機械学習を用いた太陽光エネルギーの出力を予測する説明的研究アプローチに関する研究

Doctoral Dissertation of Engineering (Science/Philosophy)

An Explanatory Study Approach, Using Machine Learning to Forecast Solar Energy Outcome

September 2022

by

Agada Ihuoma Nkechi

Interdisciplinary Intelligent Systems Engineering
Electrical and Electronics Engineering
Graduate School of Engineering and Science
University of the Ryukyus

Doctoral Dissertation of Engineering (Science/Philosophy)

An Explanatory Study Approach, Using Machine Learning to Forecast Solar Energy Outcome

September 2022

by

Agada Ihuoma Nkechi

Interdisciplinary Intelligent Systems Engineering
Electrical and Electronics Engineering
Graduate School of Engineering and Science
University of the Ryukyus

Supervisor: Prof. Nagata Yasunori

We, the undersigned, hereby, declare that we have read this thesis and we have attended the thesis defense and evaluation meeting. Therefore, we certify that, to the best of our knowledge this thesis is satisfactory to the scope and quality as a thesis for the degree of Doctor of Engineering (Science/Philosophy) in Area of study Interdisciplinary Intelligent System Engineering under Major Electrical and Electronics Engineering, Graduate School of Engineering and Science, University of the Ryukyus.

DISSERTATION REVIEW & EVALUATION COMMITTEE MEMBERS

Signature:_________________________

(Chairman) Professor Nagata Yasunori

Signature:_________________________

(Committee) Professor Urasaki Naomitsu

Signature:_________________________

(Committee) Asst. Professor Yona Atsushi

Abstract

Artificial intelligence (AI) techniques play a crucially important role in predicting the expected energy outcome and its performance, analysis, modeling and control of renewable energy. Solar energy usage has grown exponentially over the years. In the face of global energy consumption and increased depletion of most fossil fuel, the world is faced with the challenges of meeting the ever-increasing energy demands, also utility companies who provide solar energy have a challenge of unstable input of solar energy to the grid due to its intermittent nature, unlike other sources, hence the difference be- tween expected generation and actual generation, demand and supply can lead to an unbalanced grid. Therefore, incorporating accurately machine learning technology to predict the expected outcome of solar energy from the intermittent solar radiation will be crucial to keep a balance grid operation between supply and demand, production planning and energy management especially during installations of a photo-voltaic power plant. However, one of the major problems of forecasting is the algorithms used to control, model, and predict performances of the energy systems which are complicated and involves large computer power, differential equations, and time series. Also having unreliable data (poor quality) for solar radiation over a geographical location as well as insufficient long series can be a bottleneck to actualization. To overcome these problems, we employ the anaconda Navigator (Jupyter Notebook) for machine learning which can combine large amounts of data with fast, iterative processing and intelligent algorithms allowing the software to learn automatically from patterns or features to predict the performance and outcome of Solar Energy which in turns enables the balance between supply and demand on loads, efficient operation of the utility company as well as enhances power production planning and management.

Firstly, the thesis describes the need for alternative source of energy generation in developing countries. Most of the developing nations are facing low or no power supply especially in the rural areas. It shows the implementation of the micro-grid system, a battery storage device which is used for the modeling process to charge and discharge current, during high and low radiation of the solar energy respectively. Also, the use of a boost converter was considered to step up the DC current from the sun and maintain the output voltage, a bi-directional converter was used to allow two-way flow of current to charge and discharge the battery current. In the final stage an inverter with an inductor-capacitor-inductor filter (LCL) filter is used to convert current from DC to AC and filter harmonics considering the two main properties of the synchronous generator inertia and damping which is embedded in the controller of the inverter to provide more stability to the micro-grid.

Secondly, the thesis describes its main research which focuses on forecasting the output power of solar systems is required for the good operation of the power grid or for the optimal management of the energy fluxes occurring into the solar system. Before forecasting the solar systems output, it is essential to focus the prediction on the solar irradiance. The global solar radiation forecasting can be performed by several methods; the two big categories are the cloud imagery combined with physical models, and the machine learning models. In this thesis, the regression approach and time series methodology for prediction is used, the performance ranking of such methods is complicated due to the diversity of the data set, time step, forecasting horizon, set up and performance indicators.

Finally, the proposed methods are being summarized. Scopes of future research have also been described.

List of Publications

International Journal Paper: Published/Accepted

1. **Agada Ihuoma Nkechi**, Abdul Motin Howlader, and Atsushi Yona, “Integration of Photovoltaic Energy to the Grid, Using the Virtual Synchrnous Generator Control Technique”,*Journal of Energy and Power Engineering,* vol. 12, no. 7, pp. 329-339, July, 2018.

2. **Agada Ihuoma Nkechi**, Prof Nagata Yasunori “An Exolanatory study approach, Using Machine Learning to Forecast Solar Energy Outcome”,*Journal of Energy and Power Engineering,* vol. 16, pp.81-89, 2022 doi:10.17265/1934-8975/2022.02.004

Conference Paper: Presented

1. **Agada Ihuoma Nkechi**, Hidehito Matayoshi, and Atsushi Yona, “Evaluation of DC-DC Converter for Photovoltaic Renewable Energy, Institute of electrical engineers of Japan conference”, IEEJ 2017 Okinawa, Japan, Dec 10, 2017.

2. **Agada Ihuoma Nkechi**, Hidehito Matayoshi, Atsushi Yona and Tomonobu Senjyu “Fault Analysis of Synchrnoverter during Grid Integration”, International Workshop on Power Engineering in Remote Island International Workshop of power engineering in remote island 2017, Naha, Japan Feburary 12-15 2018.

3. **Agada Ihuoma Nkechi**, Atsushi Yona“Synchronverter Based Control Strategy for Photovoltaic Power Integration”, International Conference on Electrical Engineering 2018, Seoul, Korea, June 24-28, 2018

Award and Achievement

1. Japan International Cooperation Agency Scholarship, (JICA) 2016
2. Special Tuition Exemption for Academic Excellent Research, University of the Ryukyus, 2020

Acknowledgements

I would like to express cordial gratitude to my supervisor Prof. Nagata Yasunori for his dedicated direction, intelligent supervision and sharing of his vast knowledge of intelligent engineering system. He is a highly intellectual individual, creative, friendly, and inspiring, which keeps me motivated.

I would also like to give special thanks to my husband Mr. Chaka Benjamin, my beloved daughter Ms. Chizara Adele Benjamin for their continuous support towards my growth, well-being.

Also, I am grateful and express my cordial gratitude to Assistant Prof. Atsushi Yona for giving me directions and suggestions for improving my research and foundation from my master's degree.

I am grateful to my Parents Dr. Engr O.A Agada and Mrs. Patricia, Brother, Mr. Uchechukwu Agada and Sisters Ms. Chinyere Agada, Mrs. Uloma Achinanya and Ms. Kelechi Agada for their endless affection and support in every sphere of my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 History

The global shift towards renewable energy sources (RES) has driven the development of photo-voltaic (PV) panels. For example, the costs of producing electricity from PV panels have dropped significantly, while simultaneously increasing the energy conversion efficiency. More specifically, the leveled cost of electricity of large- scale PV panels has decreased between the years of 2010 and 2017 [1]. The decreased cost and increased efficiency have made PV panels a competitive alternative as a RES in many countries [2]. However, since PV panel energy output depend on weather conditions such as cloud cover and solar radiance, the energy output of the PV panels is unstable. To understand and manage the output variability is of interest for several actors in the energy market. In the short-term (0-5 hours), a transmission system operator is interested in the energy output from PV panels to find the adequate balance for the whole grid, since over- and under-producing electricity often results in penalty fees. On another side of the spectrum, electricity traders are interested in long time horizons, ordinarily, day-ahead forecasts since most electricity is traded on the day-ahead market. Consequently, the profitability of these operations relies on the ability to forecast the fluctuating solar PV panel energy output accurately. It is likely, as more countries decide to invest more and more in RES, that the use of solar PV panels will continue to increase. This will increase the need for suitable means of forecasting solar PV energy output. While the demand for accurate and efficient forecasts of PV panel energy output is evident, the solution is far from trivial. There are many complications that the current research within the field is handling. One evident nuisance is the inherited variation of weather, which makes accurate weather forecasting challenging. Parallel to the increased demand of PV power forecasting solutions, the means for forecasting with the help of machine learning (ML) techniques have in recent years gained in popularity relative to traditional time series predictive models. Al- though ML techniques are nothing new, the improved computational capacity and the higher availability of quality data have made the techniques useful for forecasting. This poses for an interesting area of research when forecasting the solar power output: How do machine learning techniques perform relative to traditional time series forecasting techniques?

