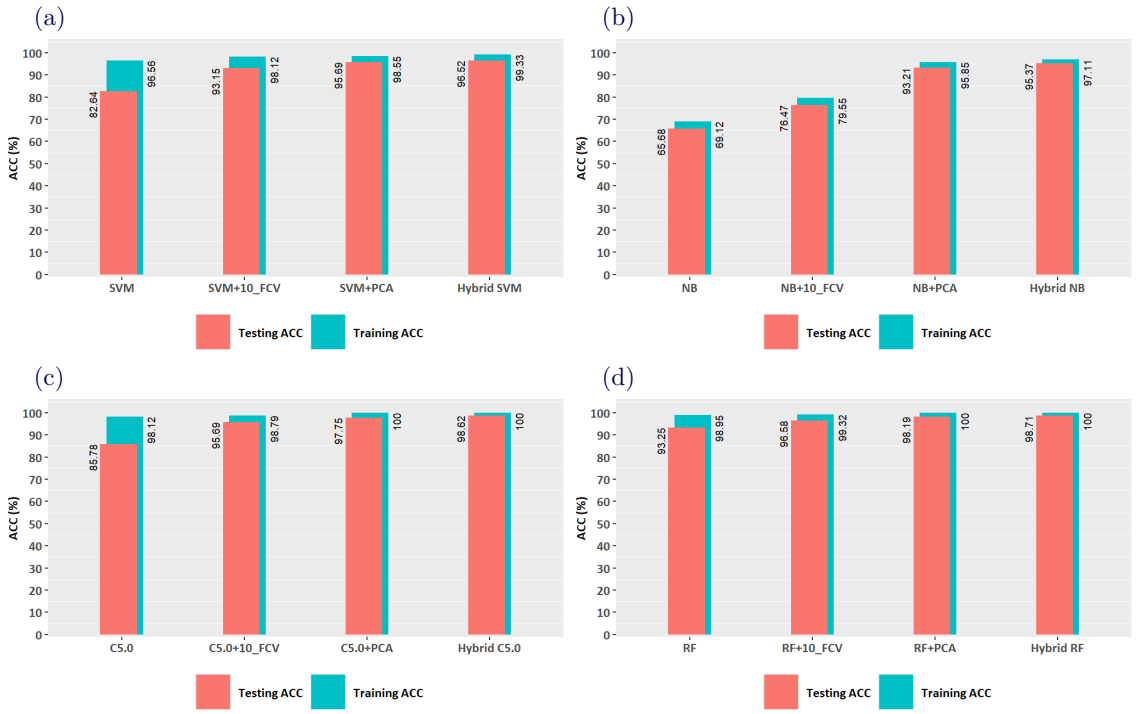


Figure 6.9: The ACC on GDS3 dataset (a) SVM; (b) NB; (c) C5.0; (d) RF

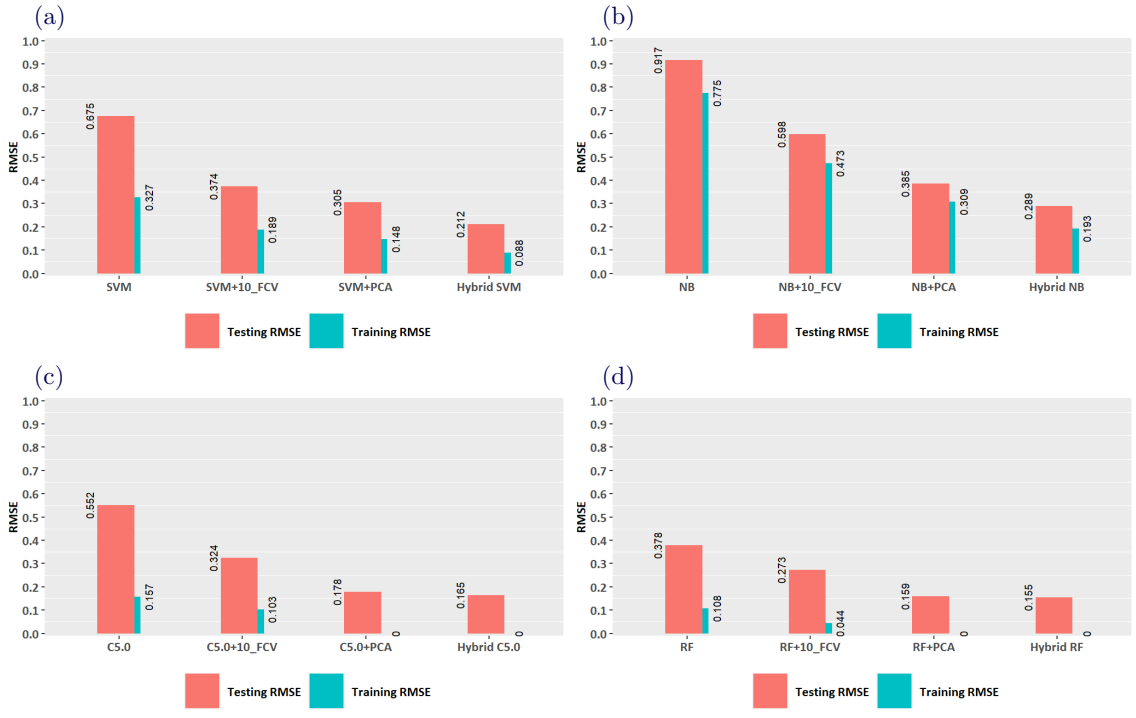


Figure 6.10: The RMSE on GDS3 dataset (a) SVM; (b) NB; (c) C5.0; (d) RF

The value of ACC and RMSE on GDS3 dataset are visualized as in Figure 6.9 and Figure 6.10. The figures perfectly show the improvement of the model performance and reducing the overfitting in the models. The gap of ACC (between training ACC and testing ACC) of SVM on baseline, SVM+10_FCV, SVM+PCA, Hybrid SVM are 13.92%, 4.97%, 2.86%, and 2.81%. The gap of ACC of NB on baseline, NB+10_FCV, NB+PCA, Hybrid NB are 3.44%, 3.08%, 2.64%, and 1.74%. The gap of ACC of C5.0 on baseline, C5.0+10_FCV, C5.0+PCA, Hybrid C5.0 are 12.49%, 3.1%, 2.25%, and 1.38%. The gap of ACC of RF on baseline, RF+10_FCV, RF+PCA, Hybrid RF are 5.7%, 2.74%, 2.81%, and 1.29%. Hence, the proposed method reduced overfitting and improve the classification performance of the models.

The experimental results on the three datasets demonstrate that the proposed hybrid models successfully solved the problem of overfitting on the four classifiers, especial our tree-based models. The proposed approach improve the performance of the models to reach the superior classification results with respect to both ACC and RMSE. In all experiment, the hybrid RF outperform the rest model. Hybrid C5.0 is the second best performer in this prediction problem.

6.5 Chapter Summary: Summary of Experimental Results

This study aim at giving the model development or improvement of the leading models obtained the comparative study in the previous Chapter 5. The tree-based models (RF and C5.0) are found to be the best two model in prediction student performance. To improve, the performance of the tree-based models, we proposed two unsupervised machine learning models to the baseline models to solve missclassification and overfitting problem. The four baseline models are SVM NB, C5.0, and RF. The unsupervised methods are dimensional reduction with PCA and resampling method with 10-fold CV. The hybrid machine models is named after combined the baseline models with the

proposed unsupervised learning models. The experiment of the baseline model was execute to check the overfitting and performance of the models. The experiment of the proposed mothods is subsequently implemented. Based on the both parameters of evaluation metrics, it shows that the proposed hybrid models by conjunction with PCA and resampling 10_FCV produced very satisfying results. Hybrid RF and Hybrid C5.0 standout the other models and produce superior classification result. These improved machine learning algorithms can be treated as very good classification models for classifying and predicting academic performance.

CHAPTER 7

THE PROPOSED OPTIMIZATION OF DEEP BELIEF NETWORKS FOR STUDENT PERFORMANCE PREDICTION IN CONTEXT OF LARGER DATASETS

7.1 Introduction

Machine learning was effectively used in educational setting. However, once the data get more complex and dataset increase, we believe that deep learning framework will solve the problem more effectively with regard to predicting and classification performance and computational time. Deep learning (DL) is one of the hottest trends in machine learning at the moment, and there are many problems where DL shines, such as robotics, image recognition and Artificial Intelligence (AI). The merits of DL framework will bring much better result both in term of classification performance and computational times. DL techniques aim to learn attribute hierarchies with an attribute from higher levels that is formed by the combining of other low features. Educational intuitions currently store an extensive collection of data and uncovered information. The analysis require effective approaches in leverage this type of dataset once the size and dimension of the data grow bigger. That why here, we try to improved the deep belief network (DBN) and treat as one of the effective model in our classification problem.

The experimental results from comparative study indicate the potential of DBN in this prediction problem in context of larger datasets. DBN is chosen among other advanced deep learning model in our study due to it suitability with characteristics of our dataset and the preliminary checking of algorithm spot-checking. Another main advantage of DBN is that it is hybrid model that carries the important characteristics of pretraining to avoid the problem of vanishing gradient and the problem of falling into local optimum by optimizing the initial parameters. The chapter give a study the optimization of DBN model for academic performance prediction in context of larger datasets.

7.2 Characteristic Descriptions of Deep Belief Network Model

DBN is a probabilistic generative multilayer neural network compose of several stacked Restricted Bolzmann Machines (RBM). In a DBN, every two sub-sequential hidden neural layers form an RBM. The input layer of the current RBM is actually the output features of a previous layer. Each model performs nonlinear transformation in its input vectors and produces as output, the vectors that will be used as input for the next model in the subsequence. Thus each RBM in the stack that makes up a DBN receives a different representation of the data. The basic DBN classification model involves preprocessing of data, pretraining, and fine-tuning. The difference between DBN and a neural network is the introduction of pretraining. The initial weight value can be close to the global optimization in the DBN model. This illustrate the merits of DBN via the specific characteristics of RBM. In Chapter 3, we present a section introducing the deep learning framework that proposed in our study. We initially introduce RBM and DBM models which are preliminaries to our proposed method. Recall that in RBM training process, the Minimizing Contrastive Divergence (MCD) training rule for an RBM replace the computationally expensive relaxation search of the Boltzmann Machine with a single step of Gibbs sampling and it does the good job for many applications [116][117]. In each

iteration, the weight w can be updated based on the MCD rule as follows:

Algorithm 2 : The Procedure of MCD algorithm

1. Step 1: Initialize the number of visible layer units n, the number of hidden units m , the number of training data N , the weight matrix w , the visible bias a , the hidden bias b , and the learning late η.

2. Step 2: Assign a sample x from the training data to be the initial state $v^{(0)}$ of the visible layer.

3. Step 3: Calculate $P(h_j^{(0)} = 1|v^{(0)})$ using equation (3.36), and extract $h_j^{(0)} \in \{0, 1\}$ from the computed conditional probability $P(h_j^{(1)} = 1|v^{(0)})$ where $j = 1, 2, ..., m$.

4. Step 4: Calculate $P(v_i^{(1)} = 1|h^{(0)})$ using (3.37), and extract updated $v_i^{(1)} \in \{0, 1\}$ from the computed conditional probability $P(v_i^{(1)} = 1|h^{(0)})$, where $i = 1, 2, .., n$.

5. Step 5: Calculate $P(h_j^{(1)} = 1|v^{(1)})$ from the updated $v^{(1)}$ using equation (3.36).

6. Step 6: Update the parameters using the equations:

$$w := w + \eta \left(P(h^{(0)} = 1|v^{(0)})(v^{(0)})^T - P(h^{(1)} = 1|v^{(1)})(v^{(1)})^T\right) \tag{7.1}$$

$$a := a + \eta \left(v^{(0)} - v^{(1)}\right) \tag{7.2}$$

$$b := b + \eta \left(P(h^{(0)} = 1|v^{(0)}) - P(h^{(1)} = 1|v^{(1)})\right) \tag{7.3}$$

7. Step 7: Assign another sample from the training data to be the initial state v_0 of the visible layer, and iterate Step 3 to 7 until all the N training data have been used.

The DBN is chose model in this study because DBN is competitive for three reasons: DBN can be fine-tuned as neural networks, DBN have many non-linear hidden layers, and DBN are generatively pretrained as shown in Algorithm 2. In addition, the real performance of DBN mentioned in many applications in literature review encourages using DBN. The proposed model consists of four stages: (1) A standard restricted Boltzmann machine is pretrained using a large number of unlabeled data to initialize its parameters; (2) the distinct number of different number of neurons in each hidden layers are trained for competitive learning; (3) competitive training and back-propagation algorithms are used to update the parameters to accomplish the optimal model; (4) by applying the recursive greedy-layer training and supervised fine-tuning, a deep neural network is built.

Even if DBN model holds many advantages, yet it has some disadvantageous that need a proper consideration and model setting up in order to get high classification performance. DBN implements pretraining in unsupervised learning and proposes layer-by-layer greedy learning strategy to better initialize the network that the training may be computationally expensive due to the initialization process and sampling. The model requires a proper selection of hyperparameters to achieve high performance. In addition, it is possible to get overfitting if we stack a large number of RBMs.

7.3 Optimization Approach for Deep Belief Network Model

For solving the problems mentioned above, we proposed three main approaches: feature selection, hyperparameter optimization, and regularization. To design a proper DBN-based students' achievement prediction models, we optimize the DBN model with (i) feature selection in preprocessing step to improve the model performance and reduction computation time, (ii) we optimized the performance of DBN with proper selection on hyperparameters to obtain the high classification performance (iii) Regularization methods to avoid overfitting and improve model performance. The details of these

proposed approaches are presented as the following.

7.3.1 Feature Selection

Feature selection is an integral task in supervised learning [45]. One main objective of the study is to identify the highly influenced features of the student's performance in high schools. In classification problems, once we have the table that contains attributes and class of attributes, we can measure the homogeneity (or heterogeneity) of the data based on classes. We say table or group is pure (homogeneity) if it contains only a single class, otherwise it is called impure (or heterogeneity). There are several indices to measure the impurity. Entropy and information gain are common statistical techniques for choosing informative attributes.

We use information gain (IG) to extract the high quality features used in this prediction. The details of this feature selection method was described in Chapter 3. Using this method, we can exclude the irrelevant, unimportant, and redundant variables, we can finally obtain a subset of predictive features with high quality that can boost the performance of prediction models. The experiment is executed on the new feature subset of lower dimension obtained from the feature selection method. This approach is proposed for not only to enhance model performance, but also to reduce computational cost on the lower dimension.

7.3.2 Hyperparameter Optimization

DBN model includes several important hyperparameters that need a proper selection or setting to obtain a high performance. Those hyperparameters can be set by hand (manual), selected by search algorithms (e.g grid search and random search) or optimized by applying model-based methods [201][202]. There is no exact rule to select the exact hyperparameter values in neural network; it depends on model architectures and the study problem. Proper hyperparameter selection requires a certain practical amount of

experience and experiment. Hence, the major task use in our problem is a proper selection of hyperparameters to obtain the robustness and high performance of the proposed DBN model. We aim to optimize DBN model with respect to different effective hyperparameters and the effects of weights and biases initialization on the output of the network. The performance of the model based on various hyperparameters is analyzed in this section. The highly influenced parameters considered in this literature are categorized into three main categories: those related to initialization step such as biases and weight, those related with model architecture, and those related to training process, as in follows.

- Weight and Biases Initialization: Offset (bias vector) and weight initialized in RBM.
- Model-based Hyperparameters: Deepth of network (number of hidden layers) and width of the network (number of units or neurons in each hidden layer)
- Training hyperparameters: Learning rate, batch size, number of epochs, and momentum.

To obtain the optimal values of this hyperparameters, we firstly give investigation on various literature views and do a preliminary experiment to select a range of values of each hyperparameter to obtain a structure of the best performance. We subsequently apply the hyper-parameters selection with three methods of hyperparameter searching: manual search, grid search, and random search, to obtain the optimal values. We run several iterations to confirm the stability of the model and get the average value of the evaluation metrics. The training was repeated for the randomly selected values in the range until the optimal values that produces the highest classification performance was obtained. Although not all the studies analyzed in this section provide the details about the hyperparameter used, references are provided when available.

Weight and Biases Initialization

Weights and biases initialization has been widely recognized as one of the most effective approaches in speeding up the training of a neural network. In fact, it influences not only the speed of convergence, but also the probability of convergence and the generalization [203]. Using too small or too large values could speed up the learning, but at the same time, it may end up performing worse. In addition, the number of iterations of the training algorithm and the convergence time would vary depending on the initialized values. Three different initialization methods have been applied during the training; small random values between 0 and 1, large random values between 1 and 10, and initializing with zeros. Initializing the weights with small random values between 0 and 1 results the least error [108][204]. Starting with the same weights makes the neurons follow the exact same updating rule, and will always end up doing the same. Using large values helps the algorithm to converge faster, however, the results are not as good as using small initial values. In this study, the training of each RBM wa carried out using MCD algorithm mentioned in Algorithm 2. At the beginning of learning, the weight are initialized to values sampels from a Gaussian distribution with mean 0 and standard deviation of about 0.01. The visible and hidden biases were initialized to 0.

Model-based Hyperparameters

In each RBM, the input data of the visible layer are reconstructed and mapped to the hidden layer again, and the reconstruction error is calculated based on the difference between the reconstructed output data and the initial training data. There is relationship between reconstructed error and network mechanism, and it is reasonable to reply on reconstruction error to determine the network depth of DBN autonomously. The number of neurons in each layer also has an impact on the network. At present, there is lack of a clear theory to prove that the appropriate number of cells is set and the improvement is achieved. DBN structure focus on the ability to determine the depth of a network, and the number of neurons in each layer is fixed. The more complex the network structure of

DBN, the stronger the ability to solve complex problem. Simultaneously, the higher the number of networks layers, the hard the training will be, the greater the training error accumulates, the lower the correctness of the model. In application, in order to establish suitable DBN structure for specific tasks, due to lack of corresponding theoretical support and effective training mode, the depth of network and the number of hidden units need to be set by experience, which leads to the deviation in the modeling process and the high cost.

(i) Number of Hidden Layers

The fewer hidden layers training may be shortly trained with low cost, but result in poor performance. However, too many layers will lead to overfitting and high computational cost. Regarding with the number of layers, most of the implementation ranges from 1 to 6 layers: 1 layers [113][205][207][210], 2 layers [110][113][205][212][206][207][210], 3 layers [115][205][206][210], 4 layers [205][206][210], 5 layers [205][206], and 6 layers [205][206]. Most of these studies indicates that 2 and 3 layers produce highest performance. In our studies, we chose a range of hidden layer from one to four layers with different neuron in each layers. Most of studies The number of hidden layers are denoted as $1L, 2L, 3L$, and $4L$ for one hidden layer, two hidden layers, three hidden layers, and four hidden layers, respectively. To void complexity, we fixed the neurons in each hidden layers. We investigate manually on the performance of the proposed DBN models with different value of layers. The experimental results illustrated that the proposed DBN with $2L$ and $3L$ are slightly different, while the DBL with $1L$ and $4L$ perform poorer.

(ii) Number of Neurons in Hidden Layers

Similarly to number of hidden layers, too little or too many of neurons (units) in each layer will lead to poor performance [108]. To obtain the better performance of the DBN, we investigate several cases of neuron in each hidden layers. There is no exact rule

for selecting number of hidden neurons, different literature recommend different values. In this study, we select all recommended value and thump rule for selecting number of hidden neurons from the studies in [208][209][210]. From these studies, we obtained a set of hidden neurons $\{25, 50, 100, 150, 200\}$. We fixed the same number of neurons on each hidden layers then do permutation, then apply the grid search on these range. The hidden layers with less than 50 and larger than or equal to 150 was found to generate the lower performance.

Training Hyperparameters

(i) Learning Rate

The function of learning rate effectively investigates how much to update the weights in the training process. Generally, large value of learning late can learn the training quickly, yet may not converge or lead to poor performance. Conversely, too small value will require much time and high energy consumption for training. Several studies proposed various optimal values of learning rate in the range $\eta \in (0, 1]$ [108][209][206][207]. The study in [210] give analysis on learning rate in DBN and investigate on the search space $(0, 2]$. In our experiment, we executed the proposed DBN with the random search of learning rate ranges from 0 to 2. We found the learning rate role very important in our DBN model, the low performance was found at learning rate lower than 0.5 and the high performance was obtained at the learning rate between 0.7 to 1.5.

(ii) Mini-Batch Size

Mini-batch divides a dataset into small chunks and performs the learning in each chunk [211]. The batch size role importantly in this performance, we run several iterations of different batch sizes and found that when the batch size is too small or too big, the performance is low [108]. It found that the performance is high when it is 0.01% to 0.05%

of the training sets [210]. The experimental results shows that the mini-batch size of lower than 5 and higher than could not produce a satisfying result, and we found the batch of size from 5 to 30 gave higher accuracy of classification.

(iii) Number of Epochs

In each RBM training, the MCD algorithms require a certain number of iterations, so that it can converge to the optimal value. The more number of interactions perceptively yield better result, but in fact, it will not only consume much energy, but also end up with overfitting in data training. Several studies proposed the optimal values of the number of epoch in DBN [110][206][208]. We observe different epochs with the grid search of 5, 10, 20, 50, 100, 150, and 200. The experimental results indicated that the epoch that produces higher accuracy ranges from 75 to 150.

(iv) Momentum

Momentum is a popular extension of backpropagation that help to prevent the network from failing into local optimum [211]. The techniques add the fraction of the previous weight update to the current update. Using momentum when updating parameters is a useful tool to avoid local minima during training. In a model using momentum, the gradient for each update effects the current velocity of a parameter instead of the parameter itself. The recommended value of momentum to start with 0.5 and then search between 0 and 0.9 [108]. The momentum coefficient used in this study is randomly searched from the search space $[0, 0.9]$.

7.3.3 Regularization Methods

Regularization method is used to continually shrink weights to ensure that they don't grow too large. One of the first methods we should try when we need to reduce overfitting is regularization and improve the performance of neural network [108][212]. It involves

adding an extra element to the loss function, which punishes our model for being too complex or, in simple words, for using too high values in the weight matrix. In our optimization approach, the model is possible to be overfitting when we over trained the model. In this study we use the simplest penalty function *L2 regularization*. We add this regularization term to loss function or cross entropy cost function to account for weight that have grown large, and it can be computed as haft of sum of squared weights multiply with coefficient λ:

$$J'(x, y, \theta) = J(x, y, \theta) + \lambda \frac{||w||^2}{2}, \tag{7.4}$$

where λ is the regularization parameter, which is used to control how quickly the weights decay. In order to analyze the effectiveness and robustness of IDBN, we study different value of the regularization parameter and we found the best fixed of parameter at 0.005.

We observed all the hyper-parameter and summarized the optimal values of all parameters in Table 7.1 below:

Table 7.1: The Optimal Parameters

Pretraining	Value	Fine-tuning	Value
Hidden Layers	2L	L2 Regularization: λ	0.005
Neurons in Each Layer	100-100	Momentum	0.01
Number of Epochs	20	Number of Epochs	100
Learning Rate	0.5	Learning Rate	0.8
Mini-batch Size	20	Mini-batch Size	5

7.4 Experimental Result

We proposed the improved DBN (IDBN), and compare it with the classical DBN (DBN), and the other two neural networks: multiplayer perceptron (MLP) and learning vector quantization (LVQ), and a superior Hybrid RF which was found to be the best model in Chapter 6. In our experiments, we processed it in two phases. The first phase,

Phase I, is to implement our experiment using the actual datasets obtained from the data collection in Cambodia; we named this dataset as ADS. The second phase is to execute our experiment using four artificial datasets. The artificial datasets were generated based on information from several literature reviews and information from our actual dataset. With the merits of deep learning framework, we aim to generate the optimal DBN model for predicting academic performance with the larger datasets. Hence, in the experimental of Phase II, we aim at confirming the robustness and effectiveness of proposed improvement methods of IDBN by testing with four synthetic datasets of the increasing sizes; small dataset GDS1 (1000 samples) for the average population of high school students in one commune; meduim dataset GDS2 (2000 samples) for the average population of high school students in one town; larger dataset GDS3 (4000 samples) for the average population of high school students in one district; and large dataset GDS4 (10000 samples) as the average high school student population in one province in Cambodia. In measuring and evaluating the performance of our proposed models, we used two standard evaluation metrics: accuracy (ACC) and root mean square error (RMSE).