1.1.2 The necessity to predict solar radiation or solar production

One of the most important challenges for the near future global energy supply will be the large integration of renewable energy sources (particularly non-predictable ones as wind and solar) into existing or future energy supply structure. An electrical operator should ensure a precise balance between the electricity production and consumption at any moment. As a matter of fact, the operator has often some difficulties to maintain this balance with conventional and controllable energy production system, mainly in small or not interconnected (isolated) electrical grid (as found in islands). The reliability of the electrical system then become dependent on the ability of the system to accommodate expected and unexpected changes (in production and consumption) and disturbances, while maintaining quality and continuity of service to the customers. Then, the energy supplier must manage the system with various temporal horizons. The integration of renewable energy into an electrical network intensifies the complexity of the grid management and the continuity of the production/consumption balance due to their intermittent and unpredictable nature. The intermittence and the non-controllable characteristics of the solar production bring a number of other problems such as voltage fluctuations, local power quality and stability issues [3, 4]. Thus, forecasting the output power of solar systems is required for the effective operation of the power grid or for the optimal management of the energy fluxes occurring into the solar system [5]. It is also necessary for estimating the reserves, for scheduling the power system, for congestion management, for the optimal management of the storage with the stochastic production and for trading the produced power in the electricity market and finally to achieve a reduction of the costs of electricity production. Due to the substantial increase of solar power generation the prediction of solar yields becomes more and more important [8]. In order to avoid large variations in renewable electricity production it is necessary to include also the complete prediction of system operation with storage solutions. Various storage systems are being developed and they are a viable solution for absorbing the excess power and energy produced by such systems (and releasing it in peak consumption periods), for bringing very short fluctuations and for maintaining the continuity of the power quality

1.1.3 Objectives

Our objective is to automate generating prediction models for smart homes that include on-site renewable. Prediction models are used by both the grid and the individual smart homes for advanced planning of electricity generation and consumption. Smart homes can use the models to potentially plan their consumption pattern to better match the power they generate on-site. The grid can use the models to plan generator dispatch schedules in advance as the fraction of renewable increases in the grid. Energy prediction must be done accurately to avoid shortage and surpluses. The more efficiently the predictions are the more efficiently the utility companies can operate. Nowadays solar panels are widely used to generate solar energy which is a very promising renewable energy source. With regards to solar electricity providers and a grid operator, it is critical to accurately predict solar power generation for supply-demand planning in an electrical grid, which directly affects their profit. Predicting solar output is, however, very difficult because solar power generation is dependent numerous weather features which is uncontrollable. This research proposes the technology of data mining using the anaconda navigator to be able to predict daily solar energy generation of any system. In this case we concentrated automatically generating models that correctly predicts solar energy generation based on the previous National Weather Service (NWS) weather forecast. Also, degradation of the solar panels is taking into consideration as this affects the outcome of solar energy generation.

It is also important to benchmark different forecasting techniques of solar energy output. Towards this end, machine learning and time series techniques can be used to dynamically learn the relationship between different weather conditions and the energy output of solar systems.

1.1.4 Applications of Data Mining

There is a rapidly growing body of successful applications in a wide range of areas as diverse as:

1. **weather forecasting**
2. **Analysing satellite imagery**
3. **Analysis of organic compounds**
4. **Automatic abstracting**
5. **Credit card fraud detection**

6. **Electric load prediction**
7. **Financial forecasting**
8. **Medical diagnosis**
9. **Predicting share of television audiences – product design**
10. **Real estate valuation**
11. **Targeted marketing**
12. **Text summation**
13. **Thermal power plant optimization**
14. **Toxic hazard analysis** and many more. Some examples of applications (potential or actual) are:
15. **a supermarket chain mines its customer transactions data to optimise targeting of high value customers**
16. **a credit card company can use its data warehouse of customer transactions for fraud detection**
17. **a major hotel chain can use survey databases to identify attributes of a high-value prospect**
18. **predicting the probability of default for consumer loan applications by im- proving the ability to predict bad loans**
19. **reducing fabrication flaws in VLSI chips**
20. **data mining systems can sift through vast quantities of data collected during the semiconductor fabrication process to identify conditions that are causing yield problems**
21. **predicting audience share for television programmes, allowing television ex-executives to arrange show schedules to maximize market share and increase advertising revenues**
22. **predicting the probability that a cancer patient will respond to chemotherapy, thus reducing health-care costs without affecting quality of care**
23. **analyzing motion-capture data for elderly people**
24. **trend mining and visualization in social networks**

Applications can be divided into four main types: classification, numerical prediction, association, and clustering. Each of these is explained briefly below. However first we need to distinguish between two types of data.

In general, we have a data set of examples (called instances), each of which comprises the values of a number of variables, which in data mining are often called attributes. There are two types of data, which are treated in radically different ways. For the first type there is a specially designated attribute, and the aim is to use the data given to predict the value of that attribute for instances that have not yet been seen. Data of this kind is called labeled. Data mining using labeled Introduction to Data Mining 5 data is known as supervised learning. If the designated attribute is categorical, i.e., it must take one of a number of distinct values such as very good, good or poor, or (in an object recognition application) car, bicycle, person, bus or taxi the task is called classification. If the designated attribute is numerical, e.g., the expected sale price of a house or the opening price of a share on tomorrow's stock market, the task is called regression. Data that does not have any specially designated attribute is called unlabeled. Data mining of unlabeled data is known as unsupervised learning. Here the aim is simply to extract the most information we can from the data available.

1.1.5 Research Question

Based on our introductory discussion, the problem can be summarized with the following research question: How do time series and machine learning techniques perform in short-term (0-5 hours) forecasting of solar PV energy output?

1.1.6 Limitations

Some limitations were done to clarify the scope of the study. Five established prediction models were chosen beforehand. The models that will be implemented and compared are Lasso, ARIMA, K-Nearest Neighbors (KNN), Gradient Boosting Regression Trees (GBRT), and Artificial Neural Networks (ANN). These models have been selected based on their tendency to perform well in previous research of energy forecasting. The focus will be placed in bench marking ML and time series techniques. Many of the above models are generic and therefore do most of them have a wide range of different model set-ups. The aim is to give a general overview of the relative performance of the methods rather than investigating a specific model in depth.

1.1.7 General research on Solar Power Forecasting

Solar power forecasting is the process of gathering and analyzing data in order to predict solar power generation on various time horizons with the goal to mitigate the impact of solar intermittency. Solar power forecasts are used for efficient management of the electric grid and for power trading.

As major barriers to solar energy implementation, such as materials cost and low conversion efficiency, continue to fall, issues of intermittency and reliability have come to the fore. The intermittency issue has been successfully addressed and mitigated by solar forecasting in many cases. Information used for the solar power forecast usually includes the Sun´s path, the atmospheric conditions, the scattering of light and the characteristics of the solar energy plant.

PV production mainly depends on the amount of solar global irradiation incident on the panels, but that irradiation is not uniform over time. Solar resource variability and the uncertainty associated to forecasts are behind most of the problems that must be handled to maintain the stability of the power grid. A part of the fluctuations is deterministic and explained by the rotational and translational movements of the Earth with respect to the Sun, which are accurately described by physical equations. However, there also exists unexpected changes in the amount of solar irradiance arriving at the Earth's surface, mainly derived from the presence of clouds, which stochastically block the Sun's rays and grant PV power forecasting a certain level of uncertainty.

The ability of precisely forecasting the energy produced by PV systems is of great importance and has been identified as one of the key challenges for massive PV integration (EPIA, 2012, PV GRID, 2014). It is decisive for grid operators, since deviations between forecasted and produced energy must be supplied by the rest of technologies that form the energy portfolio. Some of the units that build the electric system act as operating reserve generators. Thus, a proper PV forecast would be able to lower the number of units in hot standby and, consequently, reduce the operation costs. In general, there are various forecasting techniques used for energy forecasting. The study emphasizes that many different time series and ML techniques have been used in various energy contexts. The study is based on a com- petition for energy forecasting and evaluates the methods used in the competition (Global Energy Forecasting Competition 2014). The main conclusions are that for load demand forecasting, both machine learning and traditional time series techniques have been widely used. Similarly, for electricity prices, a broad mixture of time series and ML techniques have been employed. The article also provides an outline of techniques used in the competition for windmills and solar PV panels power forecasting. In contrast to load demand and electricity prices, the span of evaluated forecasting methods for windmills and solar PV panels power output is narrower. There has been some use of ML techniques while traditional time

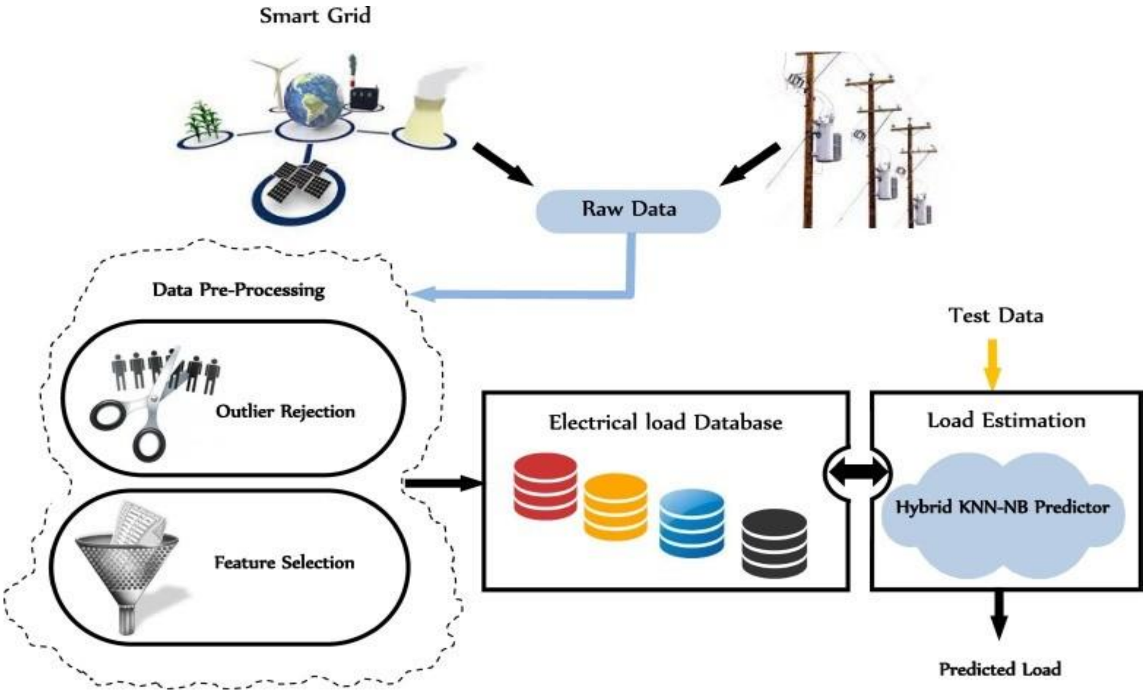


Figure 1.1: A Data Mining based Load Forecasting Strategy for Smart Electrical Grid.

series have been used to a great extent. For solar PV panels, the Random Forest algorithm, and the Support Vector Machine (SVM) algorithm were used.

1.1.8 The economics of forecasting

The main purpose of improving the accuracy of solar power forecasts is to reduce the uncertainties related to this type of variable energy source, which would directly result in a safer and easier grid management. Moreover, curtailment applied to photovoltaics could be reduced (Bird et al., 2014). Plant managers also find motivation in issuing better predictions as they can better plan maintenance stops and generate more precise bids.

1.1.9 Time series Algorithm

Reikard [7] presents a work regarding short-term horizons (0 to 4 hours), in which a benchmark of various time series models and ML models is provided. The main finding is that for short-term prediction, time series methods such as ARIMA out- performed ML and other time series models. According to Reikard, the main reason to why the model worked better is due to its ability to capture the transitions in radiance over a 24-hour period. Also, in a short-term perspective, the lagged radiance value does often reflect the momentary weather conditions to a large extent. The momentary weather condition is unlikely to change dramatically over a short period (in particular during days with stable weather) and does, therefore, provide a relatively accurate proxy of the future short-term energy output. In conclusion, for short-term forecasting, ARIMA and ARIMA with exogenous inputs will be relevant models to investigate as well as considering the use of lagged solar energy outputs. Madsen et al. [8], compare an auto-regressive model (AR), a linear regression (LR) model and an AR model with exogenous input (ARX model) and found a superiority of the ARX model over the other models when forecasting solar power output. In the study, the LR and ARX model included NWP as input data while AR only used historical solar PV panel energy output data. This study also highlights the importance of lagged power value in the short-term, as concluded in the work of Reikard [7].

1.1.10 Machine Learning Technique

Coimbra and Pedro [9] conducted a study where ARIMA, ANNs and KNN models were benchmarked and found, in contrast to the findings of Reikard, better accuracy with the ANN than the ARIMA in short-term forecasting. Coimbra and Pedro did, however, predict power output of solar PV panels rather than solar irradiance. In the study, only historical output levels of the panels were used, i.e., no exogenous NWP data was used as input. The authors also managed to attain accuracy improvements for the ANN model when using a genetic algorithm (an optimization algorithm inspired by the natural selection process) as their optimization algorithm. Ultimately, Coimbra and Pedro acknowledge the varying forecasting accuracy in different weather regimes. They suggest a division of data for different weather regimes when modeling. Here, the authors believe that fitting a specific model to a specific weather regime dataset may improve predictive results compared to using one fitted model to a data-set for all different weather regimes. Other researchers, such as Andrade et al. [2], explored ML techniques and evaluated them in combination with developing features that supposedly should improve the performance. The main approaches used in the study were Principal Component Analysis (PCA) and a feature engineering methodology in combination with a Gradient Boosting Tree model. Furthermore, the authors used different smoothing approaches to create features from their NWP data. Specifically, the authors used a grid of NWP data around the location of the PV installation and computed spatial averages and variances of weather parameters. Besides creating features based on a local grid of points, the authors also computed variances for different predictors for different lead times. When constructing variance features based on lead times, the underlying idea was that the feature would indicate the variability of the weather. The main conclusion is improved results from using both PCA and feature engineering. According to the authors, there is a twofold knowledge gap for the further research. The first aspect concerns feature management (feature engineering feature selection) and more concretely regarding how to create meaningful features that improve the forecast. The second aspect concerns the issue of further exploring ML modeling techniques that can be implemented in combination with informative features. Their final comment is that deep learning techniques will be an interesting path to pursue in combination with proper feature management. Davò et al. [10] used PCA combined with the techniques ANN and Analog Ensemble (AnEn) to predict solar radiance. With the aim to reduce the dimension of the dataset, PCA was used as a feature selection method. The dataset consists of the aggregated daily energy output of the solar radiation, measured over eight years. A comparison between using and not using PCA showed that using PCA in combination with ANN and AnEn enhances the prediction accuracy. Chen et al. present results on long-term (up to 100 hours) forecasting. The authors employed an ANN as their forecasting method with NWP data as input. The model was sensitive to prediction errors of the NWP input data and also showed a deterioration when forecasting on rainy days in particular. During cloudy and sunny days, the ANN model produced results with MAPEs of around 8 Persson et al. used Gradient Boosted Regression Trees (GBRT) to forecast solar energy generation 1-6 hours ahead. The data used was historical power output as well as meteorological features for 42 PV installations in Japan. Concerning RMSE, the GBRT model performed better than the adaptive recursive linear AR time series model, persistence model and climatology model on all forecast horizons. For shorter forecast horizons, it was shown that lagged power output values had a larger predictive ability. Similarly, for longer forecast horizons, the weather forecasts increased in importance. [12] Shi et al. propose an algorithm for weather classification and SVM to forecast PV power output on 15-minute intervals for the next-coming day. The weather is classified into four classes: clear sky, cloudy day, foggy day, and rainy day. The rationale behind the classification is correlation analysis on the local weather fore- casts and the PV power output. The data is thereafter normalized to reduce error and improve accuracy whilst still keeping correlation in the data. Four SVM models with radial basis function kernel are thereafter fitted to the four different weather classes. In conclusion, the result shows a way of using SVM models to train models on specific climatic conditions. [13]