7.4.1 Experimental Results of Phase I

In this section we gave the experimental result of the proposed models on the actual data (ADS) that is collected from Cambodia with the size of 1204 samples.

Table 7.2: Performance comparison of the proposed models of ADS

Model	Lowest ACC	Highest ACC	Average ACC	Std.	Lowest RMSE	Highest RMSE	Average RMSE	Std.
MLP	68.79	72.56	70.53	1.322	0.782	0.998	0.881	0.082
LVQ	82.32	85.34	84.16	0.903	0.312	0.723	0.520	0.154
DBN	75.00	79.20	76.38	1.337	0.712	0.926	0.849	0.067
IDBN	85.35	87.61	86.59	0.740	0.327	0.566	0.419	0.082
Hybrid RF	99.01	100	99.72	0.357	0.000	0.078	0.045	0.031

The experimental results is presented in Table 7.2, ACC of the proposed models ranges from 53.15% to 99.72%. ACC of MLP, LVQ, DBN, IDBN, and Hybrid RF are 70.53%, 84.16%, 79.20%, 86.59%, and 99.72% respectively. The IDBN outperforms other models, and the classical DBN was improved from 75.07% to 84.25% by using feature selection technique and optimized reconstruction error via the hyper parameters.

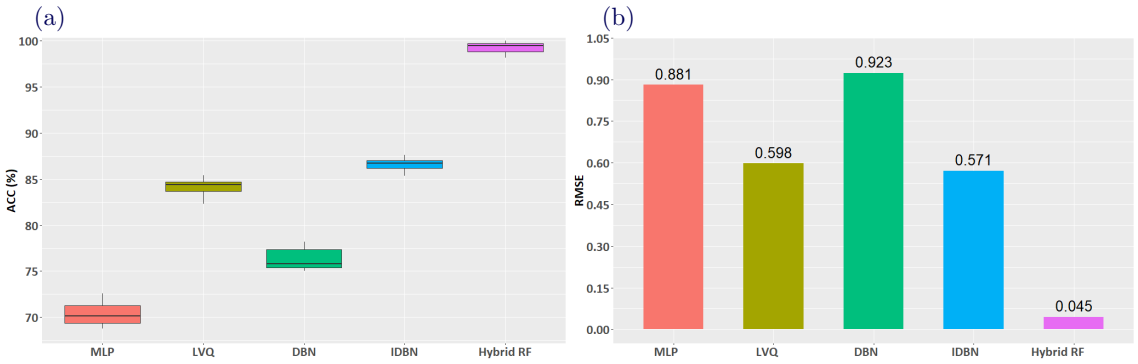


Figure 7.1: (a) ACC of the proposed models on dataset ADS; (b) RMSE of the proposed models on dataset ADS

The performance base on ACC and RMSE of the proposed model are visualized in Figure 7.1. Hybrid RF is the best prediction model followed by IDBN as the second best model as shown in Figure 7.1 (a). From Figure 7.1 (b), RMSE of the five algorithms. IDBN algorithm generates the second smallest RMSE follows Hybrid RF.

In this phase, we found that IDBN performs the highest accuracy smallest RMSE comparing to other proposed neural network models, and follows the Hybrid RF model. We observed the performance of IDBN with larger datasets in the next experiment. In dataset ADS, the accuracy of the proposed models ranges from 53.15% to 99.72%. The ACC of MLP, LVQ, DBN, IDBN, and Hybrid RF are 70.53%, 84.16%, 79.20%, 86.59%, and 99.72% respectively. IDBN outperforms other models, and the classical DBN was improved from 75.07% to 84.25% by using feature selection technique and optimized reconstruction error via a proper hyperparameters selection and network regularization.

In this experiment, the IDBN produced the highest accuracy and the smallest RMSE compared to other proposed neural network models and follow the Hybrid RF model. The performance of IDBN with larger datasets is assessed in the next experiment.

7.4.2 Experimental Results of Phase II

In this section, we give a comparative study of the performance of IDBN with respect to the increasing size of datasets. Due to the high cost with a short time for data collection in Cambodia, we could not obtain a larger size of data. However, in improving academic performance in the educational system, the analysis of big data will be required. Hence, in this phase, we proposed four generated artificial datasets. The four datasets are denoted as GDS1 (1000 samples) for average population of high school students in one commune, GDS2 (2000 samples) for average population of high school students in one town, GDS3 (4000 samples) for average population of high school students in one district, and GDS4 (10000 samples) as average high school student population in one province in Cambodia. It will be larger for some provinces, the whole country and so on, yet it is enough to state with these four sizes in our experiment. We executed our experiment of each model independently with 50 times run. The statistics of performance based on classification accuracy and RMSE of IDBN and other comparative models with respect to four datasets are shown in Table 7.3, Table 7.4, Table 7.5, and Table 7.6, respectively.

Table 7.3: Performance Comparison of the Proposed Models of GDS1

Model	Lowest ACC	Highest ACC	Average ACC	Std.	Lowest RMSE	Highest RMSE	Average RMSE	Std.
MLP	59.38	68.59	64.59	2.238	0.996	1.126	1.067	0.046
LVQ	73.60	83.20	78.13	2.684	0.409	0.551	0.479	0.034
DBN	64.60	78.80	73.48	3.707	0.707	1.015	0.874	0.077
IDBN	74.67	85.36	80.44	2.692	0.732	0.948	0.781	0.054
Hybrid RF	91.67	96.60	93.67	1.282	0.173	0.314	0.275	0.022

Table 7.4: Performance Comparison of the Proposed Models of GDS2

Model	Lowest ACC	Highest ACC	Average ACC	Std.	Lowest RMSE	Highest RMSE	Average RMSE	Std.
MLP	76.79	82.72	79.78	1.415	0.713	0.884	0.801	0.037
LVQ	84.20	89.80	87.37	1.484	0.319	0.497	0.395	0.050
DBN	82.20	87.40	84.82	1.470	0.435	0.764	0.649	0.064
IDBN	93.17	97.17	95.01	1.054	0.173	0.501	0.368	0.060
Hybrid RF	96.11	99.93	98.09	1.201	0.022	0.511	0.171	0.164

Table 7.5: Performance Comparison of the Proposed Models of GDS3

Model	Lowest ACC	Highest ACC	Average ACC	Std.	Lowest RMSE	Highest RMSE	Average RMSE	Std.
MLP	90.19	95.82	93.24	1.391	0.244	0.615	0.474	0.075
LVQ	87.20	92.60	90.24	1.615	0.260	0.386	0.315	0.044
DBN	88.40	92.30	90.83	1.181	0.397	0.705	0.577	0.080
IDBN	97.92	99.75	99.02	0.545	0.050	0.263	0.161	0.059
Hybrid RF	97.00	99.95	98.71	0.732	0.044	0.456	0.169	0.097

Table 7.6: Performance Comparison of the Proposed Models of GDS4

Model	Lowest ACC	Highest ACC	Average ACC	Std.	Lowest RMSE	Highest RMSE	Average RMSE	Std.
MLP	92.39	96.82	94.34	1.404	0.282	0.460	0.374	0.051
LVQ	88.00	92.01	90.39	1.424	0.266	0.384	0.327	0.034
DBN	93.17	96.84	95.06	1.118	0.277	0.675	0.414	0.126
IDBN	98.48	99.87	99.37	0.378	0.055	0.183	0.121	0.040
Hybrid RF	97.79	99.95	98.88	0.668	0.054	0.270	0.158	0.055

From Table 7.3, the accuracy of the proposed models ranges from 64.59% to 93.67%. The accuracies of MLP, LVQ, DBN, IDBN, and Hybrid RF are 64.59%, 78.13%, 73.48%, 80.44%, and 93.67%, respectively. The average RMSE of the four models are 1.067, 0.479, 0.874, 0.781, and 0.275. The statistics with respect to ACC of these models is illustrated in the Figure 7.2 (a). The IDBN outperforms other three neural network models, and by

and be the second best model follow the Hybrid RF. The proposed IDBN boost up the performance of the classical DBN with the accuracy of 7% increased.

Simultaneously, Figure 7.2 (b) shows the RMSE metric for dataset GDS1, we visualize the comparison of RMSE of our algorithms. The Hybrid RF generate the smallest RMSE, and follow by LVQ and IDBN as the second and third smallest, respectively. The RMSE of LVQ is smaller than the IDBN, yet it produced smaller standard deviation, which shown that it is more accurate and stable.

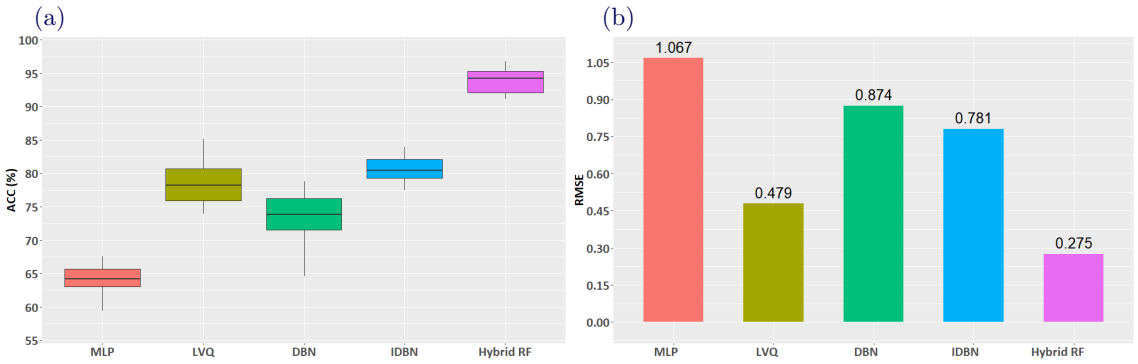


Figure 7.2: (a) The ACC of the proposed models on dataset GDS1; (b) RMSE of the proposed models on dataset GDS1

Table 7.4 summarizes ACC and RMSE of the models. ACC of the proposed models ranges from 79.78% to 98.02%. ACC of MLP, LVQ, DBN, IDBN, and Hybrid RF are 79.78%, 87.37%, 84.82%, 95.01%, and 98.09%, respectively. The proposed DBN outperforms other models, and by applying the feature selection and using optimal parameter, the improved DBN boost up the performance of the classical DBN with the accuracy of 11% increased. The improved DBN still performance as the second runner as illustrated in Figure 7.3 (a).

The comparative results of RMSE of our prediction models on dataset GDS1 is shown in Figure 7.3 (b). The MLP produces the highest error, while the Hybrid RF generates the lowest value of RMSE. The improved DBN algorithm remain generate the second smallest value of RMSE that outperforms the other neural network models.

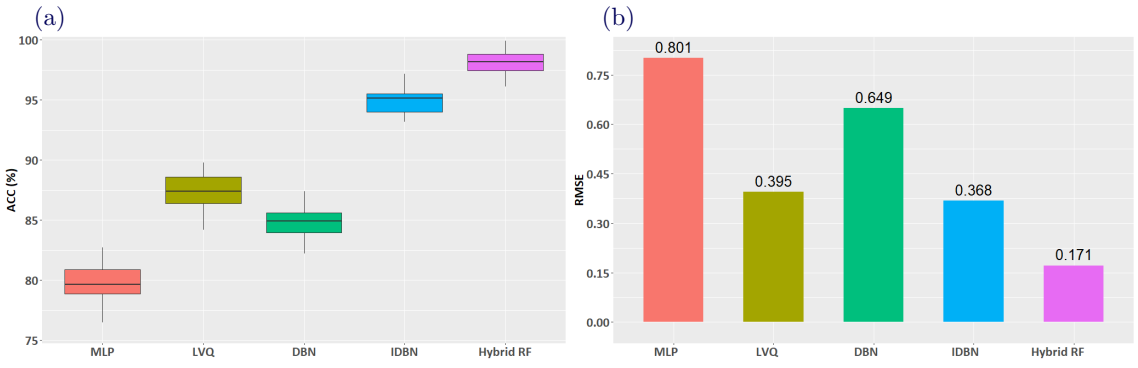


Figure 7.3: (a) The ACC of the proposed models on dataset GDS2; (b) RMSE of the proposed models on dataset GDS2

ACC and RMSE of dataset GDS3 is shown in From Table 7.5. The accuracy of the proposed models ranges from 90.24% to 99.02%. The accuracies of MLP, LVQ, DBN, Improved DBN, and Hybrid RF are 93.24%, 90.24%, 90.83%, 99.02%, and 98.71%, respectively. IDBN and Hybrid RF generate the highest accuracy. However, the Improved DBN produced the more accurate and robust as shown in the Figure 7.4 (a).

RMSE of our proposed models corresponding to dataset GDS3 is shown in Figure 7.4 (b). Base on the figure, IDBN and Hybrid RF produce the small values of RMSE. The experimental results show that these two model generate the smallest error in predicting student learning outcomes.

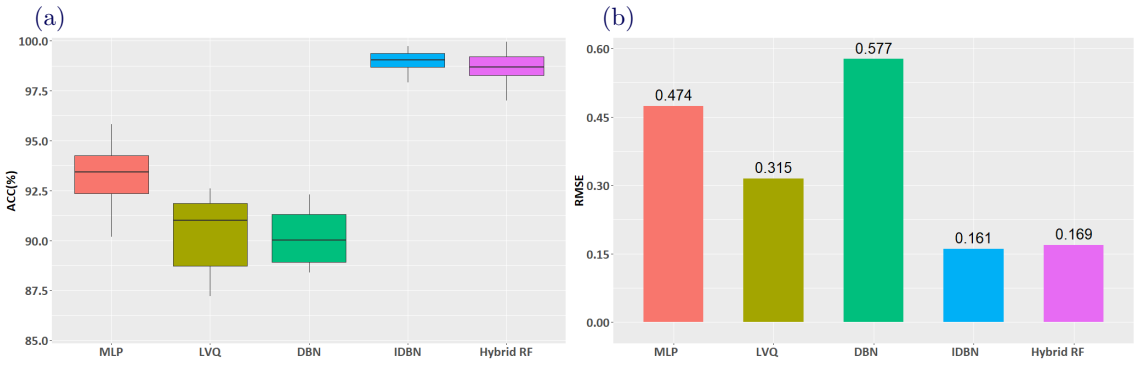


Figure 7.4: (a) ACC of the proposed models on dataset GDS3; (b) RMSE of the proposed models on dataset GDS3

ACC and RMSE of dataset GDS4 is shown in From Table 7.6. The accuracy of the proposed models ranges from 90.39% to 99.37%. The accuracies of MLP, LVQ, DBN, IDBN, and Hybrid RF are 94.34%, 90.39%, 95.06%, 99.37%, and 98.88%, respectively. IDBN and Hybrid RF generate the highest accuracy. However, IDBN produced the more accurate and robust as shown in Figure 7.5 (a). With respect to the RMSE metric, the comparison of RMSE of our algorithm is shown in Figure 7.5 (b). IDBN algorithm remain generate the smallest value of RMSE.

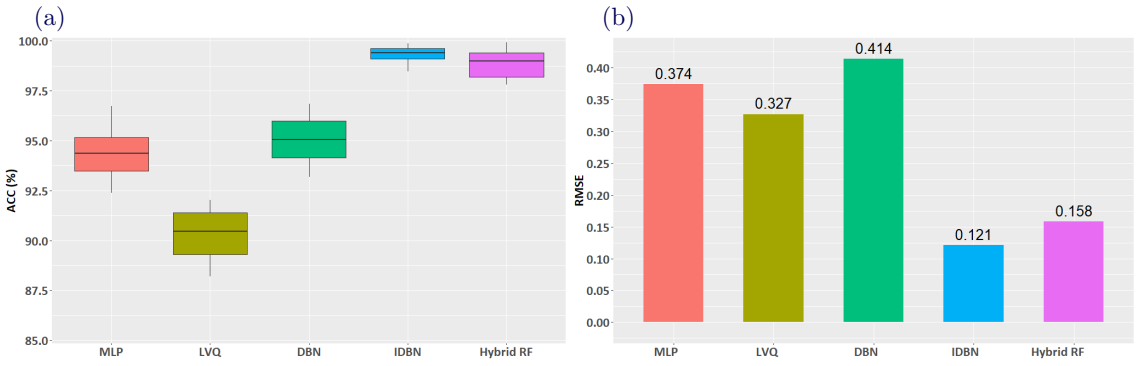


Figure 7.5: (a) ACC of the proposed models on dataset GDS4; (b) RMSE of the proposed models on dataset GDS4

7.4.3 Comparative Results and Discussion

The purposed of this research is to give the development and application of deep belief network as a trend in predicting academic performance. We proposed an optimization approach for improving the classification performance of the proposed IDBN by using feature selection method to obtain the high quality of data, optimizing DBN model with the optimal value of hyper parameters, and applying L2-regularization method for weight decay. To ensure the effective and robustness of the proposed model once we have bigger size of data, then we introduce four datasets of increasing size 1000, 2000, 4000, and 10000, respectively.

The proposed optimization approach boosted up the accuracy of traditional DBN

with accuracy of 10% increased. We observe the performance of the proposed model with larger data size. We observed that once, we have big enough of data size, the proposed models tend to converges and stable. IDBN converges to the superior performance with accuracy of 99.37%. IDBN outperforms the Hybrid RF with respected to robustness and effectiveness in the larger data size. The performance of the four models base on accuracy is shown in Figure 7.6.

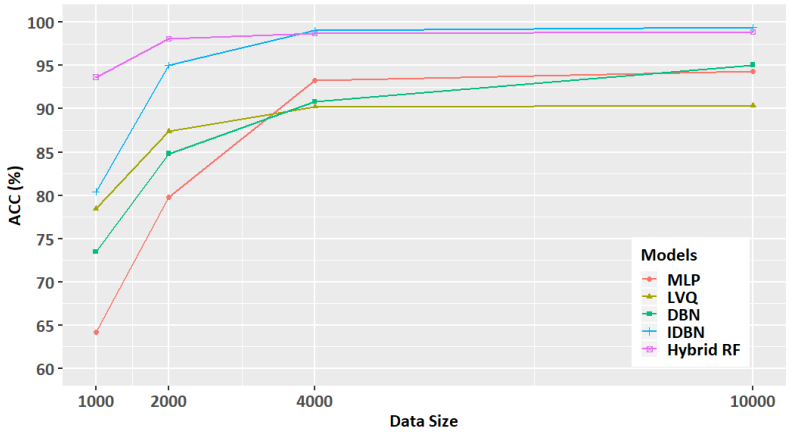


Figure 7.6: Comparative performance based on accuracy of the proposed models with respect to the increasing size of datasets

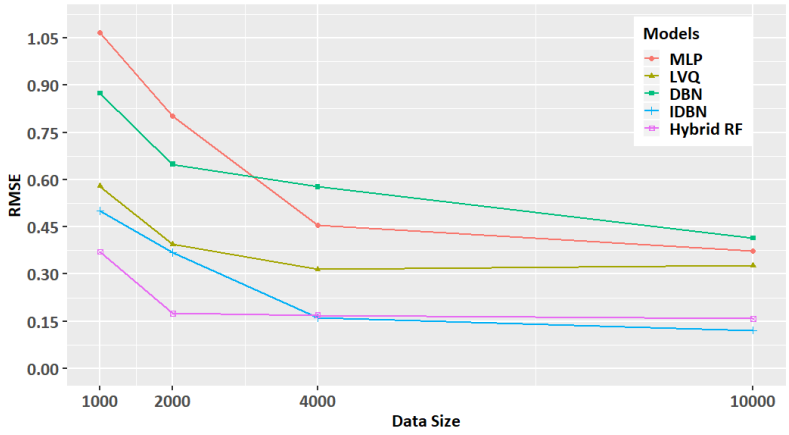


Figure 7.7: Comparative performance based on RMSE of the proposed models with respect to the increasing size of datasets

Simultaneously, Figure 7.7 shows RMSE metric of the proposed model for datasets GDS1, GDS2, GDS3, and GDS4, respectively. In dataset GDS1, Hybrid RF generates the smallest RMSE and follow by LVQ and IDBN as the second and third smallest, respectively. RMSE of LVQ is smaller than IDBN, yet it produced smaller Std. which shows that it is more accurate.

In dataset GDS2, IDBN become the second best performance which generate the second smallest RMSE follow the Hybrid RF. When the dataset increase (GDS3 and GDS4), IDBN generate the smallest RMSE, outperform Hybrid RF, which shown itself as the potential prediction models with the larger size of dataset.

7.5 Chapter Summary: Summary of Experimental Results

The main purpose of this study is to gives the development optimization of DBN applied for educational field. We proposed an optimization approach to improve the performance of DBN for classification and prediction of academic performance. The optimization approach composed of using feature selection method to obtain the high quality of data, optimizing DBN model with the optimal value of hyper parameters, and applying L2-regularization method for weight decay.

The experiment was conducted with two phases. Phase I is the experiment with actual datasets, and Phase II is the experiment with three artificial datasets of increasing size. In the experiment Phase I, IDBN outperforms other model, boosted the traditional DBN with accuracy of 10% increased, and as the second best model follow Hybrid RF. In the experimental phase 2, the comparative study with respect to larger dataset indicated that IDBN could generate the highest ACC, the smallest RMSE, and more accurate and robust comparing Hybrid RF. Hence, in the real case of educational settings, the analysis for the larger datasets will be required, IDBN will be a recommended approach for predicting and classifying the student performance.

CHAPTER 8

DETERMINING DOMINANT FACTORS FOR ACADEMIC PERFORMANCE PREDICTION

8.1 Introduction

Determining the optimal feature subset is important because a reduced dataset will enhance effectiveness in modeling in terms of computer resource saving and suitability for use on a mobile phone. In this study, the optimal feature subset will be more suitable to implement on the Academic Performance Prediction System (APPS) since the users might feel so bored if they have to input so many information or features. The purpose of this chapter to selection the optimal set to use in our prediction system. The purpose of this chapter is to select a dominant feature subsets that can effectively give higher prediction results with smaller set.

Even if several works of literature have studies on using feature selection (FS) algorithms, but still, a lot of attention is required to build academic performance prediction with the analysis and help of FS algorithms. The primary purpose of our study is to introduce a study of FS algorithms to extract the dominant factors that are necessary and sufficient to evaluate the success of students' performance. The study proposed a novel FS algorithm to improve the performance of predictive models. We extract an optimal and effective subset that improves the classification performance of classifiers. Consequently, we can obtain the potential prediction model and the dominant

factors that maximize and control the evaluation of student performance.

It is clear that thanks to today technology, gathering data stops being a problem, and in the focus of interest comes their analysis and obtaining valuable information from the data (knowledge). Central for that problem is a process of knowledge discovering from database (KDD). As the process of knowledge discovery represents a computer processing data from different perspectives in order to extract implicit and interesting patterns, trends or information from data, it can help every participant of educational process in order to improve decision making and optimize performing of the students. By determining dominant factors of student performances, it is important for all participants of educational system, because it can help with better understanding of learning patterns and it focuses on discovering, recognizing and explaining educational phenomenon and ultimately improving them. Hence, the consideration or analysis on the important or highly influencing factors are very important for educational stakeholder to give more focus on those factors to improve student performance. Here, we produced multi approach of technique to obtain the optimal result. That is why the last objective of our study is to introduce the techniques in identify the highly influenced features of the student's performance in high schools.

8.2 Feature Selection Algorithms

Feature selection has been an active and fruitful field of research applications in pattern recognition, machine learning, statistics, and data mining communities [45]. This study proposes methods as a combined-method for boosting up the performance of the predictive models. The study proposed comparative study of four baseline methods, and pick up the best methods as baseline algorithms for the proposed algorithm.

8.2.1 Existing Feature Selection Algorithms

FS is often straightforward when working with real-valued data, such as using the Pearson's correlation coefficient, but can be challenging when working with categorical data. The fours most commonly used methods for categorical input data when the target variable is also categorical (e.g. classification predictive modeling) are the chi-squared statistic and the mutual information statistic [138]. The concept and characteristics of these four methods are mentioned in Table 8.1. The four effective filter-based approaches are utilized.

Table 8.1: Comparison of Feature Selection Algorithms

Consideration	IG	SU	MI	CHI
Concept	mutual information	mutual information	mutual information	Univariate statistical test
Features	categorical	categorical	categorical	categorical
Target	multi-class	multi-class	multi-class	multi-class

Information Gain (IG)

IG is one of the popularly used methods in data mining. IG measures the relevance of a feature by separating the training samples of input features to its target class. The algorithms use the concept of Shannon's entropy in information theory to rank the importance of input features. The feature importance scored by IG is presented in Figure 8.2 (a). The features set was reduced to 29 features.