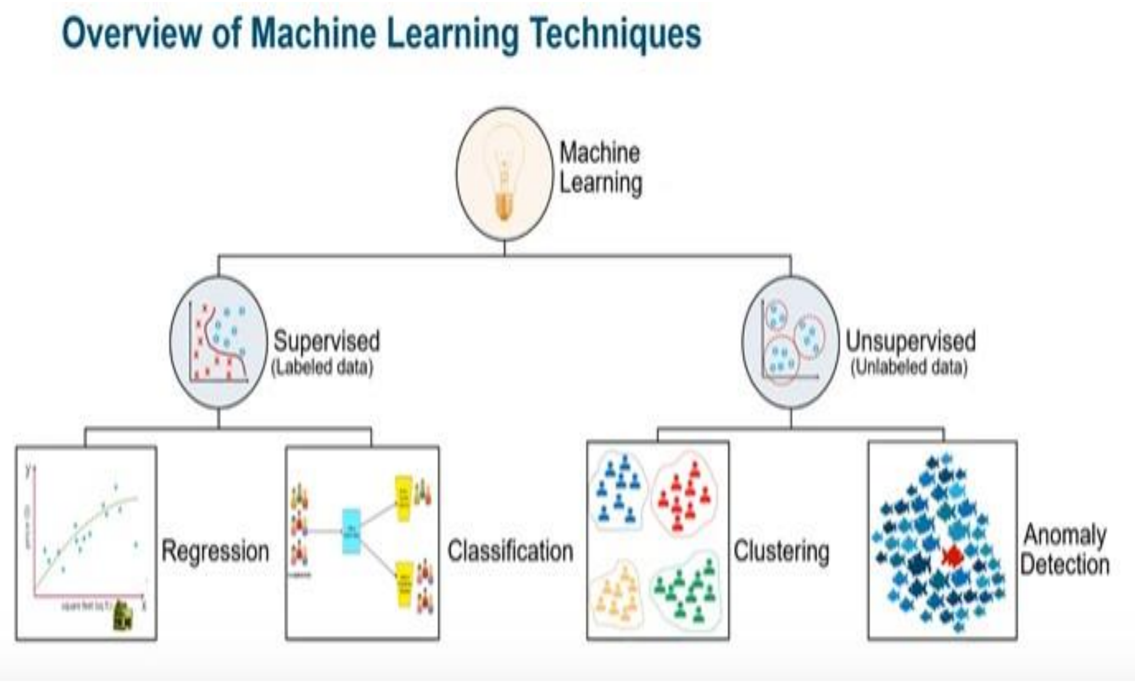


Figure 1.2: Overview of Machine Learning Technique.

1.1.11 Time Series Forecasting in Machine Learning Technique

Time Series is a certain sequence of data observations that a system collects within specific periods of time — e.g., daily, monthly, or yearly. The specialized models are used to analyze the collected time-series data — describe and interpret them, as well as make certain assumptions based on shifts and odds in the collection. These shifts and odds may include the switch of trends, seasonal spikes in demand, certain repetitive changes, or non-systematic shifts in usual patterns, etc.

All the previously, recently, and currently collected data is used as input for time series forecasting where future trends, seasonal changes, irregularities, and such are elaborated based on complex math-driven algorithms. And with machine learning, time series forecasting becomes faster, more precise, and more efficient in the long run. ML has proven to help better process both structured and unstructured data flows, swiftly capturing accurate patterns within massifs of data.

It is safe to say that time series machine learning principles basically outperform the classical time series forecasting approach. Thus, traditional methods are limited to processing only previously collected, readily available demand history. In turn, ML autonomously defines points of interest in the unlimited flow of data to then align them with customer data insights at hand and conduct what-if analysis. This results in particularly efficient takes on stimulating the demand in the commercial sector, for instance. However, this intricate predictive approach is beneficially used across numerous aspects of business management and optimization in most various niches.

1.1.12 Applications of Machine Learning Time Series Forecasting

Companies or organization dealing with constantly generated data and the need to adapt to operational shifts and changes can use time series forecasting. Machine learning serves as the ultimate booster here, allowing to better handle:

1. **Stock prices forecasting** the data on the history of stock prices combined with the data on both regular and irregular stock market spikes and drops can be used to gain insightful predictions of the most probable upcoming stock price shifts. Demand and sales forecasting — customer behavior patterns data along with inputs from the history of purchases, timeline of demand, seasonal

Demand and sales forecasting
Climate and weather prediction
Stock prices forecasting
Scientific studies forecasting
Web traffic forecasting
Demographic and economic forecasting

Figure 1.3: Applications of Machine Learning Time Series Forecasting

impact, etc., enable ML models to point out the most potentially demanded products and hit the spot in the dynamic market.

2. **Web traffic forecasting** Common data on usual traffic rates among competitor websites is bunched up with input data on traffic-related patterns in order to predict web traffic rates during certain periods.

3. **Climate and weather prediction** Time-based data is regularly gathered from numerous interconnected weather stations worldwide, while ML techniques allow to thoroughly analyze and interpret it for future forecasts based on statistical dynamics.

4. **Demographic and economic forecasting** There are tons of statistical inputs in demographics and economics, which is most efficiently used for ML-based time-series predictions. As a result, the most fitting target audience can be picked and the most efficient ways to interact with that TA can be elaborated.

5. **Scientific studies forecasting** Machine Learning and deep learning principles accelerate the rates of polishing up and introducing scientific innovations dramatically. For instance, science data that requires an indefinite number of analytic iterations can be processed much faster with the help of patterns automated by machine learning.

1.1.13 Legacy Method of Time Series Forecasting

1. **Recurrent Neural Network (RNN)** RNNs process a time series step-by-step, maintaining an internal state from time-step to time-step. Neural networks are great in this application as they can learn the temporal dependence from the given data. And considering input sequences from the temporal perspective opens horizons for more precise predictions. However, the method is considered a legacy because the “education” of neural networks can be too time-consuming.

2. **Long Short-Term Memory (LSTM)** It's kind of RNN, but while maintaining RNN's ability to learn the temporal dynamics of sequential data, LSTM can furthermore handle the vanishing and exploding gradients problem. Thus, complex multivariate data sequences can be accurately modeled, and the need to establish pre-specified time windows (which solves many tasks that feed-forward networks cannot solve). The downside of overly time-consuming supervised learning, however, remains.

1.1.14 Classic Methods of Time-Series Forecasting

1. **Multi-Layer Perceptron (MLP)** Univariate models can be used to model univariate time series forecasting problems. Multivariate MLP models use multivariate data where there is more than one observation for each time step. Then, there are multistep MLP models — there is little difference to the MLP model in predicting a vector output that represents different output variables or a vector output that represents multiple time steps of one variable. This is a very widely used method that even outperforms LSTM in certain autoregression instances.

2. **ARIMA)** Autoregression employs observations collected during previous time steps as input data for making regression equations that help predict the value to be generated at the next time step. ARIMA or an Autoregressive Integrated Moving Average model combines autoregression and moving average principles, making forecasts correspond to linear combinations of past variable values and forecast errors, being one of the most popular approaches due to that.

3. **Bayesian Neural Network (BNN)** BNN models involve constructing a prior distribution and updating this distribution by conditioning on the actual data. This is particularly useful for financial data because of its volatile nature, as nonlinear time series forecasting with machine learning is enabled. BNN treats network weights or parameters as random variables, being among the most universally used models out there.

4. **Radial Basis Functions Neural Network (RBFNN)** RBF Neural Network is based on the function approximation theory or supervised and unsupervised manner was used together. Similar to BNN, RBF models are used for forecasting nonlinear time series. RBFNN model proves to be best for predicting daily network traffic, which makes it pretty popular among commercial forecasting applications.

5. **Kernel regression or Generalized Regression Neural Network (GRNN)** Generalized regression neural network (GRNN) is a branch of the RBF neural network. Recent research activities in forecasting with GRNN suggest that GRNN can be a promising alternative to the traditional time series model. It has shown great ability in modeling and forecasting nonlinear time series, and it is gradually entering the lines of multipurpose, commonly used methods.

6. **K-Nearest Neighbor Regression Neural Network (KNN)** The k-nearest neighbor (k-NN) algorithm is one of the most popular non-parametric approaches used for classification, and it has been extended to regression. KNN is a supervised machine learning method that consists of instances, features, and targets components. The selection of the number of neighbors and feature selection is a daunting task. KNN is a simple algorithm that has been effectively used in various research areas such as financial modeling, image interpolation, and visual recognition.

7. **CART Regression Trees (CART)** The technique is aimed at producing rules that predict the value of an outcome (target) variable from known values of predictor (explanatory) variables. They show good prediction accuracy performance, but they cannot detect and adapt to change or concept drift well. This approach is certainly strong in terms of unsupervised practices, but it still lacks maturity.

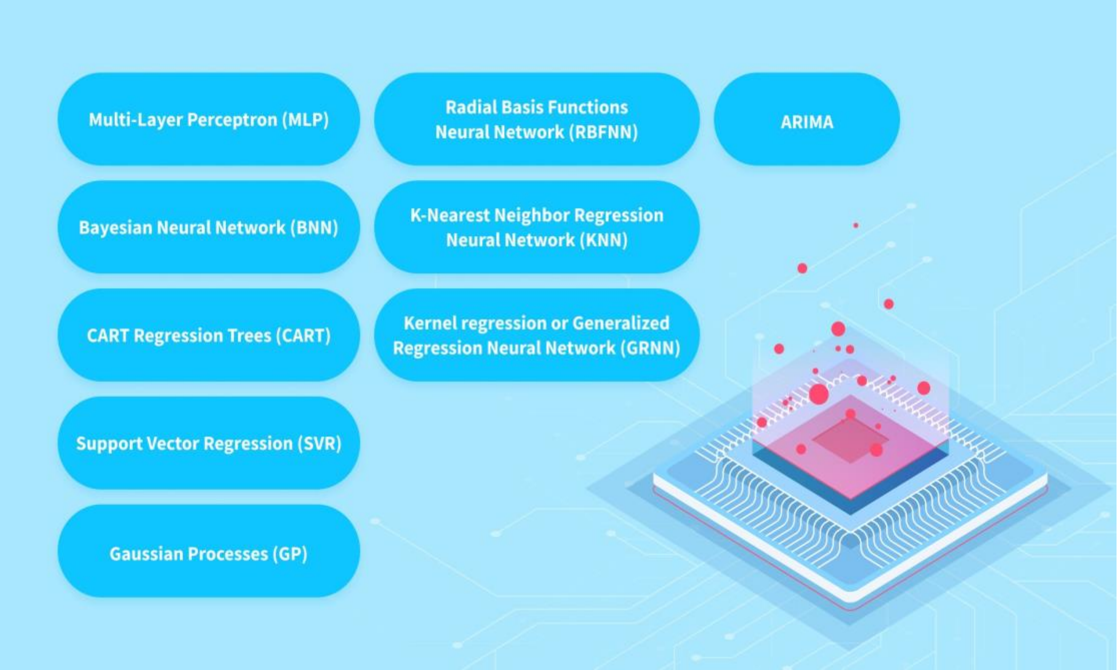

Multi-Layer Perceptron (MLP)
Radial Basis Functions Neural Network (RBFNN)
ARIMA
Bayesian Neural Network (BNN)
K-Nearest Neighbor Regression Neural Network (KNN)
CART Regression Trees (CART)
Kernel regression or Generalized Regression Neural Network (GRNN)
Support Vector Regression (SVR)
Gaussian Processes (GP)

Figure 1.4: Classic Methods of Time-Series Forecasting

8. **Support Vector Regression (SVR)** (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. The ability of SVM to solve nonlinear regression estimation problems makes SVM quite successful in time series forecasting.

9. **Gaussian Processes (GP)** Gaussian process (GP), as one of the cornerstones of Bayesian non-parametric methods, has received extensive attention in machine learning time series prediction. GP has intrinsic advantages in data modeling, given its construction in the framework of Bayesian hierarchical modeling and no requirement for a prior information of function forms in Bayesian reference.

1.1.15 Topical Methods of Time Series Forecasting

Convolutional Neural Network (CNN) Although analysis of image datasets is considered their main field of application, convolutional neural networks can show even better results than RNNs in time series prediction cases involving other types of spatial data. For one thing, they learn faster, boosting the overall data processing performance. However, CNN's can also be joined with RNNs to get the best of both worlds — i.e., a CNN easily recognizes spatial data and passes it to RNN for temporal data storing.

Attention Mechanism This is one of the basic principles of deep learning that can be adapted in terms of different forecasting models. In a nutshell, it mimics the human brain in terms of focusing attention on specific elements that stand out from a bunch. This enables a deep neural network to concentrate only on relevant data points among the barrage of various inputs, boosting the efficiency of NLP and Computer Vision.

Transformer Neural Networks A transformer neural network is an advanced architecture focused on solving sequence-to-sequence tasks. Its main goal is also to easily handle long-range dependencies. Such networks are quite popular in ML-based models, simplifying regression by simply customizing the loss function. This comes in more than handy when it comes to regressions.

Kaggle Kaggle is a coding and data processing environment where efficient web traffic time series forecasting can be carried out. This is an engine with technical capabilities contributed by an extensive community of enthusiasts over the years. This makes it an efficient tool for tackling the issue of predicting future values of multiple time series.

LightGBM This one is a widely used ML algorithm that is mostly focused on capturing complex patterns within tabular datasets. This results in quite efficient sales data predictions. In certain instances, LightGBM outruns the classical ARIMA approach in terms of making tabular-based predictions. However, both should be applied in individual situations to make out the best.

Decision Trees ML-based decision trees are used to classify items (products) in the database. Generated classes get dedicated multivariate time series models that help predict the future price of a certain item. Obviously, this one is best for commercial analyses.

XGBoost This is the applied machine learning algorithm that works with tabular and structured data. In its core, lie gradient-boosted decision trees. Working with XGBoost requires one to transform time series datasets into supervised learning problems. But the results should be worth it.

AdaBoost This type of forecasting algorithm is deemed as the best out-of-the-box classifier by many. This means that it is best used at elaborating data classifications in conjunction with other efficient algorithms. For instance, when used with decision trees, it learns to outline the hardest-to-classify data instances over time.

1.2 Step-by-Step Process of Time Series Forecasting Using Machine Learning

1.2.1 Preparation Stage

1. **Project goal definition** Start with the comprehensive outline and understanding of minor and major milestones and goals. This is where preliminary research should be carried out to most properly tackle business area specifics.

2. **Data gathering and exploration** Continuing with thorough preparation, specific data types to be analyzed and processed must be settled. Data visualization charts and plot graphs can be used for this. Data preparation and time series decomposition — specialized professionals should clean all involved data, gaining valuable insights and extracting proper variables. These variables can then be used for time series decomposition.