Symmetric Uncertainty (SU)

SU is one of the leading techniques. SU determines the correlation between a feature and target variable using entropy and information gain theory. The feature importance scored by SU is presented in Figure 8.2 (b). The features set was reduced to 29 features. The algorithm select the same number of features to IG but some differences in rank order.

Chi-square (CHI)

CHI is a statistical method that is utilized for measuring the dependency of each input feature to the target class. The algorithm uses the concept of CHI score of the classes to get the rank list of all attribute. The feature important scored by CHI are presented in Figure 8.2 (c). The features set was reduced to 29 features. The MI algorithm select the same number of features to IG and SU but rearrange the order of feature importance.

Mutual Information (MI)

The mutual information (MI) of two random variables or features is a measure quantifies the dependence measurement between those variables. MI is asymmetric measurement such that can recognize non-linear relationships between variables. The algorithm filter a feature subset of 32 informative predictive features from the original dataset. The feature importance scored by MI is presented in Figure 8.2 (d). The algorithm selects bigger size of feature subset with order of feature importance.

8.2.2 The Proposed Feature Selection Algorithm

Naturally, some feature would be more relevant to the class label more than others while some can be altogether irrelevant. Therefore, it is important to select the features that will be most helpful in identifying the correct class labels in the data at a preliminary stage. In the research literature of feature analysis, there are quite few reliable methods that have been used by researchers to select the optimal feature sets for prediction and classification problems. A major challenge in determining an optimal subset is the inconsistency of ranking the importance of features. Most of the FS algorithms (FSA) such as IG, SU, and MI are mutual information-based methods [214]. These algorithms utilize the concept of MI and information theory [138][215]. CHI is one of the robust methods that it is efficient for any dataset with categorical input features [219].

Algorithm 3 : The Proposed Feature Selection Method: **MICHI** algorithm

1. Step 0: Initilization
   - The training dataset consist of original input feature $S = (F_1, F_2, ..., F_n)$ and target class $C = \{C_1, C_2, C_3, C_4\}$
   - Feature selection algorithms: $FSA = \{MI, CHI\}$
2. Step 1: Scaling the computed scores with normalization:
   - Features subset S_{MI} and S_{CHI} are feature subsets generated by MI and CHI algorithm, respectively.
   - Scale the score of S_{MI} subset: $\widetilde{MI} = \dfrac{MI_i - MI_{min}}{MI_{max} - MI_{min}}$
   - Scale the score of S_{CHI} subset: $\widetilde{CHI} = \dfrac{CHI_i - CHI_{min}}{CHI_{max} - CHI_{min}}$
3. Step 2: Combining the scores:
   - The new feature subsets: $S_{\widetilde{MI}}$ with scaled score $\widetilde{MI}$ and $S_{\widetilde{CHI}}$ with scaled score $\widetilde{CHI}$
   - Generate a score vector combining the two scores: $MICHI = \begin{pmatrix} \widetilde{MI} \\ \widetilde{CHI} \end{pmatrix}$
4. Step 3: Taking magnitude of the score vector:
   - To get the combined-FS score, Euclidean norm is applied : $|MICHI_i| = \sqrt{(MI_i)^2 + (CHI_i)^2}$
   - The new feature subset S_{MICHI} get the new feature set with rearrange order that combine both information of MI and CHI algorithms.

MI and CHI are the two popular and effective FS algorithms; however, we believe that the combined-FS algorithm is better than trusting on a single algorithm. In this study,

we proposed a MICHI method as a combination of MI and CHI algorithms based on the ranked vector score. Since, different methods generate their feature score differently, before combining them, we first normalize the scores of both MI and CHI scores into the same format scale. The procedure of our proposed method is shown in Algorithm 3. MI and CHI are the two best FS methods from the primary checking. Our proposed method combine these two method base on feature rank score. From this combination method, this means that the score of a feature in the MICHI algorithm containing the score vector generated by the MI and CHI algorithms. The manitute of the core vector can be used to compare feature to one another. The new features rearrange the order of importance of features, feature with greater value of $|MICHI_i|$ will be ranked higher. Unlike other ways of combining scores from different methods such as AND and OR, our approach yields a true metric on the space for all pairs of scores [216]. There are experiments showed a minor improvement or no improvement in classification performance when more than three methods were combined [217][218]. This allows for a mathematical structure for analyzing the space of combined scores.

The goal is to combines the score from the two popular s methods in an attempt to develop an enhanced metric for feature filtering. Unifying and normalizing score of MI and CHI measures are to come up with a new feature rank based on the computed scores. This indeed may place features within their true rank and enhance the chances of certain significant feature to being detected for selecting the dominant feature set. Each method filter a new feature subset by dropping noisy and irrelevant features and keep the important features which is beneficial for prediction modeling.

8.3 The Consideration on the Ranking of Feature Importance

This section gives comparison of the feature ranking or feature important that generated by proposed methods to those generated by decision tree models C5.0 and random forest

(RF). The scale of scores weight by each algorithms are somehow different, so we use the rank of features to compare the different rank of a feature on each feature subsets selected by each FS and tree-based models.

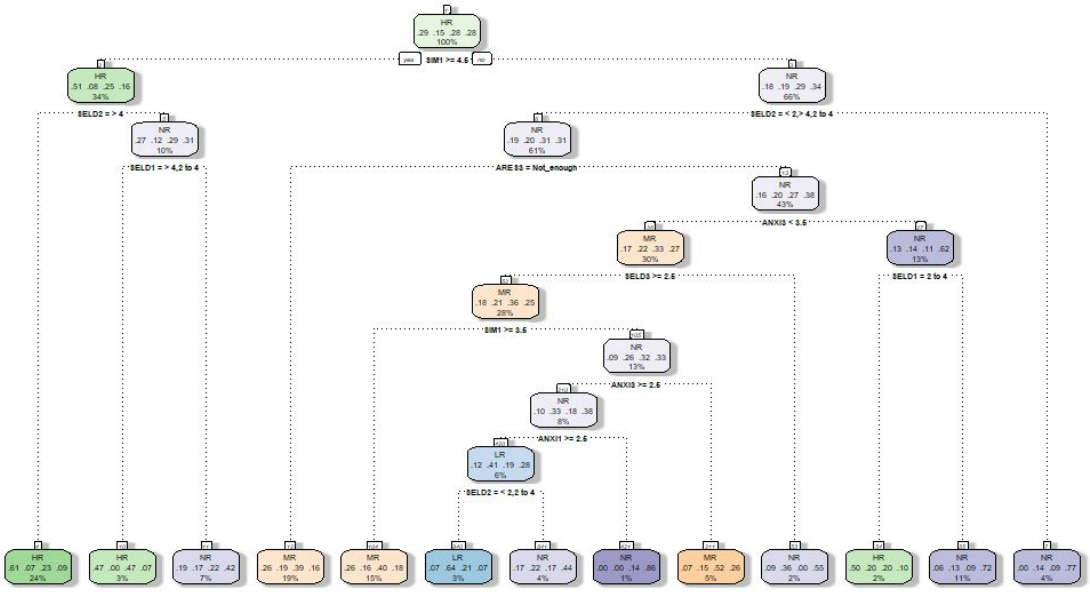


Figure 8.1: Sample of feature importance by splitting rule in decision tree model

A standard tree induced with a decision tree classifier (C5.0 or possibly ID3, C4.5 or CART) consists of a number of branches, one root, a number of nodes and a number of leaves. Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value. The sample is shown in Figure 8.1. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature. RF is an ensemble of trees trained on random samples and random subsets of features. Thus, we can calculate the feature importance for each tree in the forest and then average the importance across the whole forest.

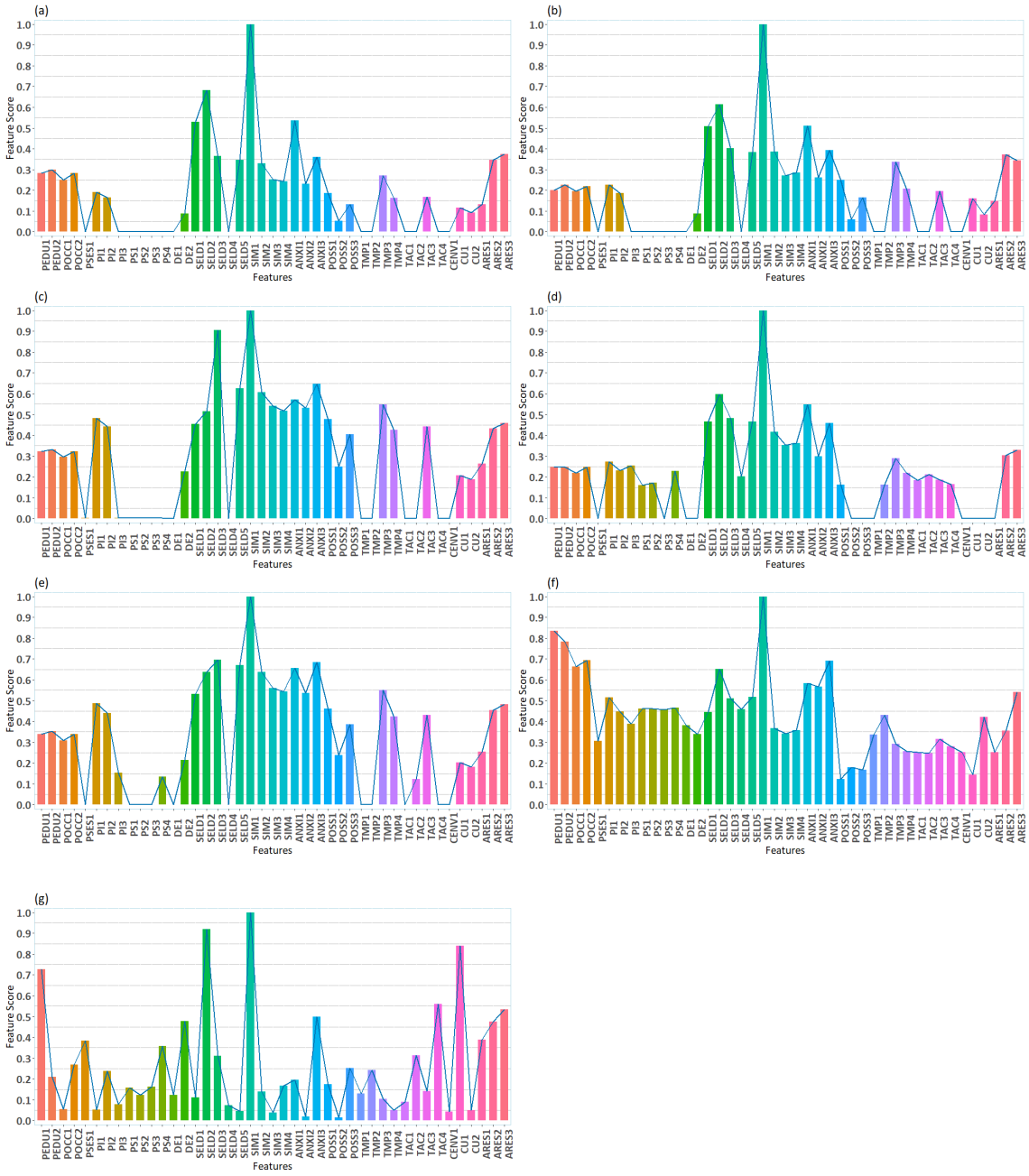


Figure 8.2: Feature Importance Rakning: (a) IG; (b) SU; (c) CHI; (d) MI; (e) MICHI; (f) RF; and (g) DT

This section gives a consideration or analysis on the ranking of feature importance. Recall that the main purpose of this study to extract the optimal feature subsets

that produce the highest performance and extract the dominant factors for educational purposes. The procedure select the feature subsets containing the important factors and rearrange the feature decreasingly for giving analysis on each subset. Hence, the ranking or ordering of the features is consider in this study. The comparison of feature importance ranking is presented in Figure 8.2. For the sake of comparison and analysis, we transform feature score of each algorithm to the same scale. Figure 8.2 (a), (b), and (c) represents the feature scoring of IG, SU, and CHI, respectively. The three algorithm select the same numbers and the same set of important features, however, the trend of feature ranking or feature importance is almost the same, yet it is give a bit different levels of importance for some features in each algorithm. The feature importance generated CHI and MI algorithms are significantly different as shown in Figure 8.2 (c) and (d). Hence, the combined method MICHI generate a new feature rank as shown in Figure 8.2 (e).

The study aims at comparing the ability of the tree-based models and the proposed FS methods. We want to see if the feature ranking or feature importance are the same or not to conclude the model performance of the feature subsets. The feature ranking of C5.0 and RF are represented in Figure 8.2 (f) and (g). The trend of feature ranking in of these two models are similar, yet the there are different for some features. However, the feature sets of C5.0 and RF select all features, while others FS drop some low influencing features. Comparing the ranking of predictive features in C5.0 and RF to those in MICHI, the trend is different. MICHI algorithm drop some irrelevant features as dropped by MI and CHI algorithm. However, the tree-based models give importancy to all features. The experimental results of prediction performance is shown in Chapter 5 and 6, and improve in this chapter with the feature subset selected by these FS methods. This section give comparison of feature ranking generated by the proposed MICHI to C5.0, RF, and four existing FS method.

Table 8.2: The Comparative Ranking on the Importance of Predictive Features

Features	IG	SU	CHI	MI	MICHI	RF	DT
PEDU1	12	20	23	18	23	2	4
PEDU2	11	16	21	16	21	3	19
POCC1	20	15	24	22	24	5	35
POCC2	29	29	26	19	29	4	15
PSES1	30	30	30	38	42	31	11
PI1	19	17	12	15	13	11	36
PI2	22	23	16	20	17	18	18
PI3	31	31	31	17	30	22	33
PS1	32	32	32	32	38	14	24
PS2	33	33	33	28	35	15	28
PS3	34	34	34	39	42	17	23
PS4	35	35	35	21	31	13	12
DE1	36	36	36	36	40	23	29
DE2	28	27	27	41	27	28	8
SELD1	4	4	15	6	12	19	30
SELD2	2	2	11	2	7	6	2
SELD3	6	5	2	4	2	12	14
SELD4	37	37	37	25	33	16	34
SELD5	9	8	4	5	4	10	39
SIM1	1	1	1	1	1	1	1
SIM2	10	7	5	8	6	24	26
SIM3	15	13	8	10	8	27	41
SIM4	17	12	10	9	10	25	22
ANXI1	3	3	6	3	5	7	20
ANXI2	18	14	9	13	11	8	42
ANXI3	7	6	3	7	5	5	7
POSS1	20	15	13	30	15	43	21
POSS2	29	29	26	42	26	42	43
POSS3	24	24	20	34	20	40	16
TMP1	38	38	38	33	39	29	27
TMP2	39	39	39	31	37	20	17
TMP3	14	11	7	14	9	32	31
TMP4	23	19	19	23	19	34	37
TAC1	40	40	40	27	34	35	32
TAC2	41	41	41	24	32	38	13
TAC3	21	21	17	26	18	30	25
TAC4	42	42	42	29	36	33	5
CENV1	43	43	43	43	43	37	40
CU1	26	25	28	37	28	41	3
CU2	27	28	29	40	29	21	38
ARES1	25	26	25	35	25	36	10
ARES2	8	9	18	12	16	26	9
ARES3	5	10	14	11	14	9	6

Note: Blue color denotes feature in rank 1 to 10; cyan in rank 11 to 20; the fuchsia in rank 21 to 32; the black for the less important; and the gray for the irrelevant features which are dropped.

The ranks of predictive features of each feature subsets are summarized in Table 8.2. From the table, we conclude that the feature rank and feature important selected by tree-based models and our proposed methods are different. The optimal set generated by variable importance of tree-based models is the full selected set. Hence, there are two possibilities that experimental results our feature subset selected by proposed method will be poorer or better than on tree-based models. The comparative experimental results are shown in the section 8.5.

8.4 Classification Models

There are two main views to have a good classification in supervised learning. Those are high-quality datasets and classifiers that best suit the given data. In general, there is no best classifier nor best dataset; both are needed to be investigated. There are several educational data mining (EDM) technique that can be used as prediction models but heavily depends on the characteristics of datasets. Once, we have done applying for filtering relevant features, then we focus on considering potential classifiers that are expected to well leverage our dataset. From the our comparative study in Chapter 5, k-Nearest Neighbors (KNN) is one of the lead models and it is kept to study in this chapter since KNN is suggest to be benefit from dataset of lower dimensional obtained from FS. Our development mentioned in Chapter 6, in Hybrid models has developed the four baseline models, Hybrid C5.0 and Hybrid RF are comparatively better better than the other classifiers. We then developed an Improved Deep Belief Network (IDBN) for prediction in the larger datasets in Chapter 7. Hence, we will choose KNN, Hybrid C5.0, Hybrid RF, and IDBN as our superior models.

8.4.1 k-Nearest Neighbors Model

KNN is one of the effective non-parametric EDM models used for classification tasks. KNN is an effective classifier and produces higher classification results in Chapter 5. Like many other classifiers, KNN classifier is noise-sensitive [220]. Its accuracy highly depends on the quality of the training data. Noise and mislabeled data, as well as outliers and overlaps between data regions of different classes, lead to less accurate classification. It is found to perform much better with the dominant set of important features.

8.4.2 The Proposed Hybrid Machine Learning Models

In Chapter 6, we have proposed hybrid machine learning models which are the combination of baseline classifiers (support vector machine (SVM), naïve Bayes (NB), C5.0, and random forest (RF)) with principal component analysis (PCA) and a resmapling method 10-fold cross-validation (10_FCV). Hybrid C5.0 and Hybrid RF were found to be the best classifiers in our classification problem.

8.4.3 The Proposed Improved Deep Belief Network Model

In Chapter 7, we proposed another prediction model IDBN. IDBN is the optimization approach of deep belief network (DBN) model. We proposed an optimization approach composed of (i) method, (ii) optimization of hyper-parameter, and (ii) Regularization method. The developed model has introduced. The improved model was found to produce the most classification results when using larger datasets.

8.5 Experimental Results

This section reported the performance evaluation of methods in selecting the dominant factors for predicting academic performance. We executed the proposed our optimal classifiers using subsets that were obtained from each method. After applying the

FS algorithms, from the original datasets each algorithm captures the a subset of top N features descent from the most important to less important. FS algorithm selects the relevant factors to the target variables, then we rank the feature weight denoting the importance of features from sets selected by each FS algorithm decreasingly. *Dominant factors* or *Dominant set* is defined as the optimal set of predictive features that are necessary and sufficient to evaluate the success of student performance. The dominant set was extracted for two purposes: improving prediction model and giving informative learning patterns for intervention. The feature subset D of a full selected set S is called dominant set if the performance p of a prediction model on D with respect to target C produce higher result i.e $p(D|C) \geq p(S|C)$. The experiment was proposed to get both best classifier and the dominant set, as shown in Figure 8.3.

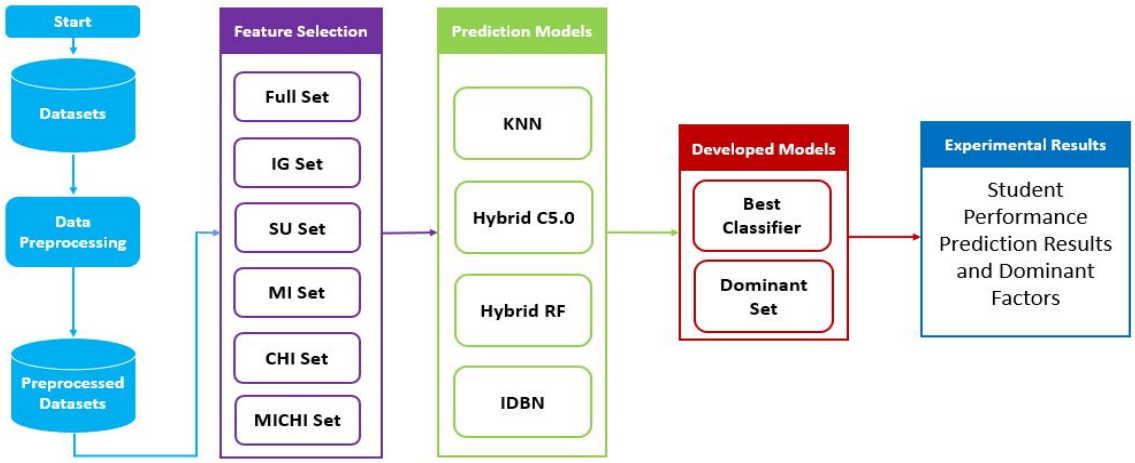


Figure 8.3: Illustration of proposed framework

Once, each FS algorithm done on selecting its feature subset, then we rearrange the feature subsets from the most important feature decreasingly to the lest important feature base on feature scores, as shown in Table 8.3. Different FS algorithms generate different ranges of feature subsets. The purpose of this study is to extract the feature subset containing the most important features that produce the highest classification performance.

Table 8.3: Feature Ranking of the Proposed Feature Selection Methods

Rank	IG		SU		CHI		MI		MICHI	
1	SIM1	1.000	SIM1	1.000	SIM1	1.000	SIM1	1.000	SIM1	1.000
2	SELD2	0.681	SELD2	0.615	SELD3	0.905	SELD2	0.599	SELD3	0.697
3	ANXI1	0.536	ANXI1	0.510	ANXI3	0.646	ANXI1	0.547	ANXI3	0.684
4	SELD1	0.530	SELD1	0.509	SLED5	0.626	SELD3	0.481	SELD5	0.670
5	ARES3	0.375	SELD3	0.402	SIM2	0.607	SELD5	0.467	ANXI1	0.656
6	SELD3	0.366	ANXI3	0.393	ANXI1	0.572	SELD1	0.466	SIM2	0.637
7	ANXI3	0.361	SIM2	0.385	TMP3	0.549	ANXI3	0.459	SELD2	0.636
8	ARES2	0.347	SELD5	0.383	SIM3	0.540	SIM2	0.416	SIM3	0.560
9	SELD5	0.346	ARES2	0.371	ANXI2	0.531	SIM4	0.363	TMP3	0.550
10	SIM2	0.330	ARES3	0.345	SIM4	0.518	SIM3	0.354	SIM4	0.544
11	PEDU2	0.300	TMP3	0.336	SELD2	0.515	ARES3	0.330	ANXI2	0.356
12	PEDU1	0.283	SIM4	0.285	PI1	0.481	ARES2	0.305	SELD1	0.531
13	POCC2	0.282	SIM3	0.271	POSS1	0.478	ANXI2	0.298	PI1	0.486
14	TMP3	0.271	ANXI2	0.262	ARES3	0.45	TMP3	0.285	ARES3	0.482
15	SIM3	0.252	POSS1	0.250	SELD1	0.454	PI1	0.275	POSS1	0.461
16	POCC3	0.249	PEDU2	0.227	PI2	0.442	PEDU2	0.255	ARES2	0.454
17	SIM4	0.243	PI1	0.226	TAC3	0.442	PI3	0.249	PI2	0.441
18	ANXI2	0.232	POCC2	0.220	ARES2	0.434	PEDU1	0.249	TAC3	0.432
19	PI1	0.192	TMP4	0.209	TMP4	0.426	POCC2	0.248	TMP4	0.423
20	POSS1	0.187	PEDU1	0.200	POSS3	0.404	PI2	0.232	POSS3	0.387
21	TAC3	0.167	TAC3	0.197	PEDU2	0.331	PS4	0.229	PEDU2	0.353
22	PI2	0.165	POCC1	0.196	POCC2	0.323	POCC1	0.219	POCC2	0.340
23	TMP4	0.164	PI2	0.187	PEDU1	0.322	TMP4	0.218	PEDU1	0.340
24	POSS3	0.133	POSS3	0.166	POCC1	0.297	TAC2	0.213	POCC1	0.309
25	ARES1	0.133	CU1	0.166	ARES1	0.297	SELD4	0.213	ARES1	0.255
26	CU1	0.117	ARES1	0.149	POSS2	0.251	TAC3	0.187	POSS2	0.238
27	CU2	0.092	DE2	0.088	DE2	0.226	TAC1	0.185	DE2	0.215
28	DE2	0.088	CU2	0.083	CU1	0.209	PS2	0.173	CU1	0.202
29	POSS2	0.053	POSS2	0.059	CU2	0.190	TAC4	0.165	CU2	0.182
30	PSES1	0	PSES1	0	PSES1	0	POSS1	0.164	PI3	0.154
31	PI3	0	PI3	0	PI3	0	TMP2	0.082	PS4	0.092
32	PS1	0	PS1	0	PS1	0	PS1	0.051	TAC2	0.024
33	PS2	0	PS2	0	PS2	0	TMP1	0	SELD4	0
34	PS3	0	PS3	0	PS3	0	POSS3	0	TAC1	0
35	PS4	0	PS4	0	PS4	0	ARES1	0	PS2	0
36	DE1	0	DE1	0	DE1	0	DE1	0	TAC4	0
37	SELD4	0	SELD4	0	SELD4	0	CU1	0	TMP2	0
38	TMP1	0	TMP1	0	TMP1	0	PSES1	0	PS1	0
39	TMP2	0	TMP2	0	TMP2	0	PS3	0	TMP1	0
40	TAC1	0	TAC1	0	TAC1	0	CU2	0	DE1	0
41	TAC2	0	TAC2	0	TAC2	0	DE2	0	PSES1	0
42	TAC4	0	TAC4	0	TAC4	0	POSS2	0	PS3	0
43	CENV1	0	CENV1	0	CENV1	0	CENV1	0	CENV1	0

The experiment was carried out with two phases. The first experiment was executed with the dataset ADS comprised of 1204 samples. The second experiment was with a larger dataset GDS4 comprises 10000 samples. The second experiment was carried out with the context of the larger dataset to confirm the performance of IDBN and other proposed models. The experiment was made a minimal subset of 5 features to a fully selected set. To evaluate and compare the performance of prediction models, accuracy (ACC) and root mean square error (RMSE) are measured. Recall that the higher value of ACC, the better model is. In contrast to ACC, the smaller value of RMSE, the better model is.