1.2.2 Modeling Stage

1. **Forecasting models evaluation** Based on all the preliminary research and prep data, different forecasting models are tested and evaluated to pick the most efficient one(s).

2. **Forecasting model training and performance estimation** The picked machine learning algorithms for time series are then optimized through cross-validation and trained.

1.2.3 Testing Stage

1. **Forecasting models run on testing data with known results** A step necessary for making sure the picked algorithms do their work properly.

2. **Accuracy and performance optimization** The last phase of polishing up algorithms to achieve the best forecasting speed and accuracy.

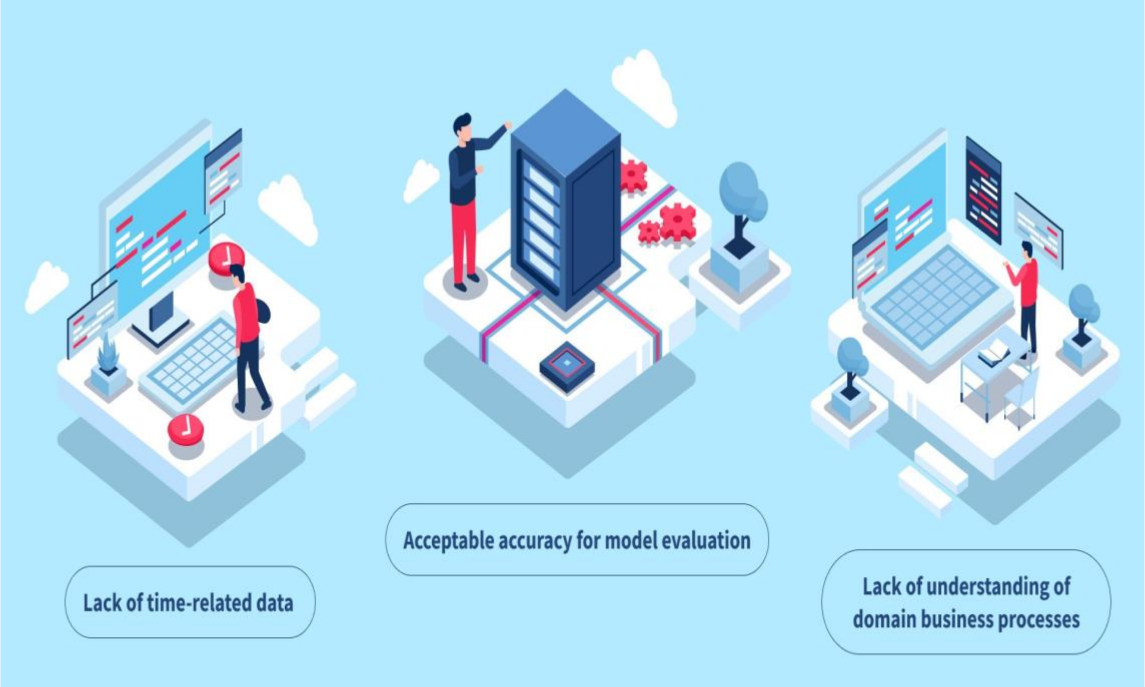

Lack of time-related data
Acceptable accuracy for model evaluation
Lack of understanding of domain business processes

Figure 1.5: Using Machine Learning for Time-Series Forecasting

1.2.4 Deployment Stage

1. **Data transformation and visualization** To integrate the resulting forecasting model(s) with the production at hand, the gathered data must be conveniently transformed and visualized for further processing.

2. **Forecasting models revision and improvements** Time series forecasting is always iterative, meaning that multiple ongoing revisions and optimizations must be implemented to continuously improve the forecasting performance.

1.2.5 Key Challenges of Forecasting Time Series with Machine Learning Models

The below section should give a basic idea of how a time series forecasting project is structured and done. However, keep in mind that you may (and will) come across certain common challenges of using machine learning for time series in the process. These include the following.

Lack of Time-Related Data The more training data a system can extract from datasets, the higher predictive accuracy can be achieved. You may, however, experience a lack of seasonality/historical data for a target variable when working with ML, which limits the system's learning capacity.

Acceptable Accuracy for Model Evaluation You may have to experiment a lot in attempts to achieve the highest forecasting efficiency. This is where a very knowledgeable approach to evaluating the most accurate predictive models is a must.

Lack of Understanding of Domain Business Processes Only experienced niche specialists can han- dle ML feature engineering. You simply cannot tackle forecasting using machine learning algorithms without proper domain expertise.

1.2.6 Related Work

Solar energy forecasts can be categorized in a variety of ways. The persistence or smart persistence model, which uses historical data to forecast future power generation over a short period of time, is the most basic method (2-3 hours). This method can be used to set a standard against which other forecasting methods can be measured. In most cases, a prediction is completed in two stages. A NWP is designed for a specified time period and location to begin with. The generated NWP is then utilized to forecast power generation using forecasting algorithms. It is possible to employ a physical model, a statistical method, or a machine learning methodology. For prediction, ML algorithms are compared to the Smart Persistence (SP) approach, with ML models outperforming the SP model. The unpredictability of so- lar resources has hampered grid management as solar diffusion rates have increased. Unpredictability and intermittent electricity delivery are two of the most difficult aspects of integrating renewables into the system. As a result, solar power forecasting is becoming increasingly important for grid stability, optimal unit commitment, and cost-effective dispatch. To overcome the problem, we employ machine learning techniques to sift through extraordinary solar radiation predicting models. For developing pre- diction models, a variety of regression algorithms are tested, including linear least squares and support vector machines with various kernel functions. We use day-ahead sun radiation data forecasts in these tests to show that a machine learning approach can correctly anticipate short-term solar power. A hybrid or mixed forecasting method was developed by combining clustering, classification, and regression approaches to produce a forecasting model. Based on the weather forecast for the next day, the model (with the closest weather condition) is chosen to forecast the power output using cluster-wise regression . Renewable energy sources are progressively being integrated into electric networks alongside nonrenewable energy sources, posing significant issues due to their sporadic and erratic nature in order to address these issues, soft-computing solutions for energy prediction are essential. We apply a number of data mining methodologies, including preparing historical load data and analyzing the features of the load time series, because electricity consumption is entangled with the usage of other energy sources like natural gas and oil. The trends in power consumption from renewable and nonrenewable energy sources were examined and contrasted. A novel machine learning-based hybrid technique (SVR) uses multilayer perceptron (MLP) and support vector regression. Using SVM regression, solar power generation produces acceptable results. However, it lacks a detailed examination of solar power generation and meteorological data, and hence is restricted in its capacity to accurately predict other data sets by merely using different SVM kernels after some basic statistical data processing.

1.3 Outline of Thesis

Considering the works have been conducted, this thesis is divided into *four* chapters (Including this chapter). Each chapter outline is following-

Chapter 2 Review previous work done on photovoltaic energy and relating it to forecasting demonstrates various types of forecasting techniques *24th International Conference on Electrical Engineering (ICEE), 2018, Korea.*

Chapter 3 The employing of the anaconda Navigator (Jupyter Notebook) for machine learning which can combine large amounts of data with fast, iterative processing and intelligent algorithms allowing the software to learn automatically from patterns or features to predict the performance and outcome of Solar Energy which in turns enables the balance between supply and demand on loads. showing results of predicted result from learned data *Applied Intelligence*.

Chapter 4 is decorated with concluding remarks and future research based on the recent interests for artificial intelligence with relation to solar energy system.

Bibliography

[1] IRENA. Renewable power generation costs in 2017. Technical report, Interna- tional Renewable Energy Agency, Abu Dhabi, January 2018.

[2] Jose R. Andrade and Ricardo J. Bessa. Improving renewable energy forecasting with a grid of numerical weather predictions. IEEE Transactions on Sustainable Energy, 8(4):1571–1580, October 2017.

[3] Rich H. Inman, Hugo T.C. Pedro, and Carlos F.M. Coimbra. Solar forecasting methods for renewable energy integration. Progress in Energy and Combustion Science, 39(6):535 – 576, 2013.