8.5.1 Experiment Results with ADS Dataset

Table 8.4 shows the performance of the four classifiers using the original datasets (without FS). The table illustrates the staitistcal results of the two metrics of KNN, Hybrid C5.0, Hybrid RF, and IDBN. The average of ACC and RMSE from serveral iteration run are recorded. The two developed tree-based models, Hybrid C5.0 and Hybrid RF generate the highest ACC and lowest RMSE.

Table 8.4: The Model Performance on Dataset ADS without Feature Selection

Models	KNN	Hybrid C5.0	Hybrid RF	IDBN
ACC	94.55	99.15	99.27	83.14
Std.	0.801	0.606	0.357	0.640
RMSE	0.261	0.073	0.041	0.659
Std.	0.016	0.045	0.029	0.031

Figure 8.4 to 8.7 illustrate the performance of proposed models on each set from each FS method. The experiment was made subset of 5 features to full selected set. Table 8.5 to Table 8.8 reports the results of the two evaluation metrics of the classifiers on full selected sets and dominant sets.

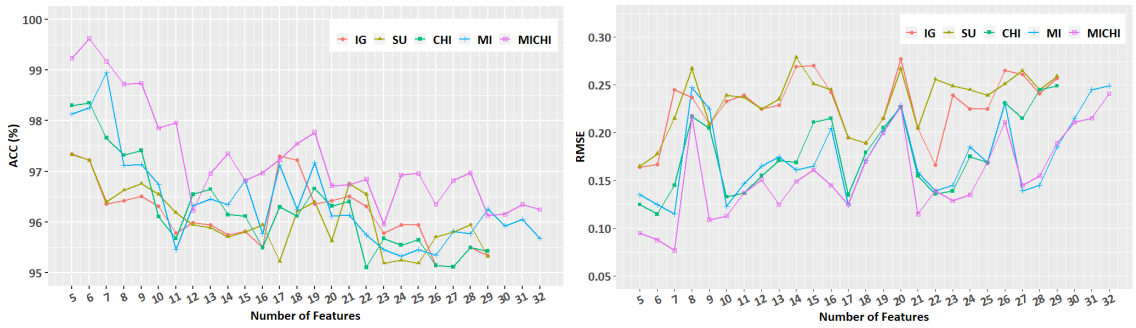


Figure 8.4: (a) Accuracy of KNN (b) RMSE of KNN

From Figure 8.4, we can obtain the results of ACC and RMSE of KNN model using selected set of each FS method, and it is summarized in Table 8.5.

Table 8.5: The KNN Performance on Dataset ADS with Feature Selection

FSA	Full Selected Set			Dominant Set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	95.34	0.257	5	97.34	0.163
SU	29	95.32	0.259	5	97.33	0.164
CHI	29	95.43	0.249	6	98.35	0.115
MI	32	95.68	0.241	7	98.94	0.077
MICHI	32	96.25	0.179	6	99.62	0.063

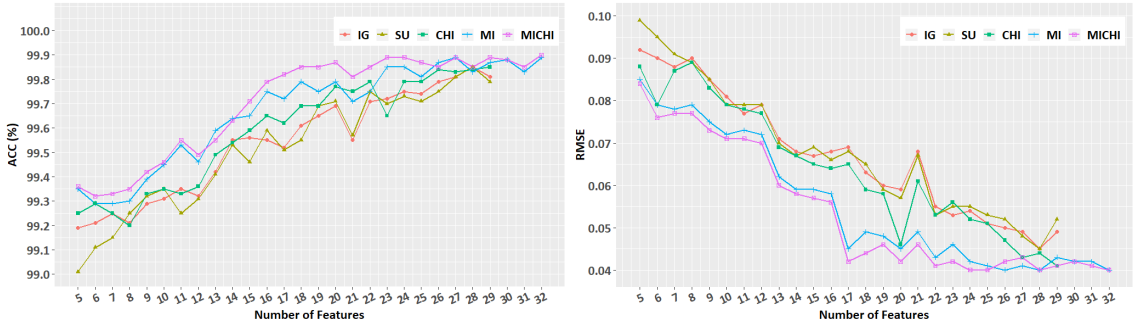


Figure 8.5: (a) Accuracy of the Hybrid C5.0 (b) RMSE of the Hybrid C5.0

Table 8.6: Hybrid C5.0 Performance on Dataset ADS with Feature Selections

FSA	Full Selected Set			Dominant Set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.81	0.049	28	99.85	0.045
SU	29	99.79	0.051	28	99.85	0.045
CHI	29	99.85	0.041	29	99.85	0.041
MI	32	99.89	0.037	32	99.89	0.037
MICHI	32	99.90	0.035	32	99.90	0.035

From Figure 8.5, we can obtain the results of ACC and RMSE of Hybrid C5.0 model using selected set of each FS method, and it is summarized in Table 8.6.

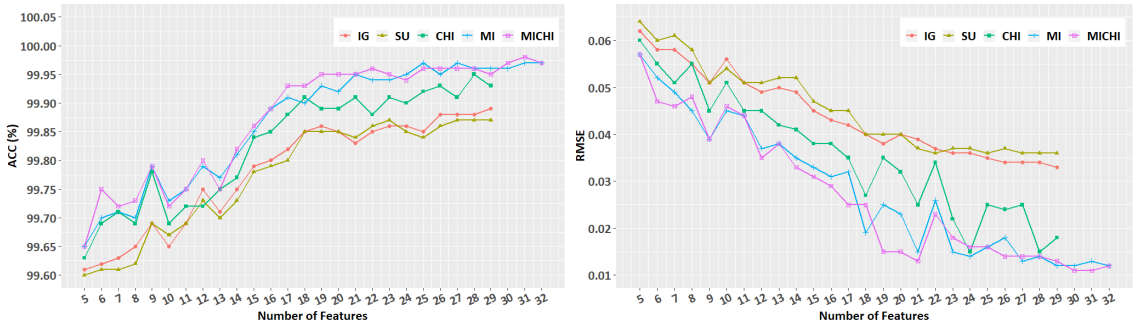


Figure 8.6: (a) Accuracy of the Hybrid RF (b) RMSE of the Hybrid RF

From Figure 8.6, we can obtain the results of ACC and RMSE of Hybrid RF model using selected set of each FS method, and it is summarized in Table 8.7.

Table 8.7: Hybrid RF Performance on Dataset ADS with Feature Selections

FSA	Full Selected Set			Dominant Set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.89	0.033	29	99.89	0.033
SU	29	99.87	0.036	28	99.87	0.036
CHI	29	99.95	0.015	29	99.95	0.015
MI	32	99.96	0.013	32	99.96	0.013
MICHI	32	99.97	0.012	31	99.98	0.011

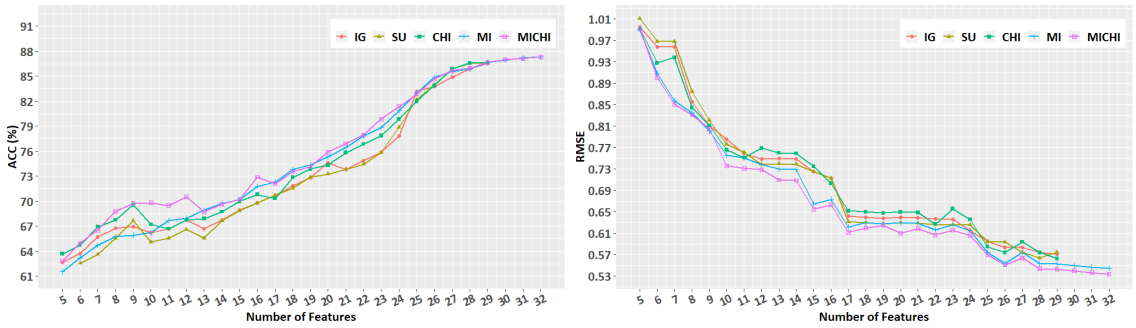


Figure 8.7: (a) Accuracy of the IDBN (b) RMSE of the IDBN

From Figure 8.7, we can obtain the results of ACC and RMSE of IDBN model using selected set of each FS method, and it is summarized in Table 8.8.

Table 8.8: The IDBN Performance on Dataset ADS with Feature Selections

FSA	Full Selected Set			Dominant Set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	86.55	0.571	29	86.55	0.571
SU	29	86.54	0.575	29	86.54	0.575
CHI	29	86.67	0.563	29	86.67	0.563
MI	32	87.11	0.545	32	87.11	0.545
MICHI	32	87.32	0.534	32	87.32	0.534

The results presented in Table 8.5 demonstrate the performance of KNN model on feature sets and dominant sets selected by IG, SU, CHI, MI, and CHIMI methods. The models work best with the low dimension of the most important features. KNN model with the dominant set of the proposed MICHI achieves the highest performance. The ACC and RMSE are improved to 99.62% and 0.063, respectively.

From Table 8.6, the performance of the Hybrid C5.0 is significantly improved when using the dominant set of the MICHI. The ACC and RMSE of the Hybrid C5.0 are improved to 99.90% and 0.033, respectively.

The results of ACC and RMSE of the Hybrid RF are shown in Table 8.7. The proposed MICHI method outperforms the rest FS methods in achieving the highest ACC

and lowest RMSE. The ACC and RMSE of the Hybrid RF are improved to 99.98% and 0.011, respectively.

Table 8.8 demonstrates the performance of the developed IDBN classifier with the input feature sets selected by the five FS methods. The ACC and RMSE of the IDBN are improved to 87.32% and 0.514, respectively.

8.5.2 Experiment Results with GDS4 Dataset

In this subsection, we do our experiment with artificial dataset of larger size of 10000 samples, dataset GDS4. We want to compare the performance of IDBN with other models using subsets of FS algorithms. The performance of proposed models with original dataset is shown in Table 8.9.

Table 8.9: The Model Performance on Dataset GDS4 without Feature Selection

Models	KNN	Hybrid C5.0	Hybrid RF	IDBN
ACC	95.12	98.55	98.88	97.01
Std.	0.942	0.578	0.312	0.666
RMSE	0.093	0.163	0.161	0.195
Std.	0.016	0.049	0.028	0.015

Table 8.9 illustrate the staitistcal results of the two metrics of KNN, Hybrid C5.0, Hybrid RF, and IDBN. The average of ACC and RMSE from serveral iteration run using GS4 dataset are recorded. The two developed tree-based models, Hybrid C5.0 and Hybrid RF generate the highest ACC and lowest RMSE, followed by the IDBN model.

Figure 8.8 to 8.11 illustrate the performance of proposed models on each set from each FS method. The experiment was made subset of 5 features to full selected set. Table 8.10 to Table 8.13 reports the results of the two evaluation metrics of the classifiers on full selected sets and dominant sets, which are implied from Figure 8.8 to 8.11, respectively.

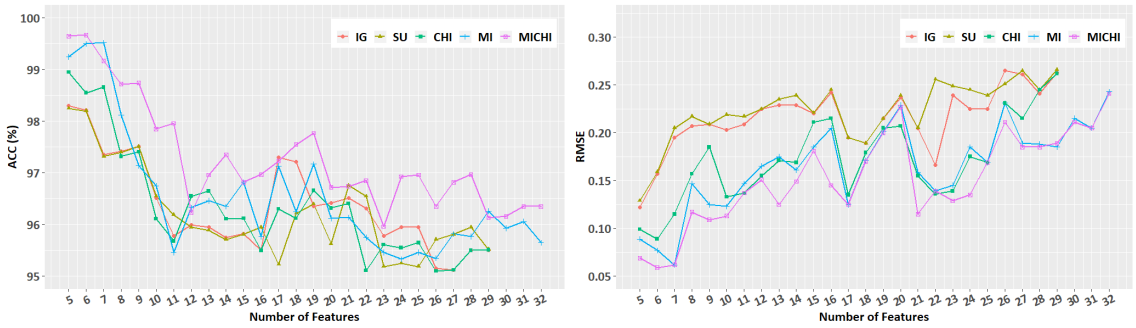


Figure 8.8: (a) Accuracy of KNN (b) RMSE of KNN

From Figure 8.8, we can obtain the results of ACC and RMSE of KNN model using selected set of each FS method, and it is summarized in Table 8.10.

Table 8.10: KNN Performance on Dataset GDS4 with Feature Selections

FSA	Full Selected Set			Dominant Set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	95.50	0.263	5	98.30	0.122
SU	29	95.52	0.266	5	98.25	0.129
CHI	29	95.51	0.262	6	98.95	0.089
MI	32	95.65	0.243	7	99.52	0.062
MICHI	32	96.35	0.175	6	99.67	0.059

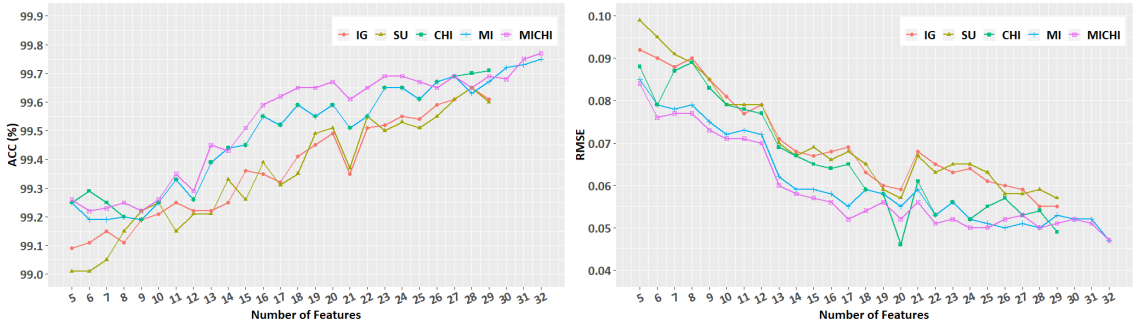


Figure 8.9: (a) Accuracy of the Hybrid C5.0 (b) RMSE of the Hybrid C5.0

From Figure 8.9, we can obtain the results of ACC and RMSE of Hybrid C5.0 model using selected set of each FS method, and it is summarized in Table 8.11.

Table 8.11: Hybrid C5.0 Performance on Dataset GDS4 with Feature Selections

FSA	Full Selected Set			Dominant Set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.61	0.059	28	99.65	0.055
SU	29	99.60	0.058	28	99.63	0.057
CHI	29	99.71	0.049	29	99.71	0.051
MI	32	99.75	0.047	32	99.75	0.047
MICHI	32	99.75	0.045	32	99.75	0.045

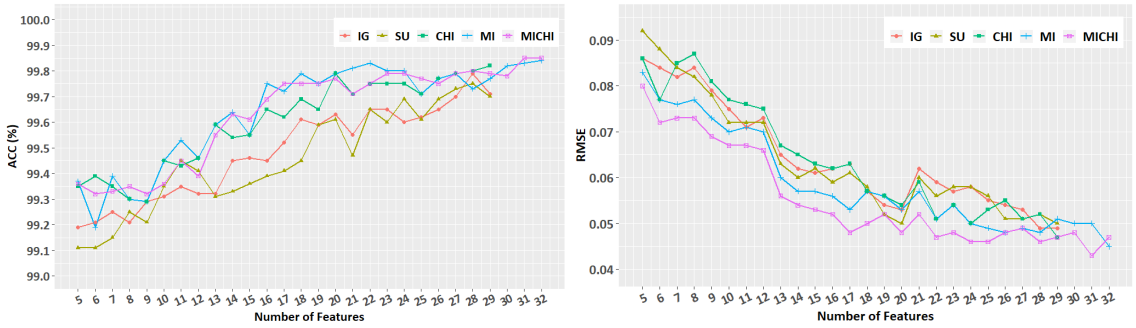


Figure 8.10: (a) Accuracy of the Hybrid RF (b) RMSE of the Hybrid RF

From Figure 8.10, we can obtain the results of ACC and RMSE of Hybrid RF model using selected set of each FS method, and it is summarized in Table 8.12.

Table 8.12: Hybrid RF Performance on Dataset GDS4 with Feature Selections

FSA	Full Selected Set			Dominant Set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.73	0.051	28	99.79	0.049
SU	29	99.73	0.052	28	99.75	0.050
CHI	29	99.82	0.047	29	99.82	0.047
MI	32	99.84	0.045	32	99.84	0.045
MICHI	32	99.85	0.043	32	99.85	0.043

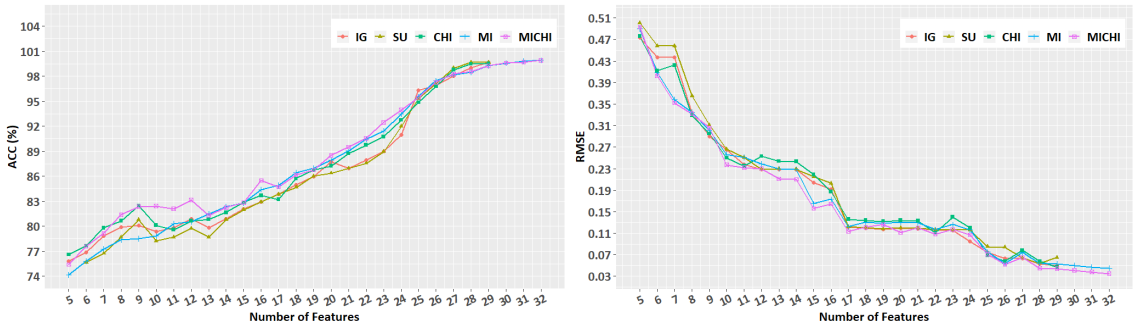


Figure 8.11: (a) Accuracy of the IDBN (b) RMSE of the IDBN

From Figure 8.11, we can obtain the results of ACC and RMSE of IDBN model using selected set of each FS method, and it is summarized in Table 8.13.

Table 8.13: IDBN Performance on Dataset GDS4 with Feature Selections

FSA	Full Selected Set			Dominant Set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.65	0.052	28	99.67	0.050
SU	29	99.65	0.052	28	99.67	0.050
CHI	29	99.77	0.047	29	99.77	0.047
MI	32	99.79	0.046	32	99.79	0.046
MICHI	32	99.83	0.044	32	99.83	0.044

The results presented in Table 8.10 demonstrate the performance of KNN models on feature sets and dominant sets selected by IG, SU, CHI, MI, and CHIMI methods on the GDS4 dataset. KNN model with the dominant set of the proposed MICHI achieves the highest performance. The ACC and RMSE of KNN are improved to 99.67% and 0.059.

From Table 8.11, the performance of the Hybrid C5.0 is significantly improved when using the dominant set of the MICHI. The ACC and RMSE of the Hybrid C5.0 are improved to 99.75% and 0.045, respectively.

The results of ACC and RMSE of the Hybrid RF are shown in Table 8.12. The proposed MICHI method outperforms the rest FS methods in achieving the highest ACC and lowest RMSE. The ACC and RMSE of the Hybrid RF are improved to 99.85% and

0.043, respectively.

Table 8.13 demonstrates the performance of the developed IDBN classifier with the input feature sets selected by the five FS methods. The ACC and RMSE of the IDBN are improved to 99.83% and 0.044, respectively.

8.6 Discussion

This study aims to boost up the performance of the proposed classifiers to reach the most classification results. Hence, the optimal models are then combined with dominant sets, which is believed to significantly improve the performance of prediction models and selected the highly influencing factors in academic performance.

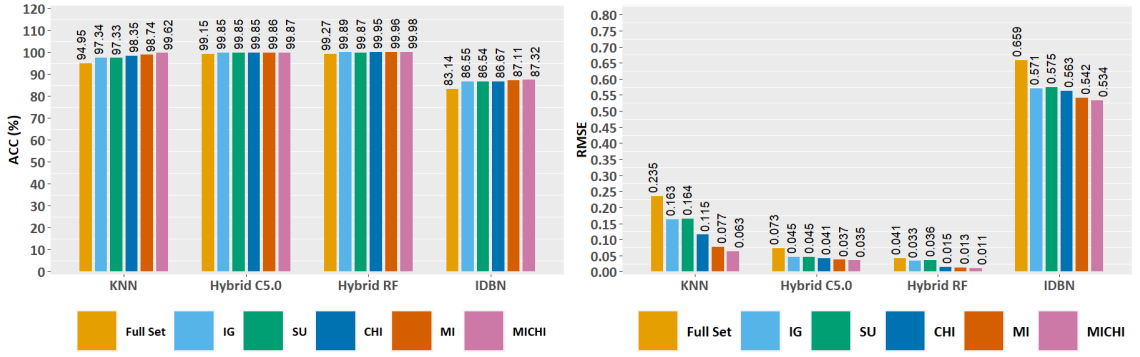


Figure 8.12: ACC and RMSE comparison using ADS dataset

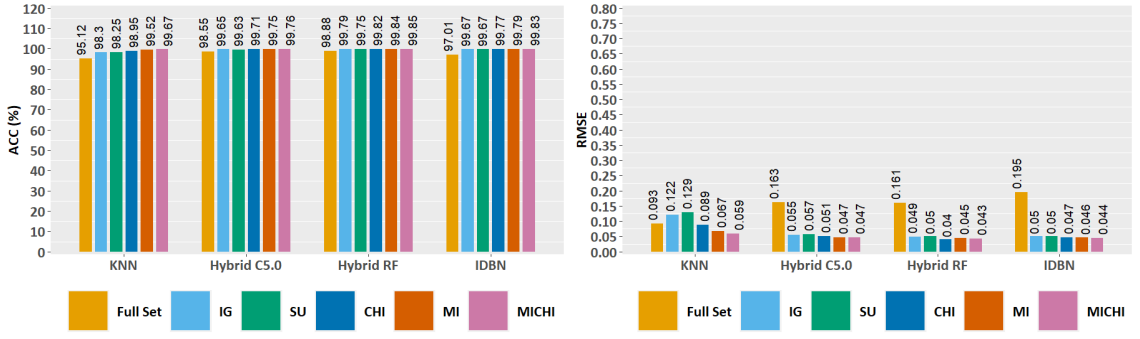


Figure 8.13: ACC and RMSE comparison using GDS4 dataset

From Table 8.4 to Table 8.13, we can summarize the performance of KNN, Hybrid C50, Hybrid RF, and IDBN on dominant sets of IG, SU, CHI, MI, and CHIMI methods. Figure 8.12 and Figure 8.13 summarize the value of ACC and RMSE of each classifier on each FS method on data ADS and GDS4, respectively.