[4] “J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. Solar Energy, 136:78–111, October 2016.

[5] Kostylev, A Pavlovski, et al. Solar power forecasting performance–towards industry standards. In 1st international workshop on the integration of solar power into power systems, Aarhus, Denmark, 2011.

[6] Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. International Journal of Forecasting, 32(3):896 – 913, 2016.

[7] Gordon Reikard. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Solar Energy, 83(3):342 – 349, 2009.

[8] Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online short-term solar power forecasting. Solar Energy, 83(10):1772 – 1783, 2009.

[9] Hugo T.C. Pedro and Carlos F.M. Coimbra. Assessment of forecasting tech- niques for solar power production with no exogenous inputs. Solar Energy, 86(7):2017 – 2028, 2012.

[10] Federica Davò, Stefano Alessandrini, Simone Sperati, Luca Delle Monache, Da- vide Airoldi, and Maria T. Vespucci. Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. Solar Energy, 134:327 – 338, 2016.

[11] Changsong Chen, Shanxu Duan, Tao Cai, and Bangyin Liu. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Solar Energy, 85(11):2856 – 2870, 2011.

[12] Caroline Persson, Peder Bacher, Takahiro Shiga, and Henrik Madsen. Multi-site solar power forecasting using gradient boosted regression trees. Solar Energy, 150:423 – 436, 2017.

[13] J. Shi, W. J. Lee, Y. Liu, Y. Yang, and P. Wang. Forecasting power out- put of photovoltaic systems based on weather classification and support vector machines. IEEE Transactions on Industry Applications, 48(3):1064–1069, May 2012.

[14] Peter J Brockwell. Introduction to Time Series and Forecasting. Springer Texts in Statistics. Springer, 3rd ed. 2016. edition, 2016.

[15] T. Hastie J. Gareth, D. Witten and R. Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer texts in statistics. An introduction to statistical learning. Springer, 2013.

[16] T. Hastie, R. Tibshirani, and J.H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statis- tics. Springer, 2001.

[17] Jerome H. Friedman. Stochastic gradient boosting. Computational Statistics Data Analysis, 38(4):367 – 378, 2002. Nonlinear Methods and Data Mining.

[18] D.Randall Wilson and Tony R. Martinez. The general inefficiency of batch training for gradient descent learning. Neural Networks, 16(10):1429 – 1451, 2003.

[19] Christoph Bergmeir and Jose´ M. Ben´ıtez. Neural networks in R using the stuttgart neural network simulator: RSNNS. Journal of Statistical Software, 46(7):1–26, 2012.

[20] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. caret: Classification and Regression Training, 2017. R package version 6.0-78.

Chapter 2

Previous Research and Methods of Data Mining

2.1 Background

Data accumulation of solar energy generation has been on a steady rise at an almost unimaginable rate from a very wide variety of sources from general solar radiation emitted every day to its usage by solar converters for energy generation. The use of data mining cannot be overemphasized alongside advances in storage technology, which increasingly make it possible to store such vast amounts of data at relatively low cost whether in commercial data warehouses, scientific research laboratories or elsewhere, has come a growing realization that such data contains buried within it knowledge that can be critical to a company growth or decline, knowledge that could lead to important discoveries in science, knowledge that could enable us accurately to predict the weather, solar outcomes, natural disasters and more. Yet the huge volumes involved mean that most of this data is merely stored never to be examined in more than the most superficial way, if at all. It has rightly been said that the world is becoming data rich but knowledge poor. Solar energy is increasing exponentially making it one of the most popular renewable forms of energy. The irradiance is a measure of the energy available to enter a solar PV system, if the irradiance of a location is known over a period, it can be used to predict future solar energy generation at that location.

An electrical operator should ensure a precise balance between the electricity production and consumption at any moment. This is often very difficult to maintain with conventional and controllable energy production system, mainly in small or not interconnected (isolated) electrical grid (as found in islands). Many countries nowadays consider using renewable energy sources into their electricity grid. This creates even more problems as the resource (solar radiation, wind, etc.) is not steady. It is therefore very important to be able to predict the solar radiation effectively especially in case of high energy integration [1].

2.1. The necessity to predict solar radiation or solar production. One of the most important challenges for the near future global energy supply will be the large integration of renewable energy sources (particularly non-predictable ones as wind and solar) into existing or future energy supply structure. An electrical operator should ensure a precise balance between the electricity production and consumption at any moment. As a matter of fact, the operator has often some difficulties to maintain this balance with conventional and controllable energy production system, mainly in small or not interconnected (isolated) electrical grid (as found in islands). The reliability of the electrical system then become dependent on the ability of the system to accommodate expected and unexpected changes (in production and consumption) and disturbances, while maintaining quality and continuity of service to the customers.

The integration of renewable energy into an electrical network intensifies the complexity of the grid management and the continuity of the production/consumption balance due to their intermittent and unpredictable nature [1, 2]. The intermittence and the non-controllable characteristics of the solar production bring a number of other problems such as voltage fluctuations, local power quality and stability issues [3, 4]. Thus, forecasting the output power of solar systems is required for the effective operation of the power grid or for the optimal management of the energy fluxes occurring into the solar system [5]. It is also necessary for estimating the reserves, for scheduling the power system, for congestion management, for the optimal management of the storage with the stochastic production and for trading the produced power in the electricity market and finally to achieve a reduction of the costs of electricity production [6, 7]. Due to the substantial increase of solar power generation the prediction of solar yields becomes more and more important [8] In order to avoid large variations in renewable electricity production it is necessary to include also the complete prediction of system operation with storage solutions. Various storage systems are being developed and they are a viable solution for absorbing the excess power and energy produced by such systems (and releasing it in peak consumption periods), for bringing very short fluctuations and for maintaining the continuity of the power quality. These storage options are usually classified into three categories:

1. **Bulk energy storage** Bulk energy storage or energy management storage media is used to decouple the timing of generation and consumption.

2. **Distributed generation or bridging power** - this method is used for peaks shaving - the storage is used for a few minutes to a few hours to assure the continuity of service during the energy sources modification.

3. The power quality storage the power quality storage with a time scale of about several seconds is used only to assure the continuity of the end use power quality.

It is important to note that the energy storage acts at various time levels and their appropriate management requires the knowledge of the power or energy produced by the solar system at various horizons: very short or short for power quality category to hourly or daily for bulk energy storage. Similarly, the electrical operator needs to know the future production at various time horizons.

2.2 Review of the hybrid DC/AC solar microgrid model conversion

Power electronics devices are used to convey and harness solar energy as it ensures maximum power tracking of current and voltage. The output active power of the panels is dependent on solar radiation and temperature for power generation. The lithium battery is used for power compensation during low solar radiation. This, in turn, smoothen the real power output and minimizes adverse impact on the grid. The system comprises the use of solar panels, boost converter to step up voltage and keep it constant, battery bank, buck-boost converter to charge and discharge the battery at high and low radiation respectively as seen in Fig.1.

2.2.1 Photovoltaic System

A photovoltaic system, also PV system or solar power system, is a power system designed to supply usable solar power by means of photovoltaics. It consists of an arrangement of several components, including solar panels to absorb and convert sunlight into electricity, a solar inverter to change the electric current from DC to AC, as well as mounting, cabling, and other electrical accessories to set up a working system. It may also use a solar tracking system to improve the system's overall performance. Basic equations of the PV are used for modelling and simulation of the PV array on Simulink. The boost converter is connected to the PV to step-up and stabilize the input voltage from the PV. Although PV voltage range, has impractical limited range but the PV voltage variation value given as 350V-370V in this paper is used for simulation. The step-up converter increases and maintain the voltage for the required input voltage needed for the inverter.