The analysis of dominant subsets is considered in our analysis of the performance of our prediction problem. Determining the dominant set of input features significantly impacts the performance of prediction models. The experimental results indicate that working on the full selected sets containing only important feature, we oftentimes can improve the classification performance. However, different FS algorithm treats dataset differently. Different algorithms select different numbers of features and/or the same number but different features. The individual method selects individual rank of feature subsets. IG and SU relatively improve the performance of classifiers while MI and CHI perform comparatively better with our dataset. The proposed MICHI is the most successful FS algorithm for selecting the dominant sets. We find that the performance of each FS algorithm performs differently on each classifier.

Table 8.14: Comparison of Optimal Classifiers

Size and Dimension	KNN	Hybrid C5.0	Hybrid RF	IDBN
Dominant set size	Smaller	Any subset	Any subset	Higher
Dataset optimal size	Medium up	Medium up	Medium up	Larger

From the experimental result, we can conclude that using full selected sets from methods may not always help to improve classification performance. It may still contain any less relevant features that simultaneously generate noise in the experiment and highly affect some classifiers. Therefore, the improvement depends on determining the dominant set, and the suitable combination of method with classifier. The proposed MICHI could select the best optimal set that effectively improves the performance of classifiers. The combination of the dominant set of each FS method significantly improve the performance

of KNN, especially with the smaller set of important features.

By considering on the dominant sets, the Hybrid RF and IDBN with MICHI algorithm generate the most successful classification result. Hybrid RF perform comaratively better than IDBN when dealing with medium data size. However, with more complicated dataset with larger size, IDBN is one of the effective prediction models. When the information record systems in the secondary education is effective, the it is more comfortable to get larger size of data set and the increasing of dimension of factors affecting student performance, the proposed IDBN will be an considered model for predicting student performance.

8.7 Chapter Summary: Summary of Experimental Result

Determining the dominant set of input feature significantly impacts on the performance of prediction models. The experimental results indicate that working on the full selected sets containing only important feature, we mostly can improve the classification performance. However, different FS algorithm treats dataset differently. Different algorithms select different numbers of features and also give the different rank of feature subsets. IG and SU relatively improve the performance of classifiers while MI and CHI and MI perform comparatively better than IG and SU. The proposed MICHI is the most successful FS algorithm for improving the performance of prediction models. We observe that the performance of each FS algorithm performs differently on each classifier.

We proposed a comparative study of EDM of prediction and classification task, the outperformed prediction models are then developed and optimized to get the most successful classification results. The comparative experiment of four classifiers (KNN and Hybrid C5.0, Hybrid RF, and IDBN) is carried out using feature sets from five FS methods (IG, SU, CHI, MI, and Proposed MICHI method). The analysis of dominant factors is cooperated and combined with the best classifier. The dominant set obtained

from the proposed MICHI algorithm significantly improves the performance of prediction models and used as a set of highly influencing factors that need to be considered for intervention and improvement of student performance. The experimental outcomes indicate that Hybrid RF and IDBN perform comparatively better than the other models. The developed prediction models (Hybrid RF and IDBN) and proposed FS method (MICHI) are integrated into the web-based prediction system.

CHAPTER 9

THE DESIGN AND IMPLEMENTATION OF ACADEMIC PERFORMANCE PREDICTION SYSTEM

9.1 Introduction

The aim of Academic Performance Prediction System (APPS) is to monitor the learning process of students, identify at-risk group or poor-performing group of students, and help teachers, educators, educational stakeholder to develop improvement strategies. Educational stakeholder such as educators and institutes wish to get accurate forecasts in the timely stage to reduce the failure rate of students, however, it is difficult for them to traditionally analyze students' previous learning behaviors to estimate student learning outcomes.

Most researchers have worked on evaluating performances of students in higher education, yet the study on high school student performance evaluation is less. High school student performance is a significant indicator of developing the academic sector since it concerns the background knowledge of students for secondary education and higher education. To improve the poor performance of students in high schools, the right intervention and improvement must be made to low-performing students who are considered in the risk of failure. Poor performing students are highly possible to fail in the

national exam and find themselves harder to survive in university life. In the context of academic poor performance, educational data mining (EDM) is used for timely prediction for intervention and improvement. Many literature of related work have proposed EDM techniques to predict student performance and extract useful information for educational data. However, the performance of proposed methods are limited with acceptable performance of classification. In this study, we proposed developed models of EDM to get superior classification results and subsequently integrate the optimal model into a web-based system for predicting high school student performance. So that educational stakeholder and related individual can utilized and predict student performance more effectively for improving academic performance.

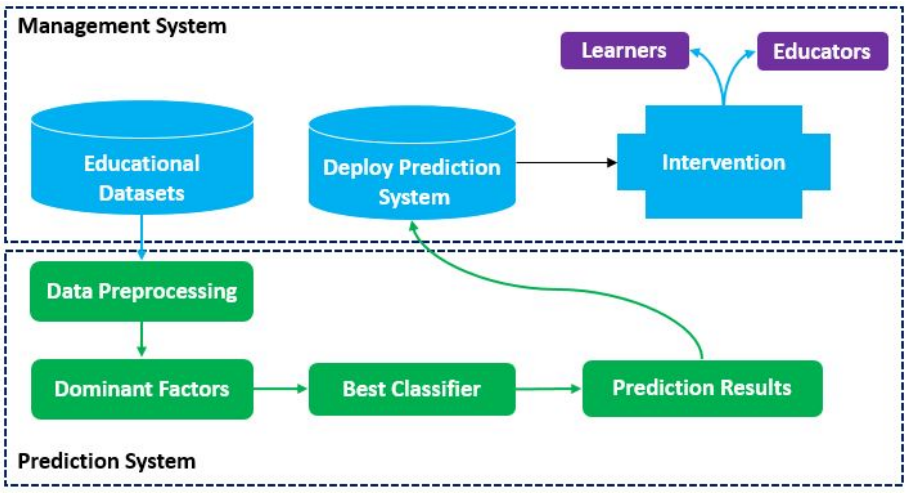


Figure 9.1: The framework of the use of an academic prediction system for educational settings.

9.2 Academic Performance Prediction Systems Modeling

Even if there were some existing works have proposed early perdition systems using popular algorithms in EDM in higher education or online courses, but still, a lot of attention is needed to build an academic performance prediction system with the

analysis and help of developed prediction methods in EDM to get high and superior classification results. This work proposes a study of developing an APPS to predict student performance in high schools. The study compost of the selection of informative data, a proposed feature selection method, developed EDM models, and a web-based support system APPS for timely-intervention to poor-performing students.

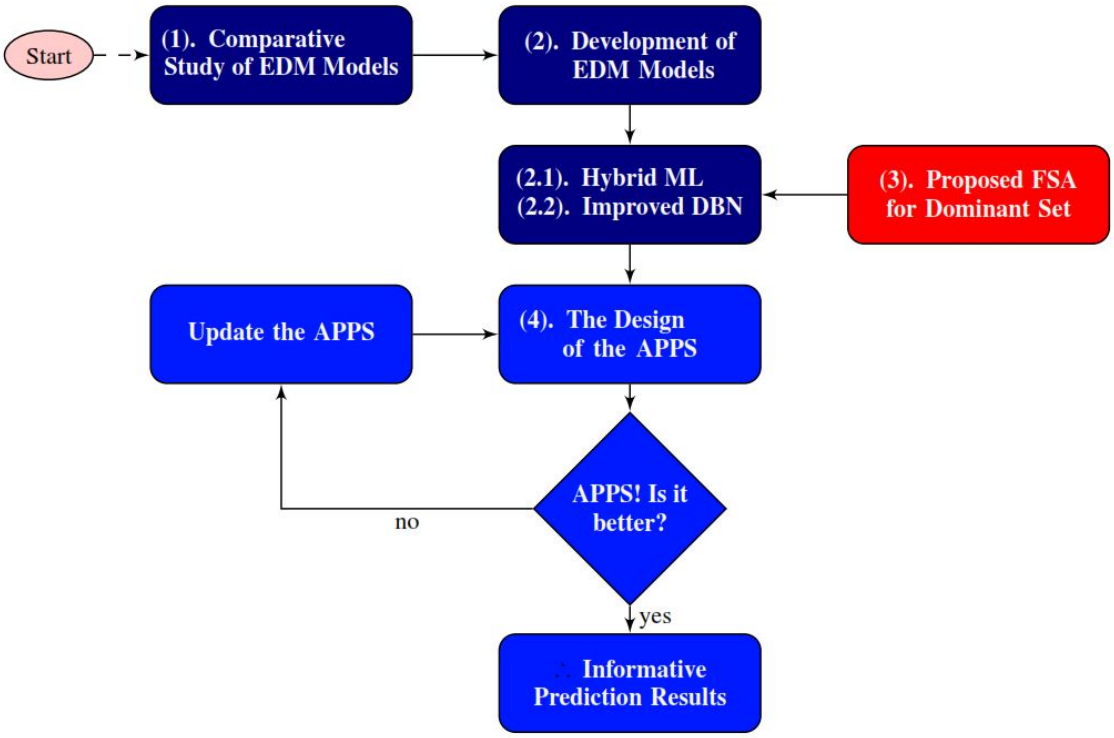


Figure 9.2: Prediction modeling of APPS

This section presents the operation flow in developing APPS. The study introduces a prediction modeling initially proposed for developing a web-based support system for predicting academic performance and it is shown in Figure 9.2. This section summarizes the process of our modeling for predicting student learning outcomes and giving intervention.

9.3 Models for Development of the System

9.3.1 Optimal Feature Subset

The performance of children, adult, adolescent can be affects by many influencing factors, especially external factors, motivation, and longitudinal factors. The main factors for students to be success, in fact, it is not about cleverness or IQ, but about discipline, motivation, and passion, which affect by environment around (themselves, parents, educators, and friends). We observe all possibly-relevant factors that highly influence, less influence, or not irrelevant to students' outcomes. Our input features are obtained for several factors that are categorized into three main factors: home or domestic factors, student or individual factors, and school factors. The total features of datasets are 43 features. There may be a larger volume of features that possibly effect student performance, however, it is not convenient and comfortable to input all the features in the prediction system. Hence, we wish to obtain the less dimension of dataset consisting of informative feature. Feature selection is a popular technique in data mining that is used to accomplish this purpose. The main objectives of feature selection are (i) to improve predictive accuracy, (ii) to remove redundant features and (iii) to reduce time consumption during analysis.

This study proposed feature selection methods to gain informative features. By gaining the optimal features subsets, it can really improve the performance of prediction models. There are three main approaches of feature selection, filter methods, wrapper methods, and embed/hybrid methods. The four effective feature approaches of filter methods are introduced. The comparative study of feature selection methods and a developed FS methods was analyzed is studied in Chapter 8. The developed MICHI method, which is the combination of MI and CHI algorithm based on feature scored was adopted due to its out-performance. The methods select the dominant factors for improving prediction and identifying factors for intervention purpose.

9.3.2 Optimal Classifier

Several EDM techniques were considered in the comparative study in Chapter 5. Three classes of statistical analysis techniques (predictive structural equation modeling), machine learning techniques (5 algorithms), and a deep learning framework called deep belief network were executed and compared in Chapter 5. The random forest was found to be the best model. The improvement of prediction models was then further improved in Chapter 6. We proposed the hybrid models by combining four baseline models (C5.0, random forest (RF), naïve Bayes (NB), and support vector machine with radial basis kernel function (SVM)) were combined with principal component analysis (PCA) and a resampling method 10-fold cross-validation (10_FCV). The hybrid RF was found to be the best model. The optimization approach of deep belief network is proposed in Chapter 7. The study discussed the merit of improved Deep Belief Network (IDBN) for predicting academic performance in context of larger dataset. In Chapter 8, the experimental results of show that IDBN and Hybrid RF are the outperform other models for predicting academic performance. Hybrid RF perform relatively better than IDBN in when combine dominant set of MICHI algorithm. However, these two models can be alternatively used in term of larger and complicated dataset. In this study, Hybrid RF is used as prediction model integrate in the system.

9.4 The Design of Web Application System

9.4.1 Component of APPS for Users

The web-based application linked to the server via the internet as illustrated in Figure 9.3. Hence, what users need are computer and internet to access to the web page of APPS. The educational stakeholder can assess the application and the menu or introduction of how to obtain the prediction results are demonstrated in APPS. The users can upload their data to obtain the results and related information such as dominant factors and

statistics of student data.

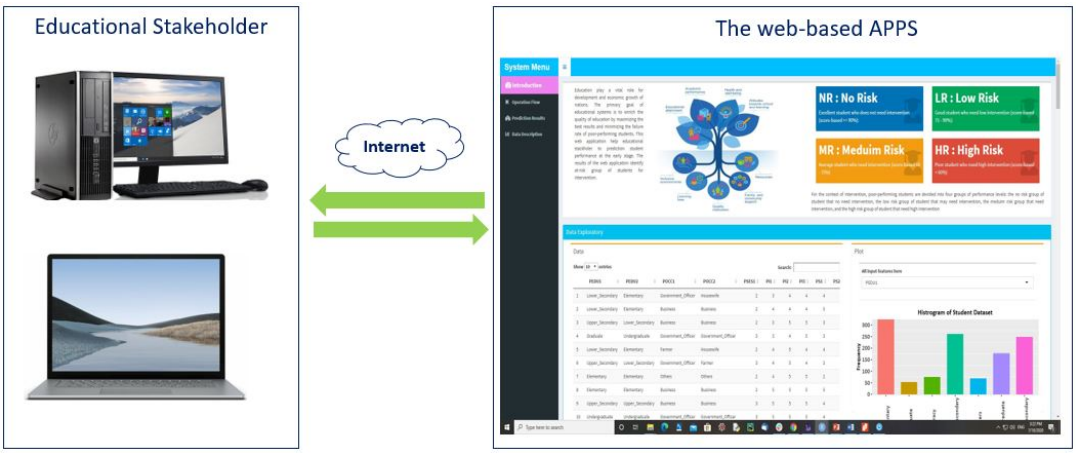


Figure 9.3: The materials for educational stakeholder to use APPS

9.4.2 The Design of APPS

This section present the design of APPS. The design process for obtaining the prediction system is as follow. The design consists of the web-based application linked to the server via internet. The design consists of the web-based application linked to the server via the internet. The clients can assess the application and upload their data to obtain the results. The web-based application was created using Shiny (an open-source R package in R language). Since our experiment is done in R, hence Shiny App is the best choice due to its simplicity. The shiny application is built using R language with its extension of the simplest code from HTML, CSS, and JavaScript. However, it is good that Shiny helps us to turn our analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge. The the component to design Shinny app is presented in Figure 9.4.

Figure 9.4: Shinny overview

9.4.3 Web Application System Interface

The architecture of the academic prediction system is shown in Figure 9.4. The recorded data that input in UI (user interface) is transferred to the server where the prediction model is executed to figure out what is the performance level of students. The results are then sent back to the client or user on the UI screen. The users are educational stakeholders such as teachers, administration office, or related individuals who have a dataset containing their student information. They can input the collected dataset and obtain the results. Figure 9.5 presents the introduction of the system.

The prototype shown as in Figure 9.6 presents the operation follow of tasks in using this prediction system. The flow chart aims to introduce to users how to use this app and the steps in using it and obtaining the prediction results.

The system gives a link to an online repository for data collection by conducting a survey questionnaire designed using Google document so that educational stakeholders can decide when is the suitable time to make an early prediction for intervention. Once the data collection is done, users can upload the data in the system (File Upload button). The results of prediction will be released in the system; in which users can download the prediction results, identify the poor performing group of students for intervention, and other useful information for future use.

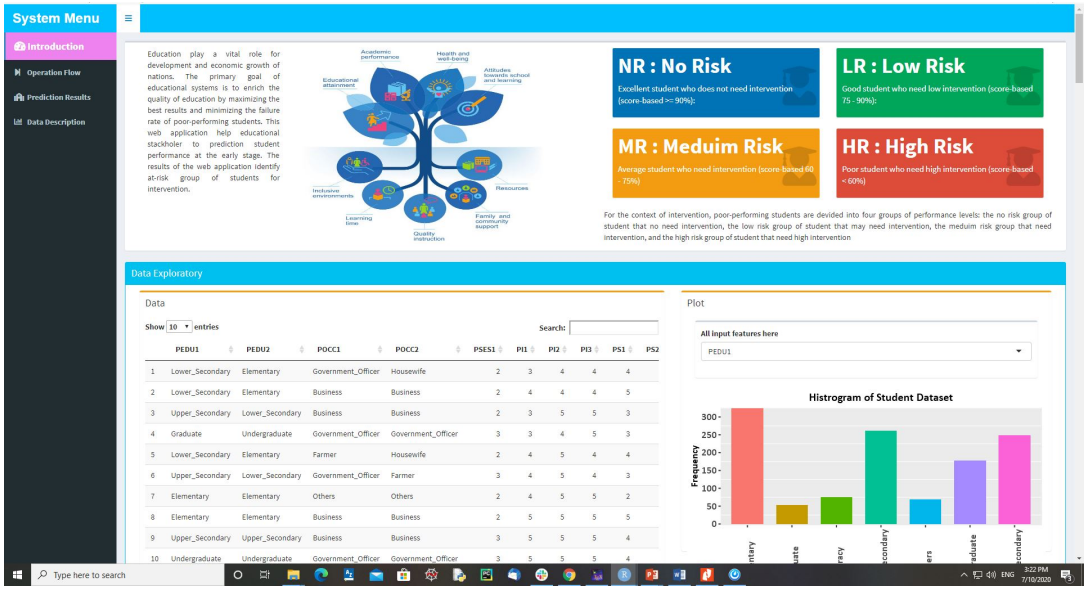


Figure 9.5: A prototype of APPS interface

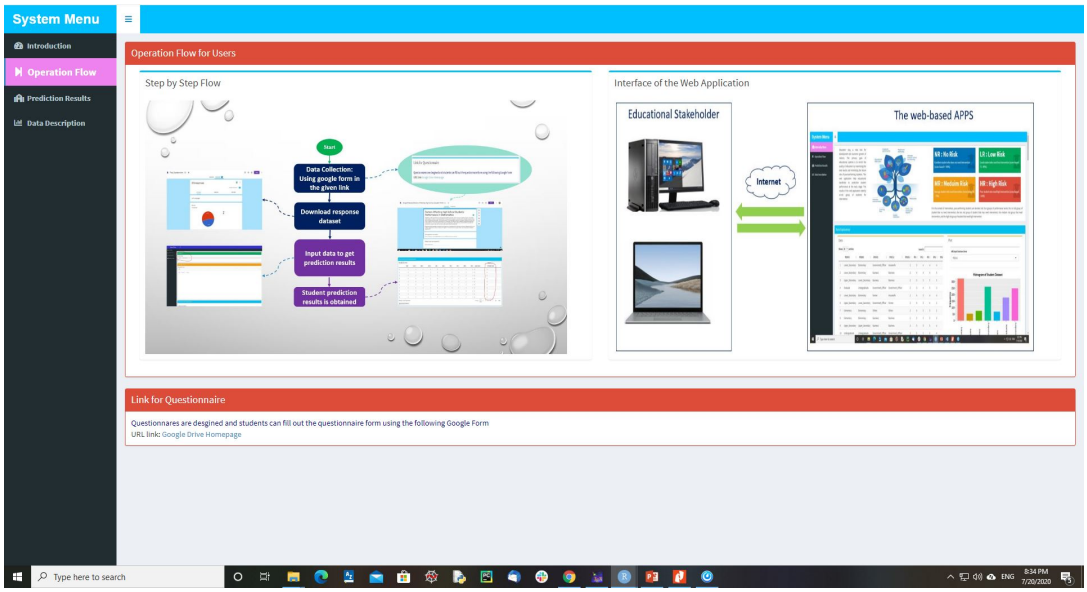


Figure 9.6: Operation flow chart for users to see the overall process to get prediction results

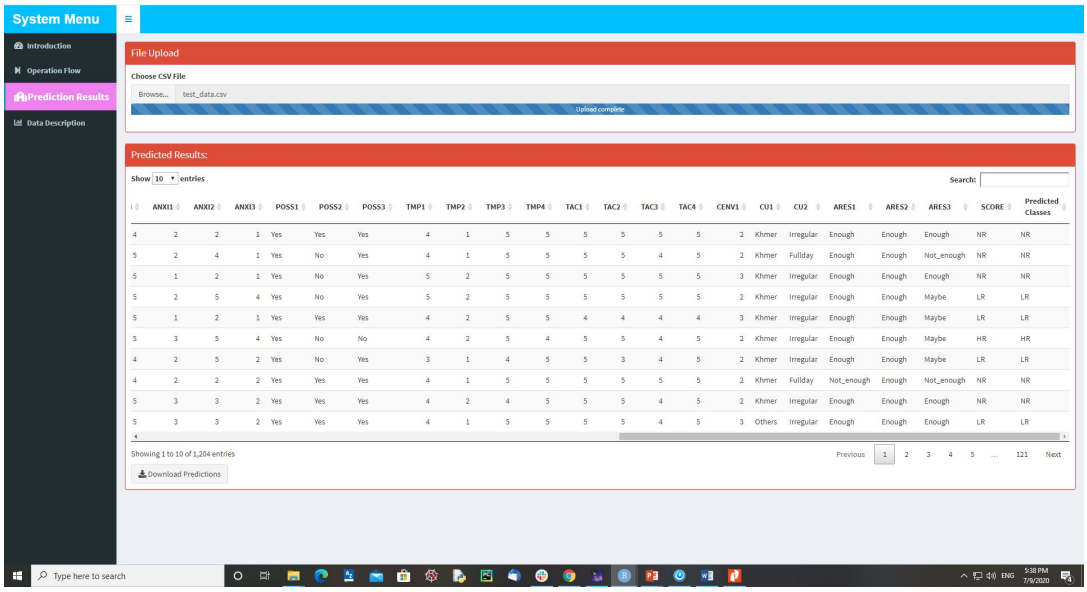


Figure 9.7: Prediction results identifying at-risk levels of poor-performing students

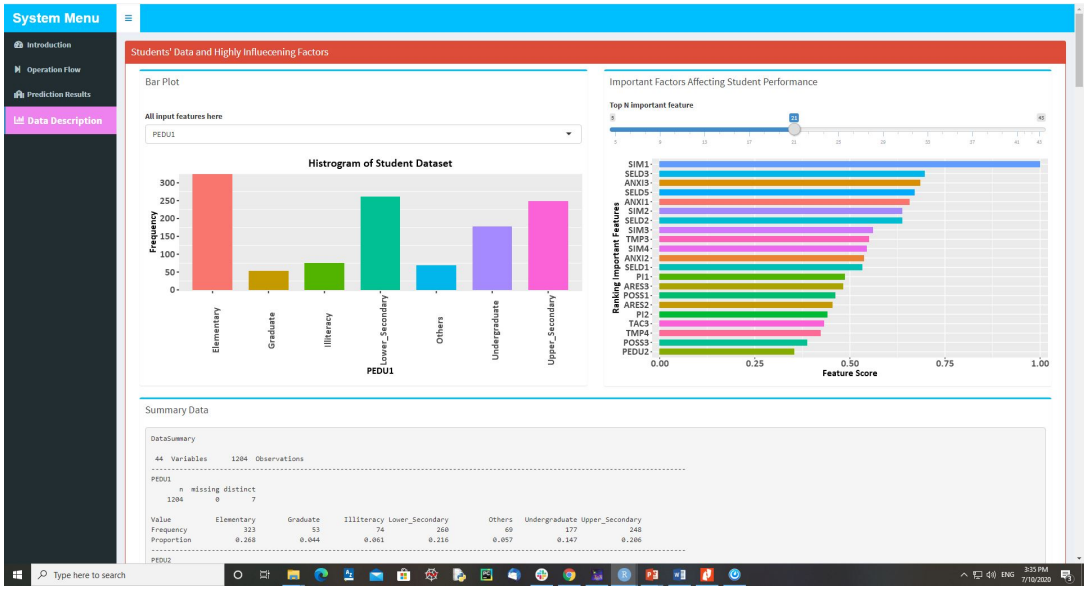


Figure 9.8: Summary of dominant set and ranking of highly influenced factors

Figure 9.8 demonstrates the information of students and the percentage of each group

of students based on their performance levels. The description of input data can be stored in the interface and viewed to understand students' information. The results of the prediction are shown in Figure 9.7 where at the bottom, there is a button that users can download the prediction results as a dataset in CSV format.

9.4.4 The System Usability

Once the development of the Shiny app (APPS) is done, it is shared or distributed with users. There is two basic option that can share. The first one is the Shiny app is shared as R scripts, users can use and edit from runGitHub. Users can use these scripts to launch the app from their R session. Users with no knowledge of programming or no care of how it works, the second type is to use the app from a web page or browser, which is from Shinyapps.io. This is definitely the most user-friendly way to share a Shiny App to users. They can navigate to our app through the internet with a web browser.

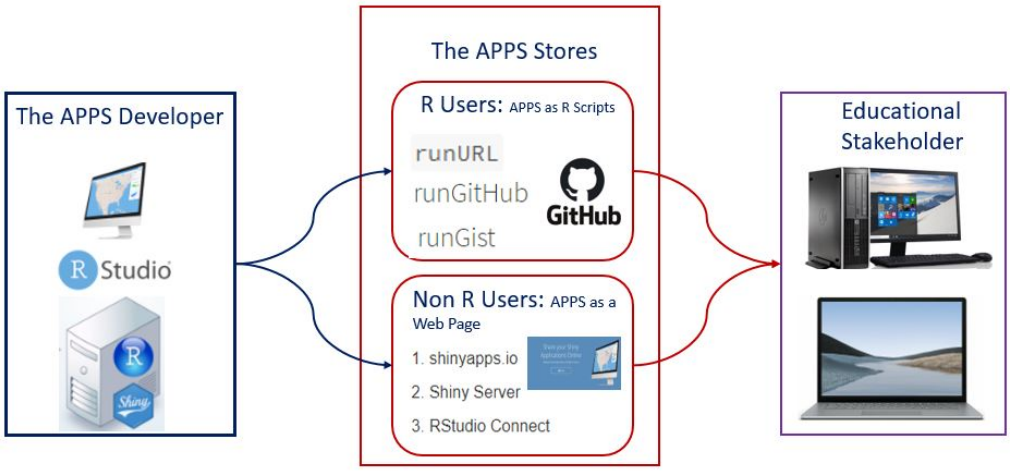


Figure 9.9: Deployment of APPS

9.4.5 The System Evaluation

In evaluating the usability of the web application of APPS, we designed a subjective questionnaire. The evaluation is carried out based on ten characteristics of the system: useful, motivating, user-friendly, relevant, reliable, efficient, organized, time cost, adaptable, and sophisticated. The questionnaire was designed with a 5-point Likert scale ranging from 1 to 5 (1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly agree). The description or statement regarding these characteristics of the systems is shown in Table 9.1. The participant of 57 students and 10 teachers are invited to join the presentation and test the system. The response regarding the survey of usability of the web application. When writing 11 (1) mean 11 students and 1 teacher agree with the given statement.

Table 9.1: Evaluation Survey Results.

Characteristics	Statement	5-point Likert Scale				
		1	2	3	5	5
Useful	System help students and instructors.	0(0)	0(0)	11(1)	31(6)	15(3)
Motivating	It is exciting to see that the system is able to give a feedback response according to the students face that affect their academic performance.	0(0)	0(0)	12(1)	37 (5)	10 (4)
User-friendly	The interface is user-friendly.	0(0)	0(0)	4 (2)	52(3)	1(5)
Relevant	It is easy to find the information I need.	0(0)	0(0)	17(3)	35(5)	15(2)
Reliable	I feel comfortable using the systems.	0(0)	0(0)	15(3)	35(4)	7(3)
Efficient	It produces results immediately after feeding in the information, and results are given correctly, easily and fast.	0(0)	0(0)	11(1)	31(6)	15(3)
Organized	It is easy to learn it use, the interface are simple and well structure.	0(0)	0(0)	4(3)	41(5)	12(2)
Time Cost	The data can obtained anytime and fast with questionnaire in Google form and results of prediction can obtain immediately after data collection.	0(0)	0(0)	5(1)	39(4)	13(5)
Adaptable	Student's weakness with be known so that the right intervention can be put in place.	0(0)	0(0)	6(2)	44(4)	7(4)
Sophisticated	This is innovative technology in educational system.	0(0)	0(0)	5(0)	46(6)	6(4)

The student participants are 35 males and 22 female students, and teacher participants are 7 male teachers and 3 female teachers. Most students and teachers agreed that the system is useful and effective for predicting student performance and identify the poor performing students for intervention. The survey result shows that 82.08% supported that the system is useful (55.22% agree, 26.86% strongly agree), 83.58% thought that the system is motivating (62.68% agree, 20.89% strongly agree), 91.04% felt that the interface is friendly (82.08% agree, 8.95% strongly agree), 85.07% believed the information was relevant (58.20% agree, 14.92% strongly agree), and 73.13% thought that the system was reliable (62.68% agree, 20.89% strongly agree), 82.08% reported the efficient of the system (55.12% agree, 26.86% strongly agree), 74.62% claimed that the system was well-organized (58.78% agree, 20.89% strongly agree), 91.04% realized that the system speed (time cost) was fast (64.17% agree, 26.86% strongly agree), 88.05% of participants perceived that the system was adaptable (62.68% agree, 20.89% strongly agree), and 92.58% felt that the system was sophisticated (77.61% agree, 14.92% strongly agree). The analysis of evaluation that summarize the percentage of agree and strongly agree on the characteristics of the system is shown in Figure 9.10. The evaluation survey indicates the effectiveness and usefulness of the developed academic prediction system.

9.5 Chapter Summary: Results of APPS Development

The APPS composes of a developed prediction model and an effective feature selection method for determining the dominant factors for the success of student performance. We proposed a comparative study of EDM of prediction and classification task, the outperformed prediction models are then developed and optimized to get the most successful classification results. The comparative experiment of four classifiers (KNN and Hybrid C5.0, Hybrid RF, and IDBN) is carried out using feature sets from four FS methods (IG, CHI, MI, and the Proposed MICHI method). The analysis of dominant

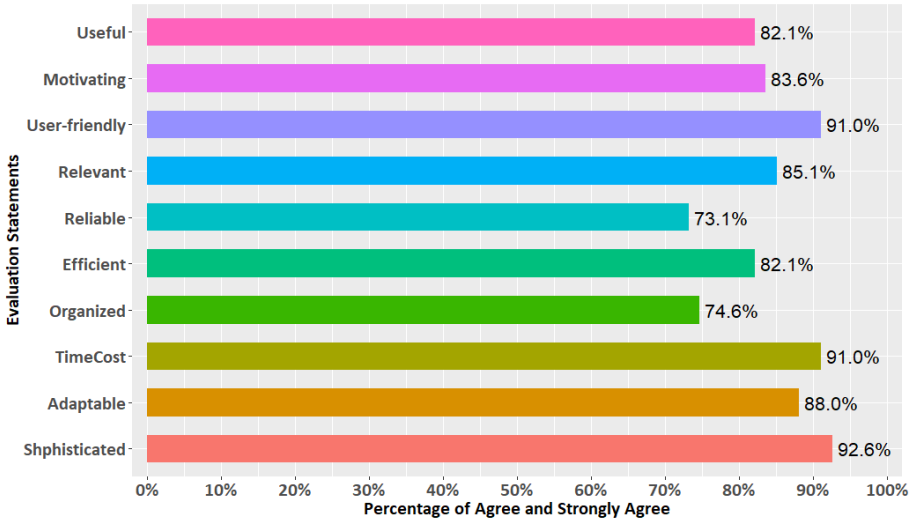


Figure 9.10: User feedback rating the characteristics of APPS

factors is cooperated and combined with the best classifier. The dominant set obtained from the MICHI algorithm significantly improves the performance of prediction models and used as a set of highly influencing factors that need to be considered for intervention and improvement of student performance. The experimental outcomes indicate that Hybrid RF outperformed the other three classifiers with the superior classification results. The developed prediction model and dominant factors are integrated into the web-based prediction system.

The findings of this research confirms the effectiveness of the prediction model and the usability of APPS. The system illustrates operation flows consist of the method of faster and more comfortable way of data collection, dominant factors that have a significant impact on the student performance, and results of prediction. According to the results from APPS, it is informative for educational institutions to carry out the right intervention for improving student performance.

CHAPTER 10

SUMMARY AND CONCLUSION

10.1 Introduction

High schools generally have intervention program and activities to support to students. However, identifying the actual group and right support activities are the challenging tasks. This dissertation give analysis cognitive and non-cognitive on factors that influence the student performance that mining factors that affect the students' success and poor performance group. The study aims at developing potential prediction models and academic performance prediction system for predicting student performance. We searched for a number of state-of-the-art models of educational data mining (EDM) then give a comparative study for extracting the leading models for predicting student performance. The methodology of the research is to develop the leading prediction models to get the most successful prediction results. This concluding chapter give a summary of proposed developed models and prediction system for academic performance prediction, discussion, and main contributions of the research.

10.2 Summary and Discussion

EDM approaches contributed a high level of extracting knowledge in the education environment. In this dissertation, we aim at proposing models and methods in EDM for predicting student performance in the purpose of intervention. Two kinds of datasets,

actual datasets and synthetic datasets were introduced. The actual datasets was collected from various high schools in Cambodia and the synthetic datasets were generated based of several studies of literature review, research in educational environment, and information from actual or original dataset. The procedure of our study is as follow:

Firstly, comparative study for model development is proposed. We produce multiple classifiers to investigate the performance of EDM models. Hence, three different main categories of research methods, predictive structural equation modeling (PSEM) of statistical analysis techniques, several high popular algorithms of machine learning, and deep belief network (DBN), a generative graphical framework of deep learning, were studied and compared. We generalize the comparative study of EDM models by using three artificial datasets. Random Forest (RF) algorithms generated the highest accuracy and smallest RMSE follows by C5.0. We subsequently obtained the actual dataset from data collection in Cambodia, we again introduce several classes of prediction models to this dataset. We applied the spot-checking algorithms process, which is to evaluate diverse algorithms on our dataset and see what work and drop what doesn't work; optimized the worked algorithms. In this work, we boost the tree-based algorithms with ensemble methods. The RF algorithms generate the highest accuracy with the lowest RMSE, which shows its effectiveness as the best algorithm.

Secondly, the optimal models are subsequently developed to get higher prediction performance. The two outperformed models (C5.0 of decision tree and random forest (RF)) from comparative study, and the other two popular model (naive Bayes (NB) and support vector machine with radial basis kernel function (SVM)). In this experiment, the hybrid machine learning models is the proposed. The developed models composed of combining the baseline machine learning models with principal component analysis, and a resampling method k-fold cross validation (10_FCV) to solve the missclassificaiton problem in baseline models. The hybrid models produced very good value of accuracy and RMSE. To confirm the robustness and effectiveness of the hybrid models, two artificial

datasets GDS2, GDS3, and ADS were used. The proposed approach boost up the prediction accuracy from 8% to 23% increase. The random forest algorithms still remain the best models with accuracy of 98.25%, and 98.71%, and 99.27% for the three datasets, respectively.

Thirdly, in the age of AI revolution, deep learning framework is a powerful tool. In the context of larger datasets, we propose an optimization approach of deep belief network (IDBN). We introduced two main experiment. The first experiment was done with actual dataset ADS comprises 1204 samples. The results show that our proposed improved DBN could produce an accuracy of 86.59% (10% increased compare to traditional DBN), and RMSE value of 0.045. In the second experiment, we implemented using four artificial datasets with increasing size. The four datasets are denoted as GDS1 (1000 samples) , GDS2 (2000 samples) , GDS3 (4000 samples), and GDS4 (10000 samples). We compared IDBN to the LVQ, MLP, DBN, and hybrid RF. The IDBN standouts the three neural network models, and follow Hybrid RF in GDS1 and GDS2. IDBN performs comparatively to Hybrid RF, and performs better in GDS4. The accuracy of IDBN on the four datasets are 80.44%, 95.01%, 99.02%, and 99.37%. The experimental results indicated that with the context of larger dataset, the IDBN could produce as best as the Hybrid RF.

Lastly, in order to extract highly influencing factors alongside with developed prediction models, the study on analysis of dominant factors is conduced. The dominant factors refer to a set of input feature that are needed and sufficient for controlling the outcome of student performance. The merit of the dominant factors are the dominant sets that improve the performance of prediction models and capture the student learning patterns. We presents four effective FS algorithms (MI, IG, SU, and CHI) and introduced a novel FS algorithm (MICHI) in our study. The experimental results of Hybrid RF on ADS give the highest accuracy of 99.98%. In dataset GDS4, the accuracies of Hybrid RF and IDBN are 99.85% and 99.83%, respectively. The experimental results from this

study indicated the best performance of developed prediction models and the dominant factors that answered or replied our research questions.

In addition, we design a web application for academic prediction. By using the proposed techniques of EDM, an Academic Performance Prediction system (APPS) is subsequently developed for educational stockholders to get an early prediction of student learning outcomes for timely intervention. Experimental outcomes and evaluation surveys report the effectiveness and usefulness of the developed academic prediction system.

10.3 Methodological Contributions

10.3.1 Developed Optimal Prediction Models

The prediction process were make it in 3 main experiment. The comparative study of the popular prediction models of EDM was conducted. The last two experiment are the experiment of model improvement or development. The second is the proposed hybrid models of machine learning algorithms, and the third experiment is the Optimization of DBN with with context of larger datasets. These two main model developments can alternatively used as the optimal prediction model to predict academic performance.

10.3.2 Proposed a Novel FS Algorithm

In the context of intervention of academic performance and improving prediction models, optimal feature sets are investigated. The purpose of determining the dominant feature are to improve the performance of prediction models and select the dominant factors for academic performance. Feature selection methods are considered in this study. The study give a comparative study of existing FS method and proposed FS method, MICHI. The proposed MICHI capture the dominants set and significantly improved the performance of prediction models.

10.4 Applicability Contributions

Identifying the performance levels of group of students combining with observing the highly influencing factors are the needed task for educational environments. Usually, there are several intervention programs and activities to support students. However, they programs and activities are self-choice where students choose whether to participate or not. Many students who need help are left behind because they do not know where to go while related individuals also do not know which exact group should be chosen. Therefore, leading targeted student to success, reducing drop out and repeated class is not just the responsibility of students themselves but it is as much a responsibility of educational stackholder and related individual such as their family, friends, teachers, and the surrounding environment.

By using the proposed techniques of educational data mining, and academic performance prediction system is subsequently developed for educational stockholders to get an early prediction of student learning outcomes for timely intervention.

10.4.1 The web-based APPS for Predicting Student Performance

High school academic performance has a greater impact on student performance in secondary education, especially for national diploma exam and further study in higher education and career skills. Poor performance of students was considered as a strong predictor of academic success, drop out, and repeated class. A major part of this study was to construct a adequate understanding of student behaviors that directly and indirectly affect their academic performance and life skills.

Interventions that increase student communication with instructors, connect students to available school service, and privde students with an *accurate assessment of how they are currently performing* can be effective at increasing both retention and academic performance. APPS can be used as a component of an early-warning system that

can equip instructors to provide interventions with these crucial characteristics. The system illustrates operation flows consist of the method of faster and more comfortable way of data collection, dominant factors that have a significant impact on the student performance, and results of prediction. According to the results from APPS, it is informative for educational institutions to carry out the right intervention for improving student performance. The developed prediction system will help educational stakeholders such as teachers, educational administrators, and policymakers, and related individuals to improve academic performance in educational institutions. The teachers can quickly adjust their teaching methods and adopt adaptive teaching approaches to meet the needs of students. Educational stakeholders and related individuals can figure out the weak points and the solution to make improvements. Therefore, overall learning quality and learning performance can be improved greatly and reduce the failure rate of poor-performing students.

10.4.2 Intervention: Recommendation and Action Strategies

This results from the APPS can be used as recipe to revise their learning behaviors. Finally, results can be mapped into some action strategies to the following individuals:

(a) **Providing Feedback for Parents:** From ranking informative features, family aspect also involvel and participate importantly in improving their kids performance. Parents' education and parents' involvement follow up their children performance role importantly in affecting student performance. That's parents should pay attention and follow up in order to improve and encourage their children performance.

(b) **Providing Feedback for Students:** The level of interest of the students in class is the most important feature since they study hard to get good score or lazy depend on their interest. Anxiety in class, the feeling of stress and pressure in math class and the feeling of helpless in class are the important factor to consider and give intervention.

Additionally, the level of preparing for exam, quiz, the frequency of doing homework are the important things that student should do. Most student perform poorly in math class due to the low level of enjoyment in lesson and less motivation to get good score in their study.

(c) **Providing Feedback for Supporting Teachers:** Teacher's teaching methods and practices are the important thing that affect the student performance. Then they should improve and enhance their teaching methods for student. Additionally, the connect or being friendly to students can reduce student anxiety and nervousness of asking for help in class. Teachers role importantly to motivate students to interest, pay attention and enjoy in class. More importantly,from the prediction results, teachers can identify the poor performance group of students and cooperate with schools for intervention.

(d) **Providing Feedback for Schools:** School should taking care of adequate use of required handout, and adequate number of teachers. Obtaining the highly influencing features, school can cooperate with family, and teachers to support students to enhance their academic achievements.

10.4.3 Managerial Settings, Planning and Scheduling

The success of student in high school role importantly as a strong backbone for these skill. The success of student in high school become an important thing to implement. Hence, the study of predicting achievement of students and mining the influence factors is a point educational settings. The results from the developed prediction models and systems can be used as information for managerial setting, planning and scheduling. More importantly, these results will be used as information to timely identify poor-performance group of students in the early stage.

BIBLIOGRAPHY

[1] Ministry of Education, Youth and Sport, Phnom Penh, Cambodia, http://www.moeys.gov.kh/en/.

[2] Pit, C., & Ford, D. (2004). Cambodian higher education: Mixed visions. *In P.G Altbach and T. Umakooshi (Eds), Asian Universities: Historical perspectives and contemporary challenges (pp. 333-362)*. Baltimore: Jonh Hopkins University Press.

[3] MoEYS (2005). *Education strategic plan 2006-2010*. Phnom Penh: Author, Cambodia.

[4] MoEYS (2018). *Education in Cambodia: Finding from Cambodia's Experience in PISA for Development*. Phnom Penh: Author, Cambodia.

[5] MoEYS (2019), *Policies on Science, Technology, and Innovation, 2020-2030*. Phnom Penh: Author, Cambodia.

[6] MoEYS (2017). *Education in Cambodia: Finding from Cambodia's Experience in PISA for Development*. Phnom Penh: Author.

[7] Fall, A. M., & Roberts, G. (2012). High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. *Journal of adolescence*, 35(4), 787-798.

[8] OECD (2016). *PISA: Low-performing students: Why they fall behind and how to help them succeed*. OECD Publishing, Paris.

[9] OECD (2013). *PISA 2012 Results: What Makes Schools Successful? Resources, Policies and Practices (Volume IV)*. OECD Publishing, Paris.

[10] Orodho, J. A., Waweru, P. N., Ndichu, M., & Nthinguri, R. (2014). Home-based challenges to effective implementation of curriculum under free primary education system in nomadic kenya. *Journal of Education and Practice*, 5(26), 134-144.

[11] Etsey, K. (2005). Causes of low academic performance of primary school pupils in the shama sub-metro of shama ahanta east metropolitan assembly (saema) in ghana. *In Regional Conference on Education in West Africa Dakar*, Senegal, 1-34.

[12] Pears, K. C., Kim, H. K., & Fisher, P. A. (2008). Psychosocial and cognitive functioning of children with specific profiles of maltreatment. *Child abuse & neglect*, 32(10), 958-971.

[13] Bae, D., & Wickrama, K. (2014). Family socioeconomic status and academic achievement among korean adolescents linking mechanisms of family processes and adolescents' time use. *The Journal of Early Adolescence*, 11.

[14] Chen, W. B., & Gregory, A. (2009). Parental involvement as a protective factor during the transition to high school. *The Journal of Educational Research*, 103(1), 53-62.

[15] Bowen, G. L., Hopson, L. M., Rose, R. A., & Glennie, E. J. (2012). Students' perceived parental school behavior expectations and their academic performance: A longitudinal analysis. *Family Relations*, 61(2), 175-191.

[16] Wegmann, K., & Bowen, G. L. (2010). Strengthening connections between schools and diverse families: A cultural capital perspective. *Prevention Researcher*, 17(3), 7-10.

[17] Herbert, K. (1978). *The New Book of Popular Science Daribury*. Connecticut: Grolier Inc.

[18] Al-Agili, M. Z. G., Mamat, M., Abdullah, L., & Maat, H. A. (2012). The Factors Influence Students' Achievement in Mathematics. *Australian Journal of Basic and Applied Science*, 17(9), 1224-1230.

[19] Al-Agili, M. Z. G., Mamat, M., Abdullah, L., & Maat, H. A. (2013). Two Models of Describing Students' Achievement in Mathematics: A Comparative Study. *Australian Journal of Basic and Applied Science*, 7(6), 184-198.

[20] Wong, N.Y. (1992). The Relationship Among Mathematics Achievement, Affective Variables and Home Background, Mathematics. *Education Research Journal*, 4(3), 32-42.

[21] Barnes, T., Dessmaris, M., Romero, C., & Ventura, S. (2009, July 1-3). Educational data mining 2009. *Proceedings of the 2nd International Conference on Educational Data Mining*, Cordoba, Spain.

[22] Pena-Ayala, A. (2014). Educational data mining: Survey and a data mining-based analysis of recent works. *Expert Systems with Application*, 41(2014), 1432-1462.

[23] Romero, C., & Ventura, S. (2006). Educational data mining: A survey from 1995 to 2005. *Expertise Systems with Application*, 33(1), 135–146.

[24] Romero, C. & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601–618.

[25] Mazidi, A. A., & Abusham, E. (2018). Study of general education diploma students' performance and prediction in Sultanate of Oman, based on data mining approaches. *International Journal of Engineering Business Management*, 10, 1-11.

[26] Romero, C., & Ventura, S. (2013). Educational data in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.

[27] Thakar, P., Mehta, A., & Manisha (2015). Performance analysis and prediction in educational data mining: A research travelogue. *International Journal of Computer Application*, 110(15), 60-68.

[28] Al-Agili, M. Z. G., Mamat, M., Abdullah, L., & Maat, H. A. (2013). Path Analysis of the Factor Influencing Students' Achievement in Mathematics. *Australian Journal of Basic and Applied Science*, 7(4):437-442.

[29] Uysal, S. (2015). Factors affecting the Mathematics achievement of Turkish students in PISA 2012. *AcademicJournals*, 10(12), 1670-1678.

[30] Nazim, A., & Ahmad, S. (2014). Modeling the influential factors of 8th grades students' mathematics achievement in Malaysia by using structural equation modeling (SEM), *International Journal of Scientific & Engineering Research*, 5(6), 950-953.

[31] Tambunan, H. (2018). The dominant factor of teacher's role as a motivator of students' interest and motivation in mathematics achievement. *International educational studies, Canadian Center of Science Education*, 11(4), 144-151.

[32] Serhat, K., & Ferit, K. (2014). A Structural Equation Modeling on Factors of How Experience Teachers Affect the Student's Science and Mathematics Achievements. *Educational Research International*, Hindavi Publishing Corporation, Article ID 49037.

[33] Kilic, S., & Askin, O. E. (2007). Parental influence on students' mathematics achievement: the comparative study of Turkey and best performer countries in TMISS. *Procedia-social and behavioral science*, 106(2013), 2000-2007.

[34] Anuradha, C., & Velmurugan, T. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students performance. *Indian Journal of Science and Technology*, 8(15), 1-12.

[35] ThaiNghe, N., Janecek, P., & Haddawy, P. (2007, October 10-13). A Comparative Analysis of Techniques for Predicting Academic Performance. *ASEE/IEEE Frontiers in Education Conference.*

[36] Kotsiantis, S., Piarrekeas, C., & Pintelas, P. (2007). Predicting Students' performance in Distance Learning using Machine Learning Techniques. *Applied Artificial Intelligence*, 18, 411-426.

[37] Minaei-Bidgoli, B., Kashy, D., Kortemeyer, G., & OPunch, W. F. (2003). Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA, *33rd ASEE/IEEE Frontiers in Education Conference.*

[38] Kovacic, Z. J. (2007). Early Prediction of Students Success: Mining Students Enrolment Data. *Information science & IT Education Conference.*

[39] Mythili, M. S., & Shanavas, M. A. R. (2014). An Analysis of Students' Performance using Classification Algorithms, *IOSR Journal of computer Engineering (IOSR-JCE)*, 16(1), 63-69.

[40] Daud, A., Aljonhani, N. R., Abbasi R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2016). Prediction student performance using advanced learning analytics. *Proceedings of 26th International Conference on World Wide Web*, Companion, Perth, Australia, 416-421.

[41] Sultana, J., Rani, M. U., & Farquad, M.A.H. (2019). Student Performance Prediction Using Deep Learning and Data Mining Methods. *International Journal of Recent Technology and Engineering (IJRTE)*, ISSN: 2277-3878, 8(1S4).

[42] Sinthia, G., & Balamurgan, M. (2019). Analyzing Student's Academic Performance Using Multilayer Perceptron Model. *International Journal of Recent Technology and Engineering*, 7(5S3).

[43] Zhang, Y., Shah, R., Chi, M. (2016). Deep Learning+Student Modeling+Clustering: A Recipe for Effective Automatic Short Answer Grading. *Proceeding of the 9th International Conference on Education Data Mining.*

[44] Estrera, P. J. M., Natan, P. E., Rivera, B. G. T., & Colarte, F. B. (2017). Student performance analysis for academic rankings using decision tree approach in the University of Science and Technology of Southern Philippines senior high school. *International Journal of Engineering and Technologies*, 3(5), 147–153.

[45] Ramaswami, M., & Bhaskaran, R. (2009). A study of feature selection techniques in educational data mining. *Journal of Computing*, 1(1), 7–11.

[46] Amjad, A. S., Al-Emran, M., & Khaled, S. (2020). Mining Student Information System Records to Predict Students' Academic Performance. *Springer Nature Switzerland AG 2020, AMLTA 2019, AISC 921*, 229-239.

[47] Dimic, G., Rancic, D., Milentijevic, I., Spalevic, P., & Plecic, K. (2017). Comparative Study: Feature Selection Methods in the Blended Learning Environment. *Automatic Control and Robotics*, 6(2), 96-116.

[48] Zaffar, M. & Savita, K.S. (2018). A study of feature selection algorithms for predicting students' academic performance. *International Journal of Advanced Computer Science and Applications*, 9(5), 541-549.

[49] Jonhson, E. S., Smith, L. & Harris, M. L. (2009). *How RTI works in secondary school*. Thousand Oaks, CA.

[50] Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online course learning performance. *Computers in Human Behaviors*, 36, 469–478.

[51] Ackcapinar, G., Altun, A., & Askar, P. (2019). Using learning analytics to develop early warning systems for at-risk students. *International Journal of Educational Technology in High Education*, 16(40), 1-20.

[52] Lee, S. & Chung, J. Y. (2019). The machine learning-based dropout early warning systems for improving the performance of dropout prediction. *Journal of Applied Science*, 9(15), 1-14.

[53] Pena-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41 (4), 1432-1462.

[54] Johnson, D., & Samora, D. (2016). The potential transformation of higher education through computer-based adaptive learning systems. *Global Education Journal*, 2016(1).

[55] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, Todd Green.

[56] Wright, S. (1921). Correlation and Causation. *Journal of Agriculture Research*, 20, 557-586.

[57] Mohamed, Z. G. A., Mustafa, B. M., Lazim, A., & Hamdan, A. M. (2013). Two Models of Describing Students' Achievement in Mathematics: A Comparative Study. *Australian Journal of Basic and Applied Science*, 7(6), 184-198.

[58] Masud, H., Ahmad, M. S., Jan, F. A., & Jamil, A. (2016). Relationship between parenting styles and academic performance of adolescents: mediating role of self-efficacy, *Asia Pafici Educ. Rev*, DOI 10.1007/s12564-015-9413-6.

[59] Anders, S., & Sophia, R. H. (2005). Structural Equation Modeling: Categorical Variables. *Wiley: Entry for the Encyclopedia of Statistics in Behavioral Science*, https://onlinelibrary.wiley.com/doi/epdf/10.1002/0470013192.bsa596.

[60] Vanitbancha, K. (2007). *Multivariate Data Analysis. Vol. 2.* Bangkok: Chulalongkorn University Book Center.

[61] Field, A. (2000). *Discovering statistics using SPSS for Windows.* Sage Publication, London-Thousand Oaks-New Delhi.

[62] Randall, E. S., & Richard G. L. (2010). *A Beginner's Guide to Structural Equation Modeling.* Taylor and Francis Group, LLC.

[63] Muthen, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent indicators. *Psychometrika*, 49, 115–132.

[64] Yu, C. Y. (2002). *Evaluation of model fit indexes for latent variable models with categorical and continuous outcomes.* Unpublished dissertation. Retrieved Jnaurary 5, 2005 from Mplus website https://www.statmodel.com/download/Yudissertation.pdf.

[65] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. L. (2010). *Multivariate Data Analysis* (7 ed.). New Jersey: Prentice-Hall.

[66] Kleine, R. B. (2005). *Principle and Practice of structural equation modeling.* The Guilford Pres, New York.

[67] Bentler, P. M., & Bonnet, D. C. (1980). Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin*, 88 (3), 588-606.

[68] Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, 107 (2), 238-46.

[69] Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics (5th ed.)*. New York: Allyn and Bacon.

[70] Hu, L. T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6 (1), 1-55.

[71] Steiger, J. H. (1990). Structural Model evaluation and modification. *Multivariate Behavioral Research*, 25, 1990.

[72] Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming*. New Jersey: Lawrence Erlbaum Associates.

[73] Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL*. London: Sage Publications.

[74] MacCallum, R. C., Browne, M. W., & Sugawara, H., M. (1996). Power Analysis and Determination of Sample Size for Covariance Structure Modeling. *Psychological Methods*, 1(2), 130-49.

[75] Joreskog, K. G., & Sorbog, D. (1982). Recent Development in Structural Equation Modeling. *Journal of Marketing Research*, 19(4), 404-416.

[76] McQuitty, S. (2004). Statistical power and structural equation models in business research. *Journal of Business Research*, 57 (2), 175-83.

[77] Hooper, D., Coughlan, J., Mullen, M. (2008). Structural Equation Modeling: Guiding for Determining Model Fit. *Electronic Journal of Business Research Methods*, 6(1), 53-66.

[78] John, D. K., Brain, M. N., & Aoife, D. (2015). *Fundamentals of machine learning for predictive data analytics*. Cambridge: The MIT Press.

[79] Burger, S. V. (2018). *Introduction to Machine Learning with R: Rigorous Mathematical Analysis*. O'Reilly Media, 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[80] Mohri, M., Afshin, R., & Ameet, T. (2018). *Foundations of Machine Learning*. The MIT Press.

[81] Wei, Y., & Chu, D. (2017). *Machine Learning with R Cookbook (2nd ed.)*. PACKT Publishing.

[82] Karthik, R., & Abhishek, S., (2019). *Machine Learning Using R: With Time Series and Industry-Based Use Cases in R*. Apress.

[83] El-Habil, A. M. (2012). An Application of Multinomial Logistic Regression Model. *Pakistan Journal of Statistics and Operation Research*, 8(2), 271-291.

[84] Chatterjee, S., & Hadi, A. (2006). *Regression Analysis by Example*. John Wiley & Sons.

[85] Farid, J., & Ahmad, A. S. (2019). Building student's performance decision tree classifier using boosting algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, 14(3), 1298-1304.

[86] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Taylor & Francis Group.

[87] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[88] Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag.

[89] Burgers, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discov,* 2(1998) 121–167.

[90] Scholkopf, B. & Smola, A.J. (2002). *Learning with Kernels*. MIT, London.

[91] Lin, H. T., & Lin, C. J. (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. *Technical Report, University of National Taiwan, Department of Computer Science and Information Engineering*, 1–32.

[92] Muller, K.R., Mike, S., Ratsch, G., Tsuda, K., & Scholkopf B. (2001). An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* 12(2001) 181–201.

[93] Humera, S., Raniah, Z., Kavitha, G. (2015). Prediction of Student Performance in Semester Exam Using a Naïve Bayes Classifier. *International Journal of Innovative Research in Science, Engineering and Technology*, 4(10), 9823-9829.

[94] Khamis, H. S., Cheruiyot, K. W., & Kimani, S. (2014). Application of k-Nearest Neighbour Classification in Medical Data Mining. *Int. J. Inf. Commun. Technol. Res.*, 4(4), 121–128.

[95] Wibowo, A. T., & Fitrianah, D. (2018). A k-nearest algorithm based application to predict SNMPTN acceptance for high school student in Indonesia. *International Research Journal of Computer Science (IRJCS)*, 5(1).

[96] Suguna, N., & Thanushkodi, K. (2013). An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. *International Journal of Computer Science Issues*, 7(4), 18–21.

[97] Giri, A., Bhagavath, M. V., Pruthvi, B., & Dubey, N. (2016). A Placement Prediction System using k-nearest neighbors classifier. *Proc. - 2016 2nd Int. Conf. Cogn. Comput, Inf. Process. CCIP 2016*, 3–6.

[98] Brownlee, J. (2019). *Mastering Machine Learning Algorithms*. In Press.

[99] Kohonen, T. (1990). The self-organizing maps. *proceeding of the IEEE*, 78, 1464-1489.

[100] Kohonen, T. (1992). New Development of Learning Vector Quantization and the Self-Oragizing Map. *Symposium on Neural Networks: Alliances and Perspectives in Senri*, Osaka, Japan.

[101] Gaur, P. (2012). Neural Network in Data Mining. *International Journal of Electronics and Computer Science Engineering (IJECSE, ISSN:2277-1956)*, 1(3), 68-84.

[102] Zare, M., Pourghasemi, H. R., Vafakhah, M., & Pradhan, B. (2013). Landslide susceptibility mapping at vaz watershed (iran) using an artificial neural network model: a comparison between multilayer perceptron (mlp) and radial basic function (rbf) algorithms. *Arabian Journal of Geosciences*, 6(8), 2873-2888.

[103] Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., & Held, P. (2013). Multi-layer perceptrons. *In Computational Intelligence*, 47-81, Springer.

[104] Panchal, G., Ganatra, A., Kosta, Y., & Panchal, D. (2011). Behavior analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *International Journal of Computer Theory and Engineering*, 3(2), 332-337.

[105] Venkatesan, P., & Anitha, S. (2006). Application of a radial basis function neural network for diagnosis of diabetes mellitus. *Current Science*, 91(9), 1195–1199.

[106] International Educational Data Mining Society. http://www.jedm.educationaldatamining.org/index.php/JEDM.

[107] Antonio, H. B., Boris, H. F., David, T., & Borja, N. C. (2019). A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Hindawi: Complexity*, 2019, Article ID 1306039.

[108] Hinton, G. E. (2010). A practical guide to training restricted Boltzmann machines. *Momentum 2010*, 9(1).

[109] Mazdak, F., Mahnood, A., Arash, A., Mahyar, S., & Philippe, D. (2016). Towards an Spiking Deep Belief Network for Face Recognition Application. *6th International Conference on Computer and Knowledge Engineering (ICCKE 2016).*

[110] Qadri, S. F., Ai, D., Hu, G., Ahmad, M., Haung, Y., Wang, Y., & Yang, J. (2019). Automatic Deep Feature learning via Pathch-Based Deep Belief Network for Vertebrae Segmentation in CT Image. *Applied Science*, 9(69), doi:10.3390/app9010069www.mdpi.com/journal/applsci .

[111] Lokare, V., Birari, S., & Patil, O. (2015). Application of Deep Belief Network for Image Compression. *International Journal of Computer Science and Information Technology*, 6(5), pp. 4799-4803.

[112] Sharikaya, R., Hinton, G. E., & Deoras, A. (2014). Application of Deep Belief Network for Natural Language Understanding. *IEEE Transaction on Audio, Speech and Language Processing*,22(4), 778-784.

[113] Kim, J., Kang, U., & Lee, Y. (2017). Statistics and Deep Belief Network-Base Cardiovascular Risk Prediction. *Healthc Inform Res.*, 23(3), 169-175.

[114] Li, H., Li, X., Ramanathan, M., & Zhang, A. (2014). Identifying informative risk factors and prediction bone decreases progression via deep belief network. *Methods*, 69, 257-265.

[115] Tamilselvan, P., Wang, P. (2013). Failure diagnosis using deep belief learning based health state classification. *Reliab Eng Syst Saf*, 115, 124-135.

[116] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14, 1711-1800.

[117] Carriera-Perpinan, M. A., & Hinton, G. E. (2005). One Contrastive Divergence. *In. AISTATS*, 10, 33-40.

[118] Hinton, G. E., Osindero, S., Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput* 18(7), 1527-1554.

[119] Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multi task learning. *In Proceedings of the International Conference on Machine Learning. ACM, 2008*, 160–167.

[120] Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pretrained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 20(1), 30–42.

[121] Pudil, P., & Novovičová, J. (1998). Novel Methods for Feature Subset Selection with Respect to Problem Knowledge. *IEEE Intelligent Systems and their Applications*, 13(2), 66-74.

[122] Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *International Conference on Machine Learning (ICML)*, Washington DC.

[123] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1), 273-324.

[124] Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2), 153-158.

[125] Collins, R. T., Liu, Y., & Leordeanu, M. (2005). Online selection of discriminative tracking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), 1631-1643.

[126] Kumar, G., & Kumar, K. (2011). A novel evaluation function for feature selection based upon information theory. *In Electrical and Computer Engineering (CCECE), 2011 24th Canadian Conference*, 000395-000399.

[127] Peteiro-Barral, D., Bolroon-Canedo, V., Alonso-Betanzos, A., Guijarro-Berdinas, B., & Sanchez-Marono, N. (2013). Toward the scalability of neural networks through feature selection. *Expert Systems with Applications*, 40(8), 2807-2816.

[128] Cawley, G. C., & Talbot, N. L. (2010). On overfitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research,* 11, 2079-2107.

[129] Sanmay, D. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. *In proceedings of the International Conference on Machine Learning 2001*, 74-81.

[130] Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3), 483-519.

[131] Saeys, Y., Inza, I., & Larra naga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics,* 23(19), 2507-2517.

[132] Lazar, C., Taminua, J., Meganck, S., Steehno, D., Coletta, A., Mlter, C., De Schaetzen, V., Dugue, R., Bersini, H., & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4), 1106-1119.

[133] Tang, J., Alelyani, S., & Liu, H. (2014). *Feature selection for classification: A review. Data Classification: Algorithms and Applications*. Data Mining and Knowledge Discovery Series, CRC Press In Chapman & Hall/CRC.

[134] Jantawan, B., & Tsai, C. F. (2014). A comparison of filter and wrapper approaches with data mining techniques for categorical variables selection. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(6), 4501-4508.

[135] Ma, L, Fu, T., Blaschke, T., Li, M., Tiede, D., Zhou, Z., Ma, Z., & Chen, D. (2017). Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine. *International Journal of Geo-Information*, 6(56), doi:10.3390/ijgi6020051.

[136] Jindal, P., & Kumar, D. (2017). A Review on Dimensionality Reduction Techniques. *International Journal of Computer Applications*, 173(2), 42-26.

[137] Phyu, T.Z., & Oo, N.N. (2016). Performance Comparison of Feature Selection Methods. *MATEC Web of Conferences42*, DOI: 10.1051/matecconf/20164206002.

[138] Bommert, A., Sun, Z., Bischl, B., Ranenfuhrer, J., Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Journal of Computational Statistics and Data Analysis*, 143.

[139] Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2007). Advancing Feature Selection Research: Asu Feature Selection Repository, TR-10-007. *School of Computing, Informatics, and Decision Systems Engineering, Arizona State University: Tempe, AZ, USA*.

[140] Liu, H., Setiono, & Chi, R. (1995, November 5-8). Chi2: Feature selection and discretization of numeric attributes. *In Proceedings of the Seventh IEEE International Conference on Tools With Artificial Intelligent*, Herndon, VA, USA.

[141] John, G. H., Kohavi, R., & Pfleger, K.(1994). Irrelevant Features and Subset Selection Problem. *Proceeding of the Eleventh International Conference of Machine Learning*, Morgan Kaufmann Publishers, San Francisco, CA, 121-129.

[142] Hall, M. A., & Smith, L.A. (1997). Feature subset selection: a correlation based filter approach. *Proceeding of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems*, Springer, 855-858.

[143] Matinez, A.M., & Kak, A.C. (2001). Principal component analysis versus linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228-233

[144] Jollifie I. T. (1986). *Principle components analysis and factor analysis (pp. 115-128)*, Springer.

[145] Duda, R. O., Hart, P. E., & Stork, D.G. (2001). *Pattern Classification (2 ed.)*. John Wiley & Sons, Inc., New York.

[146] Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *J Data Warehousing (2000)*, 5, 13—22.

[147] Oreski, D., Pihir, I., and Konecki, M. (2017, April 27-28). CRISP-DM Process Model in Educational Settings. *20th International Scientific Conference "Economic and Social Development"*, 19-28.

[148] Yaacob, W. F. W., Nasir, S. A. S., Yaacob, W. F. W., & Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(3), 1584-1592.

[149] Bali, S. J., Demo, D. H., & Wedman (1998). Family involvement with children's homework: An intervention in the Middle Grades, family relations, 47.

[150] Caro, D. H. (2009). Socioeconomic status and academic achievement trajectories from childhood to adolescence. *Canadian J. Educ.*, 32, 558-590.

[151] Clark, R. M. (1983). *Family life and school achievement: Why poor black children succeed or Fail*. Westview press.

[152] Dzever, L. T. (2015). The impact of home environment factors on academic performance of senior secondary school students in Garki area district. *University of Ibadan, NIGERIA*.

[153] Purity, N. W. (2011). Socio-Economic factors influencing students performance in Kenya certificate secondary examination in Secondary school in Igembe South District, Estern Province, KENYA. *Kenyatta University Institutional Repository*, https://ir-library.ku.ac.ke/handle/123456789/7286.

[154] Kamuti, J. K. (2015). Influence of home environment on academic performance of students in Public Secondary school in Kitut West Sub country, KENYA. Master Thesis.

[155] Fan, X. (2001). Parental Involvement and Students' Academic Achievement: A Growth Modeling Analysis. *The Journal of Experimental Education*, 70(1), 27-61.

[156] Henton, J., Lamke, L., Murphy, C., & Haynes, L. (1980). Crisis reactions of college freshmen as a function of family support systems. *The Personnel and Guidance Journal*, 58(8), 508–511.

[157] Hoffman, J. A., & Weiss, B. (1987). Family dynamics and presenting problems in college students. *Journal of Counseling Psychology*, 34(2), 157-163.

[158] Rice, K. G., Cole, D. A., & Lapsley, D. K. (1990). Separation-individuation, family cohesion, and adjustment to college: Measurement validation and test of a theoretical model. *Journal of Counseling Psychology*, 37(2), 195-202.

[159] Strage, A., & Brandt, T. S. (1999). Authoritative parenting and college students' academic adjustment and success. *Journal of Educational Psychology*, 91(1), 146-156.

[160] Lewis, R., & J. Aiken, 1970. Attitudes toward Mathematics. *Review of Educational Research*, 40(4): 551-596.

[161] Smolleck L. A., & Duffy, C. G (2017). The Role of Negative Behavior on Children's Academic Performance in Early Childhood Education. *Interdisciplinary Education and Psychology*, 1(8), 1-12.

[162] Montroy, J. J., Bowles, R. P., Skibbe, L. E., & Foster, T. D. (2015). Social skills and problem behaviors as mediators of the relationship between behavioral self-regulation and academic achievement. *Early Childhood Research Quarterly*, 29, 298–309.

[163] Aranador, L. C., Khairiree, K., Hong, K. S., Michael, L. T. L., Lasom, A. F. (1998). *Mathematical problem solving abilities of Southeast Asia students*. Penang: SEAMEO Regional Centre for Education in Science and Mathematics.

[164] Cokadar,. H., & Kulge, G. (2008). Attitudes Towards Sciences: A Case of Turkey. *World Applied Science Journal*, 3(1), 102-109.

[165] McLeod, D. B. (1992). *Handbook of Research on Mathematics teaching and learning: "Research on affect in mathematics education: A reconceptualization" (pp. 575-579)*. D.A.Grows, Editor. Macmilland: New York.

[166] Stevens, T., Olivarez, A., Lan, W. Y., & Tallent-Runnels, M. K. (2004). Role of Mathematics Self-Efficacy and Motivation in Mathematics Performance Across Ethnicity. *The Journal of Educational Research*, 97(4), 208-222.

[167] Ashcraft, M. H., (2002). Math Anxiety: Personal, Educational, and Cognitive Consequences. *Current Directions in Psychological Science*, 11(5), 181-185.

[168] Akin, A., (2008). Self-Efficacy, Achievement Goals and Depression, Anxiety, and Stress: A structural Equation Modeling. *World Applied Sciences Journal*, 3(5), 725-732.

[169] Cates, G. L., & Rhymer, K. N. (2003). Examining the Relationship Between Mathematics Anxiety and Mathematics Performance: An Instructional Hierarchy Perspective. *Journal of Behavioral Education*, 12(1), 23-34.

[170] Hiebert, J., & Grouws, D. A. (2007). The Effects of Classroom Mathematics Teaching on Students' Learning. In F. Lester (Ed.), Second Handbook of Research on Mathematics Teaching and Learning (pp. 371-404). Charlotte, NC: Information Age.

[171] Stevenson, H. W., Lee, S. Y., Stigler, J. W. (1986). Mathematics Achievement of Chinese, Japanese and American Children. Science, 231(4739), 693-699.

[172] Maeyer, S. D., Bergh, H. V. D., Rymenans, R., Petegem, P. V., & Rijlaarsdam, G. (2010). Effectiveness criteria in school effectiveness studies: Future research on the choice for a multivariate model. *Educational research review*, 5(1), 81-96.

[173] Teddlie, C., & Reynolds, D. (2000). *School effectiveness research and the social and behavioral science in C. Teddlie & D. Reynolds (Eds)*. The international handbook of school effectiveness research, pp. 301-321, London/New York: Falmer Press.

[174] Yamat, H., Maarof, N., Maasum, T. N. T. M.,Zakaria, E., & Zainuddin, E. (2011). Teacher's Cod-Switching as Scaffolding in Teaching Content Area Subjects. *World Applied Science Journal 15 (Innovation and Pedagogy for Lifelong Learning)*, 15, 18-22.

[175] Matt, S. M., Zakaria, E., Nordin, N. M., & Meerah, T. S. M. (2011). Confirmatory Factor Analysis of the Mathematics Teachers' Teaching Practices Instrument. *World Applied Science Journal*, 12(11), 2092-2096.

[176] Ale, S. O. (1989). School mathematics in the 1990's some major problems for developing countries. *International Journal of Mathematical Education in Science and Technology*, 20(5), 655-659.

[177] Yılmaz, H., Çavaş, P. H. (2008). The Effect of the Teaching Practice on Pre-service Elementary Teachers' Science Teaching Efficacy and Classroom Management Beliefs. *Eurasia Journal of Mathematics, Science & Technology Education*, 4(1), 45-54.

[178] Adeogun, A. A., & Osifila, G. I. (2008). Relationship between educational resources and students' academic performance in Lagos State Nigeria. *International Journal of Education Management (IJEM)*, 5(6), 712-722.

[179] Arslan, B.,Demirel, O. (2008). Contribution to the student achievement of using different teaching methods in the fourth science and teaching course in primary education. *World Applied Sciences Journal*, 4(1), 37-43.

[180] Scriven, M. (1990). Can research-based teacher evaluation be saved?. *Journal of Personnel Evaluation in Education*, 4(1), 19-39.

[181] Wilkins, J. (2008). The relationship among elementary teachers' content knowledge, attitudes, beliefs, and practices. *Journal of Mathematics Teacher Education*, 11(2), 139-164.

[182] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22 140, 55.

[183] Vogt, W. P., & Jonhson, R. B. (1999). *Dictionary of statistics and methodology.* Sage: Thousand Oaks, California.

[184] Mellor, D., & Moore, K. A. (2014). The Use of Likert Scales with Children. *Journal of Pediatric Psychology* 39(3), 369–379.

[185] Dandekar, A., Zen, R. A. M., & Bressan, S. (December 2017). Comparative evaluation of synthetic data generation methods. *Deep Learning Security Workshop*, Singapore.

[186] Hu, J., J. P. Reiter, and Q. Wang (2014). Disclosure risk evaluation for fully synthetic data. In J. Domingo-Ferrer (Ed.), P*rivacy in Statistical Databases, Lecture Notes in Computer Science, PSD 2014*, Volume 8744, pp. 185–199. Springer International Publishing.

[187] Hittmeir, M., Ekelhart, A., & Mayer, R. (2018). On the utility of synthetic data: An empirical evolution on machine learning tasks. *In Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES 2019)*

(ARES 19), August 26–29, 2019, Canterbury, United Kingdom. ACM, New York, NY, USA.

[188] Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., and Sanil, A.P.(2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* 60, 3 (2006), 224–232.

[189] Reiter, J., A. Oganian, and A. Karr (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis*, 53, 1475–1482.

[190] Dreschler, J. and J. P. Reiter (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *Journal of Official Statistics*, 25(4), 589–603.

[191] Slavkovic, A. and J. Lee (2010). Synthetic two-way contongency tables that preserve conditional frequencies. *Statistical Methodology*, 7, 225–239.

[192] Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluation. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 5(2).

[193] Pardos, Z. A., Wang, Q. Y., & Trivedi, S. (2012). The real world significance of performance prediction. *Proceedings of the 5th International Conference on Educational Data Mining*, 192-195.

[194] Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature. Geoscientic Model Development, 7(3), 1247-1250.

[195] Lehmann, E. L., & Casella, George (1998). *Theory of Point Estimation (2nd ed.)*. New York: Springer. ISBN 0-387-98502

[196] Goodfellow, I., & Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[197] Zhuo, Z. H. (2012). *Ensemble Methods*. Chapman and Hall/CRC:Boca Raton, FL, USA.

[198] Ghorbani, R., and Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. *IEEEAccess*, 8, 67899-67911.

[199] Hussain, M., Zhu, W., Zhang, W., Abidi, S.M.R., & Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. *Artif. Intell. Rev.*, 52(1), 381-407.

[200] Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybern. Inf. Technol.*, 13(1), 61-72.

[201] Bergstra, J., & Bengio Y. (2012). Random search for hyperparameters. *J. Mach. Learn. Res.*, 13., 281-305.

[202] Papa, J.P., Schierer, W., & Cox, D.D. (2016). Fine-tuning deep belief network using Harmony search. *Applied Soft Computing*, 46(2016), 8875-885.

[203] Fernández-Redondo, M., & Hernández-Espinosa, C. (2001, April 25-27). Weight Initialization Methods for Multilayer Feedforward. *European Symposium on Artificial Neural Networks*, 119-124.

[204] Hebbo, H., and Kim, J. W. (2013). Classification with deep belief network. https://www.ki.tu-berlin.de/fileadmin/fg135/publikationen/Hebbo_2013_CDB.pdf

[205] Pinaya, W.H.L., Gadelha, A., Dpyle, O.M., Noto, C., Zugman, A., Cordeiro, Q., Jackowski, A.P., Bressan, R.A., & Sato, J.R. (2016). Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports*, 6,38897, doi:10.1038/srep38897.

[206] Lu, P., Guo, S., Zhang, H., Li, Q., Wang, Y., Wang., Y., & Qi, L. (2018). Research on improved deep belief network-based prediction of Cardiovascular disease. *Journal of Healthcare Engineering*, 2018, ArticleID 8954878.

[207] Song, S.H., & Kim, D.K. (2017). Development of a stress classification model using deep belief networks for stress monitoring. *Healthcare Informatics Research*, 23(4), 285-292.

[208] Sheela, K.G., & Deepa, S.N. (2013). Review on methods to fix number of hidden neurons in neural network. *Mathematical Problmes in Engineering*, 2013, ArticleID 425740.

[209] Li, C., Ding, Z., Yi, J., Lv, J., Zhang, G. (2018). Deep belief network based hybird model for building energy consumption prediction.

[210] Ghasemi, F., Mehridehnavi, A., Fassihi, A., & Sanchez, H.P. (2018). Deep neural network in QSAR studies using deep belief network. *Applied Soft Computing*, 62(2018), 251-258.

[211] Blanco, A.H., Flores, B.H., Tomas, D., Colorado, B.N. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, 2019, ArticleID 1306039.

[212] Mohamed, E., Najiba, T., & Monji, K. (2017). Optimization of DBN using Regularization Methods Applied for Recognizing Arabic Handwritten Script. *Procedia Computer Science*, 108, 2292-2297.

[213] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 40(6), ID: 21355.

[214] Mazumder, D. H., & Veilumuthu, R. (2018). An enhanced feature selection filter for classification of microarray cancer data. *Wiley ETRI Journal*, 358-370.

[215] Rachburee, N., & Punlumjeak, W. (2015). A Comparision of Feature Selection Approach Between Greedy, IGratio, Chi-square, and mRMR in Educational Data Mining. *7th International Conference on Information Technology and Electrical Engineering*, 420-424.

[216] Tsai, C. F., & Hsiao, Y. C. (2010). Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. *Decis Support System*, 50(1), 258-269.

[217] Uncu, O., & Turksen, I. B. (2007). A novel feature selection approach: combining featuer wrapper and filters. *Information Science*, 177(2), 449-466.

[218] Thubaity, A., Abanumay, N., Al-Jerayyed, S., Alrukban, A., & Mannaa, Z. (2013). The effect of combining different feature selection methods on Arabic text classification. *IEEE: The 14th ACIS International Conference Software Engineering, Artificial Intelligent, Networking and Parallel/distributed Computing (SNPD),* 211-216.

[219] Bahassine, S., Madani, A., Sarem, M. A., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Science*, 32, 225-231.

[220] Oulianroglou, S., & Evangelidis, G. (2015). Dealing with noisy data in the context of k-NN Classification. *Proceeding of the 7th Balkan Conference on Information Conference*, Article ID. 28, 1-4, https://doi.org/10.1145/2801081.2801116.

APPENDIX

កត្តាជះឥទ្ធិពលលើការសិក្សាមុខវិជ្ជាគណិតវិទ្យារបស់សិស្សនៅថ្នាក់វិទ្យាល័យ

ការអប់រំបានដើរតួនាទីយ៉ាងសំខាន់សំរាប់ការផ្លាស់ប្តូរកាគិតអោយកាន់តែប្រសើរនិងជាកត្តានាំទៅរកសង្គមមួយដែលរីកចំរើននិងសំបូរបែប។ ក្នុងនោះដែរមុខវិជ្ជាគណិតវិទ្យាត្រូវបានគេដឹងថាជាមុខវិជ្ជាមួយដ៏សំខាន់ និងជាមូលដ្ឋានគ្រឹះនៃមុខវិជ្ជាវិទ្យាសាស្ត្រនាៗ ក៍ដូចជាមុខជំនាញដទៃទៀតក្នុងសម័យកាលវិត្តន៍បច្ចេកទេស និងពត៍មានវិទ្យា។ កំរងសំនួរនេះធ្វើឡើងក្នុងគោលបំណងសំរាប់ការសិក្សាវិភាគលើកត្តាទាំងឡាយណាដែលជះឥទ្ធិពលលើការសិក្សាគណិតវិទ្យា។ ជាមួយគ្នានេះដែរ យើងធ្វើឡើងដើម្បីធ្វើការទាយលើលទ្ធផលរបស់សិស្សជាមុន ក្នុងគោលបំណងកែរលំអរអោយទាន់ពេលវេលា ។ សូមមេត្តាជួយផ្តល់ពត៍មានស្របទៅនឹងសំណួរខាងក្រោមដោយគ្រាន់តែចុចទៅលើចម្លើយតាមការគិតរបស់អ្នក។
បញ្ជាក់: រាល់ចម្លើយដែលអ្នកឆ្លើយ ស្របតាមការយល់ឃើញរបស់អ្នកនិងស្ថានភាពជាក់ស្តែងរបស់អ្នកដោយមិនស្ថិតក្នុងការវាយតំលៃថាខុសឬត្រូវឡើយ ពត៍មាននេះនឹងត្រូវបានប្រើប្រាស់ក្នុងគោលបំណងនៃការសិក្សាអប់រំតែប៉ុណ្ណោះ។

* Required

ពត៍មានផ្ទាល់ខ្លួន

សូមជួយផ្តល់ពត៍មានផ្ទាល់ខ្លួនរបស់អ្នកតាមរយៈសំនួរខាងក្រោម

1. **Gender: ភេទ**
   Mark only one oval.
   - ប្រុស
   - ស្រី

2. **Age: អាយុ**
   Mark only one oval.
   - ក្រោម 15 ឆ្នាំ
   - 15
   - 16
   - 17
   - 18
   - 19 ឡើងទៅ

3. **Siblings: ចំនួនបងប្អូន**
   Mark only one oval.
   - 1
   - 2
   - 3
   - 4
   - 5នាក់ឡើងទៅ

4. **Grade: កំរិតថ្នាក់រៀន** *
Mark only one oval.

- ថ្នាក់ទី១០
- ថ្នាក់ទី១១
- ថ្នាក់ទី១២

5. **School Sector : ប្រភេទសាលា** *
Mark only one oval.

- សាលារដ្ឋ
- សាលាឯកជន
- ផ្សេងៗ

កំរិតវប្បធម៌របស់អាណាព្យាបាលរបស់អ្នក

សូមជួយផ្តល់ពត៌មានអំពីកំរិតវប្បធម៌របស់អាណាព្យាបាលរបស់អ្នកខាងក្រោម

6. **PEDU1: កំរិតវប្បធម៌របស់ឪពុក** *
Mark only one oval.

- អត់ចេះអក្សរ
- រៀនបានត្រឹមបឋមសិក្សា
- រៀនបានត្រឹមអនុវិទ្យាល័យ
- រៀនបានត្រឹមវិទ្យាល័យ
- បរិញ្ញាប័ត្រ
- អនុបណ្ឌិត-បណ្ឌិត
- ផ្សេងៗ

7. **PEDU2: កំរិតវប្បធម៌របស់ម្តាយ** *
Mark only one oval.

- អត់ចេះអក្សរ
- រៀនបានត្រឹមបឋមសិក្សា
- រៀនបានត្រឹមអនុវិទ្យាល័យ
- រៀនបានត្រឹមវិទ្យាល័យ
- បរិញ្ញាប័ត្រ
- អនុបណ្ឌិត-បណ្ឌិត
- ផ្សេងៗ

មុខរបររបស់ឪពុកម្តាយ

សូមជួយផ្តល់ពត៌មានអំពីមុខរបររបស់អាណាព្យាបាលរបស់អ្នកខាងក្រោម

8. **POCC1: មុខរបររបស់ឪពុក** *
Mark only one oval.

- គ្មានមុខរបរច្បាស់លាស់
- កសិករ
- មន្ត្រីរាជការ
- បុគ្គលិកឯកជន
- មានជំនួញ ឬមុខរបរផ្ទាល់ខ្លួន
- ជរាភាព
- ផ្សេងៗ

9. **POCC2: មុខរបររបស់ម្តាយ** *
Mark only one oval.

- មេផ្ទះ
- កសិករ
- មន្ត្រីរាជការ
- បុគ្គលិកឯកជន
- មានជំនួញ ឬមុខរបរផ្ទាល់ខ្លួន
- ជរាភាព
- ផ្សេងៗ

កំរិតជីវភាពគ្រួសារ

1: ក្រីក្រខ្លាំង 2: ក្រីក្រ 3: មធ្យម 4: ធូរធា 5:ធូរធារខ្លាំង

10. **PSES1: តើអ្នកគិតថាគ្រួសាររបស់អ្នកមានកម្រិតចំណូលកំរិតណា?** *
Mark only one oval.

	1	2	3	4	5	
ក្រីក្រ	◯	◯	◯	◯	◯	ធូរធារខ្លាំង

ការចូលរួមចំណែករបស់ឪពុកម្តាយក្នុងការអប់រំកូន

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនដែលសោះ 2 = ម្តងម្កាល, 3 = ធម្មតា , 4 = ជាញឹកញាប់, 5 = តែងតែ

11. **PI1: តើអាណាព្យាបាលរបស់អ្នកយកចិត្តទុកដាក់លើការរៀនសូត្ររបស់អ្នកកំរិតណា?** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ/ មិនដែលខាន

12. **PI2: តើអាណាព្យាបាលរបស់អ្នកចំណាយពេលវេលានិងថវិកាសំរាប់ការរៀនសូត្ររបស់អ្នកក្នុងកំរិតណា?** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ/ មិនដែលខាន

13. **PI3: គ្រួសារតែងតែលើកទឹកចិត្តអោយខំរៀនសូត្រនិងអោយតំលៃលើការអប់រំ** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ/ មិនដែលខាន

ទំលាប់នៃការចិញ្ចឹមកូន

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនដែលសោះ 2 = ម្តងម្កាល, 3 = ធម្មតា , 4 = ជាញឹកញាប់, 5 = តែងតែ

14. **PS1: ឪពុកម្តាយរបស់អ្នកឆ្លើយតបនឹងអារម្មណ៍និងតំរូវការរបស់អ្នក** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ/ មិនដែលខាន

15. **PS2: ឪពុកម្តាយរបស់អ្នកជ្រាប់ពីអារម្មណ៍របស់ពួកគាត់ចំពោះអាកប្បកិរិយាល្អឬអាក្រក់របស់អ្នក** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ/ មិនដែលខាន

16. **PS3: ឪពុកម្តាយរបស់អ្នកលើកទឹកចិត្តអ្នកជ្រាប់ពីអារម្មណ៍និងបញ្ហារបស់អ្នក** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ/ មិនដែលខាន

17. **PS4: ពួកគាត់ផ្តល់នូវការសរសើរនៅពេលអ្នកធ្វើអ្វីមួយបានសំរេច?** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ/ មិនដែលខាន

កត្តាកន្លែងស្នាក់នៅ

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនឯកភាពទាំងស្រុង 2 = មិនឯកភាព, 3 = ធម្មតា , 4 = ឯកភាព, 5 = ឯកភាពទាំងស្រុង

18. **DE1: បរិស្ថាននៅផ្ទះបង្កភាពងាយស្រួលក្នុងការរៀនសូត្រ** *

Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

19. **DE2: ចំងាយពីផ្ទះមកសាលា** *

Mark only one oval.

- ◯ ឆ្ងាយ
- ◯ ល្មម
- ◯ ជិត

វិន័យនិងទម្លាប់នៃការរៀនសូត្រ

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនដែលសោះ 2 = ម្តងម្កាល, 3 = ធម្មតា , 4 = ជាញឹកញាប់, 5 = តែងតែ

20. **SELD1: តើអ្នកចំណាយពេលប៉ុន្មានម៉ោងក្នុងមួយថ្ងៃដើម្បីរៀនដោះស្រាយលំហាត់គណិតវិទ្យាដោយខ្លួនឯងក្រៅពីម៉ោងក្នុងថ្នាក់រៀន?** *

Mark only one oval.

- ◯ មិនដែលសោះ
- ◯ តិចជាង 2 ម៉ោង
- ◯ 2-4 ម៉ោង
- ◯ ច្រើនជាង 4 ម៉ោង

21. **SELD2: ក្នុងមួយសប្តាហ៍តើអ្នករៀនគួរគណិតវិទ្យាប៉ុន្មានម៉ោង?** *

Mark only one oval.

- ◯ មិនដែលសោះ
- ◯ តិចជាង 2 ម៉ោង
- ◯ 2-4 ម៉ោង
- ◯ ច្រើនជាង 4 ម៉ោង

22. **SELD3: តើអ្នកតែងតែធ្វើកិច្ចការផ្ទះ(Homework) ដែលគ្រូដាក់អោយឬទេ?** *

Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ

23. **SELD4: តើអ្នកតែងតែអវត្តមានឬខកខានឬមកយឺតម៉ោងគណិតវិទ្យាឬទេ?** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ

24. **SELD5: តើអ្នកតែងតែរំលឹកមេរៀននិងត្រៀមខ្លួនសំរាប់ការប្រលងនាឮៗនៃមុខវិជ្ជាគណិតវិទ្យាឬទេ?** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ

ការផ្តោតអារម្មណ៍ និងកម្លាំងជំរុញចិត្តចំពោះគណិតវិទ្យា

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនដែលសោះ 2 = ម្តងម្កាល, 3 = ធម្មតា , 4 = ជាញឹកញាប់, 5 = តែងតែ

25. **SIM1: តើអ្នកចូលចិត្តមុខវិជ្ជាគណិតវិទ្យាក្នុងកំរិតណា?** *
Mark only one oval.

	1	2	3	4	5	
អត់ចូលចិត្តសោះ	◯	◯	◯	◯	◯	ចូលចិត្តខ្លាំង

26. **SIM2: តើអ្នកតែងតែជំរុញខ្លួនឯងអោយខំរៀនដើម្បីបានពិន្ទុល្អនិងពូកែលើមុខវិជ្ជាគណិតវិទ្យាឬទេ?** *
Mark only one oval.

	1	2	3	4	5	
មិនដែល	◯	◯	◯	◯	◯	តែងតែ

27. **SIM3: តើអ្នកតែងតែផ្តោតអារម្មណ៍ស្តាប់គ្រូក្នុងម៉ោងគណិតវិទ្យាឬទេ?** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ

28. **SIM4: អ្នកតែងតែខំរៀនមុខវិជ្ជាគណិតវិទ្យាព្រោះយល់ដឹងពីសារៈសំខាន់របស់គណិតវិទ្យាចំពោះមុខវិជ្ជាផ្សេងៗនិងទៅអនាគត** *
Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

ភាពតានតឹងក្នុងម៉ោងគណិតវិទ្យា

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនដែលសោះ 2 = ម្តងម្កាល, 3 = ជាមួតា , 4 = ជាញឹកញាប់, 5 = តែងតែ

29. **ANXI1: អ្នកតែងតែមានអារម្មណ៍តានតឹងនិងធុញទ្រាន់ក្នុងម៉ោងគណិតវិទ្យា** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ

30. **ANXI2: អ្នកតែងតែមានអារម្មណ៍ភ័យក្នុងពេលប្រលងគណិតវិទ្យាម្តងៗ** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ

31. **ANXI3: អ្នកមានអារម្មណ៍ថាគ្មានអ្នកជួយពន្យល់រាល់ពេលដែលអ្នកមិនយល់** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ

សំភារៈប្រើប្រាស់ក្នុងការរៀនសូត្រ

សូមជ្រើសរើសចំលើយខាងក្រោម

32. **POSS1: អ្នកមាន Internet ដែលបង្កភាពងាយស្រួលក្នុងការស្រាវជ្រាវនិងរៀនសូត្រ** *
Mark only one oval.

- ◯ មាន
- ◯ គ្មាន

33. **POSS2: អ្នកមានកុំព្យូទ័រឬទេ** *
Mark only one oval.

- ◯ មាន
- ◯ គ្មាន

34. **POSS3: តើនៅផ្ទះរបស់អ្នកមានតុសំរាប់អង្គុយរៀនឬទេ?** *
Mark only one oval.

- ◯ មាន
- ◯ គ្មាន

វិធីសាស្ត្របង្រៀន

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនឯកភាពទាំងស្រុង 2 = មិនឯកភាព, 3 = ធម្មតា , 4 = ឯកភាព, 5 = ឯកភាពទាំងស្រុង

35. **TMP1: គ្រូគណិតវិទ្យារបស់អ្នកគ្រប់គ្រងសណ្តាប់ធ្នាប់ក្នុងថ្នាក់រៀនបានល្អ** *
Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

36. **TMP2: គ្រូគណិតវិទ្យារបស់អ្នកតែងតែយឺតឬអវត្តមានម៉ោងបង្រៀនឬទេ?** *
Mark only one oval.

	1	2	3	4	5	
មិនដែលសោះ	◯	◯	◯	◯	◯	តែងតែ

37. **TMP3: គ្រូគណិតវិទ្យារបស់អ្នកតែងតែប្រើគ្រប់វិធីសាស្ត្រដើម្បីពន្យល់អ្នក** *
Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

38. **TMP4: គ្រូគណិតវិទ្យារបស់អ្នកតែងតែឱ្យតំបង្រៀនអ្នកអស់ពីសមត្ថភាព** *
Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

អាកប្បកិរិយានិងអត្តចរិករបស់គ្រូគណិតវិទ្យារបស់អ្នក

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនឯកភាពទាំងស្រុង 2 = មិនឯកភាព, 3 = ធម្មតា , 4 = ឯកភាព, 5 = ឯកភាពទាំងស្រុង

39. **TAC1: គ្រូគណិតវិទ្យារបស់អ្នកមានសមត្ថភាពគ្រប់គ្រាន់ក្នុងការបង្រៀន** *
Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

40. **TAC2: ពួកគាត់តែងតែលើកទឹកចិត្តអ្នកអោយខំរៀនសូត្រ** *
Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

41. **TAC3: ពួកគាត់យកចិត្តទុកដាក់ចំពោះសិស្សគ្រប់គ្នា** *

Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

42. **TAC4: ពួកតែងតែជួយពន្យល់រាល់ពេលសិស្សត្រូវការជំនួយ** *

Mark only one oval.

	1	2	3	4	5	
មិនឯកភាពទាំងស្រុង	◯	◯	◯	◯	◯	ឯកភាពទាំងស្រុង

បរិស្ថានថ្នាក់រៀន

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនស្អាតទាល់តែសោះ, 2 = មិនសូវស្អាត , 3 = ធម្មតា, 4 = ស្អាតសមរម្យ, 5 = ស្អាតខ្លាំត

43. **CENV1: បរិស្ថានថ្នាក់រៀនអ្នកបែបណា** *

Mark only one oval.

	1	2	3	4	5	
មិនស្អាតទាល់តែសោះ	◯	◯	◯	◯	◯	ស្អាតខ្លាំង

កម្មវិធីសិក្សា

សូមជ្រើសរើសចំលើយខាងក្រោម : 1= មិនឯកភាពទាំងស្រុង 2 = មិនឯកភាព, 3 = ធម្មតា , 4 = ឯកភាព, 5 = ឯកភាពទាំងស្រុង

44. **CU1: ម៉ោងគណិតវិទ្យារបស់អ្នករៀនជាភាសា** *

Mark only one oval.

- ◯ ខ្មែរ
- ◯ អង់គ្លេស
- ◯ ជារាំង
- ◯ ផ្សេង

45. **CU2: កម្មវិធីសិក្សារបស់អ្នក** *

Mark only one oval.

- ◯ មួយវេន(ព្រឹកឬល្ងាច)
- ◯ មួយថ្ងៃពេញ
- ◯ ព្រឹកខ្លះល្ងាចខ្លះ
- ◯ ផ្សេងៗ

ធនធានសាលា

សូមផ្តល់នៅចំលើយខាងក្រោម

46. **ARES1: សាលាមានគ្រូបង្រៀនគ្រប់គ្រាន់** *
Mark only one oval.

- គ្រប់គ្រាន់
- មិនគ្រប់គ្រាន់
- មិនប្រាកដ

47. **ARES2: សាលាមានថ្នាក់រៀនគ្របគ្រាន់** *
Mark only one oval.

- គ្រប់គ្រាន់
- មិនគ្រប់គ្រាន់
- មិនប្រាកដ

48. **ARES3: បណ្ណាល័យសាលារបស់អ្នកមានសៀវភៅគ្រប់គ្រាន់សំរាប់អានឬខ្ចីឬទេ** *
Mark only one oval.

- គ្រប់គ្រាន់
- មិនគ្រប់គ្រាន់
- មិនប្រាកដ

រង្វាយតំលៃពិន្ទុគណិតវិទ្យារបស់អ្នក

សូមបំពេញក្នុងចន្លោះក្រោមសំណួរខាងក្រោម (ឧទាហរណ៍: 85) ។ ករណីដែលមិនចាំនូវពិន្ទុពិតប្រាកដអ្នកអាចបំពេញពិន្ទុប្រហាក់ប្រហែលទៅនឹងពិន្ទុដែលអ្នកតែងតែទទួលបានក្នុងមុខវិជ្ជាគណិតវិទ្យា

49. **SCORE: ពិន្ទុគណិតវិទ្យាឆមាសទីមួយរបស់អ្នក** *

ក្រោយពេលបំពេញរួចរាល់សូមមេត្តាចុចលើពាក្យ"SUBMIT"ឬ "ដាក់ស្នើ"ខាងក្រោមដើម្បីបញ្ជូន។ សូមអរគុណយ៉ាងជ្រាលជ្រៅសំរាប់ការចំណាយពេលវេលាក្នុងការបំពេញនូវកំរងសំនួរនេះ។

This is the end of questionnaire section

Powered by
Google Forms