2.2.2 Battery Energy Storage System

Battery Energy Storage Systems (BESSs) are a sub-set of Energy Storage Systems (ESSs). ESS is a general term for the ability of a system to store energy using thermal, electro-mechanical or electro-chemical solutions. A BESS utilizes an electro-chemical solution. Essentially, all Energy Storage Systems capture energy and store it for use later. Examples of these systems include pumped hydro, compressed air storage, mechanical flywheels, and BESSs. These systems complement intermittent sources of energy such as wind, tidal and solar power to balance energy production and consumption. Energy storage results in a reduction in peak electrical system demand and ESS owners are often compensated through regional grid market programs. Regulators also offer incentives (and in some cases mandates) to encourage participation. BESSs use electro-chemical solutions and include some of the following types of batteries.

- Lithium-ion These offer good energy storage for their size and can be charged/discharged many times in their lifetime. They are used in a wide variety of consumer electronics such as smart-phones, tablets, laptops, electronic cigarettes, and digital cameras. They are also used in electric cars and some aircraft. Most home battery storage systems coming onto the market use lithium-ion (or Li-ion) technology.

- Lead-acid These are traditional rechargeable batteries and are inexpensive compared to newer types of batteries. Uses include protection and control systems, back-up power supplies, and grid energy storage. Lead acid batteries are the most common battery type. They work similarly to your car battery and have been used to provide back-up power in blackouts and in remote areas.

- Flow Battery Flow batteries are quite large and are generally used to store energy from renewable sources, they store energy in liquid electrolyte solutions. They tend to be used in commercial and large-scale applications but can be used in homes.

2.2.3 Boost Converter

Boost Converter is a DC-to-DC power converter that steps up voltage while stepping down current from the input source to its output supply. For the integration of renewable energy (Photovoltaic) to the grid, the need of a device to stabilize and step-up output voltage is very important due to the variation in solar radiation and temperature. Therefore, the boost convert is a key component for the functionality of this system This can be seen in Fig. 2.1. For the boost converter the switching device used is the IGBT (integrated gate bipolar Transistor) as we are considering a high voltage output. The IGBT is used basically when the output voltage is higher than 200V. for the Boost Converter the IGBT and MOSFET are the main switching device used. A simple explanation for the DC-DC Boost converter occurs for when the switching device (IGBT) is ON and OFF. When the transistor is ON current flows through the inductor creating a magnetic field, it is important to note that input voltage is equal to current in the inductor Ii=Il (current in series is the same) when the IGBT is turned OFF, the magnetic field created is destroyed and this induces an EMF (faraday’s first law of electromagnetic induction). the induced emf flows through the diode, at transistor OFF diode starts to conduct and the emf is stored in the capacitor. The diode prevents the charge from flowing back to the circuit and this occurs in microseconds, therefore continuous switching ON and OFF the transistor generates a higher voltage output than the voltage input supply.

2.2.4 The Buck-boost converter

The buck-boost converter is a type of DC-to-DC converter that has an output voltage magnitude that is either greater than or less than the input voltage magnitude. It is equivalent to a fly back converter using a single inductor instead of a transformer. In a buck-boost DC-DC converter topology, the output voltage can be higher or lower than the input voltage. Often used in battery-operated equipment, buck-boost

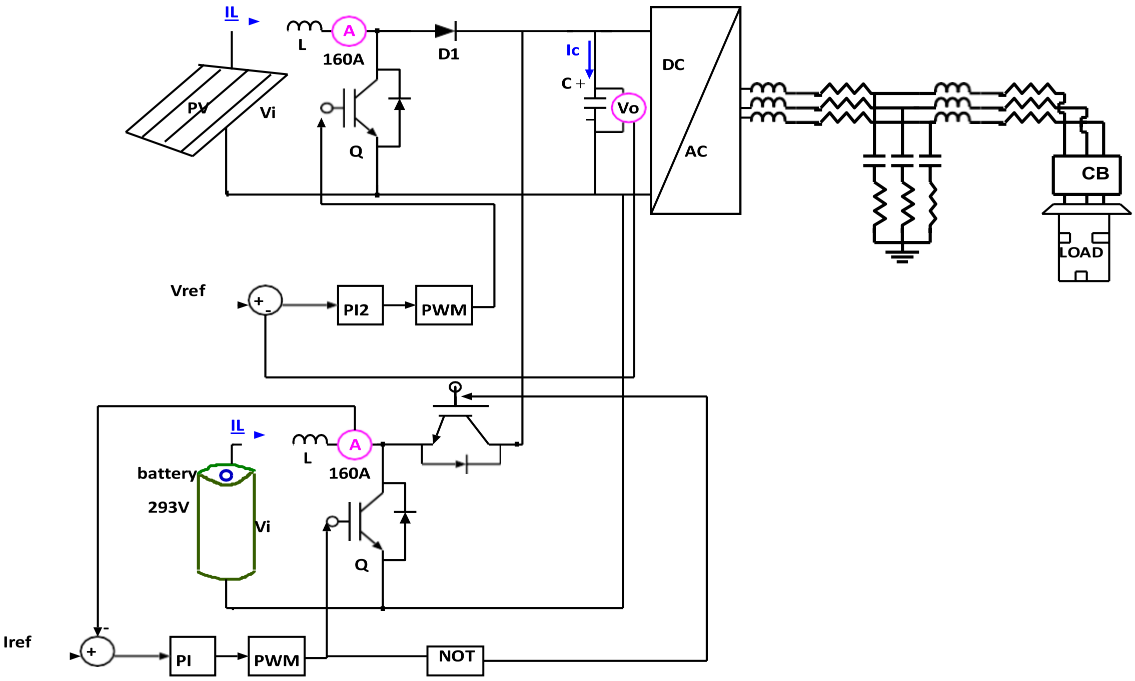


Figure 2.1: A Detailed model of the micro grid.

DC-DC converters require high-efficiency and ultra-low stand-by current as well as a small size to fit the requirements for portable and wearable devices for the Internet of Things (IoT). In the inverting topology, the output voltage is of the opposite polarity than the input. This is a switched-mode power supply with a similar circuit topology to the boost converter and the buck converter. The output voltage is adjustable based on the duty cycle of the switching transistor. One possible drawback of this converter is that the switch does not have a terminal at ground; this complicates the driving circuitry. However, this drawback is of no consequence if the power supply is isolated from the load circuit (if, for example, the supply is a battery) because the supply and diode polarity can simply be reversed. When they can be reversed, the switch can be on either the ground side or the supply side. When a buck (step-down) converter is combined with a boost (step-up) converter, the output voltage is typically of the same polarity of the input and can be lower or higher than the input. Such a non-inverting buck-boost converter may use a single inductor which is used for both the buck inductor mode and the boost inductor mode, using switches instead of diodes sometimes called a" four-switch buck-boost converter", it may use multiple inductors but only a single switch as in the SEPIC and Ćuk topologies.

2.2.5 Inverter

An inverter is one of the most important pieces of equipment in a solar energy system. It's a device that converts direct current (DC) electricity, which is what a solar panel generates, to alternating current (AC) electricity, which the electrical grid uses. In DC, electricity is maintained at constant voltage in one direction. In AC, electricity flows in both directions in the circuit as the voltage changes from positive to negative. Inverters are just one example of a class of devices called power electronics that regulate the flow of electrical power. Fundamentally, an inverter accomplishes the DC-to-AC conversion by switching the direction of a DC input back and forth very rapidly. As a result, a DC input becomes an AC output. In addition, filters and other electronics can be used to produce a voltage that varies as a clean, repeating sine wave that can be injected into the power grid. The sine wave is a shape or pattern the voltage makes over time, and it's the pattern of power that the grid can use without damaging electrical equipment, which is built to operate at certain frequencies and voltages.

2.2.6 Available Forecasting Method

The solar power forecasting can be performed by several methods; the two big categories are the cloud imagery combined with physical models, and the machine learning models. The choice for the method to be used depends mainly on the prediction horizon; actually, all the models have not the same accuracy in terms of the horizon used. Various approaches exist to forecast solar irradiance depending on the target forecasting time. The literature classifies these methods in two classes of techniques: Extrapolation and statistical processes using satellite images or measurements on the ground level and sky images are generally suitable for short-term forecasts up to six hours. This class can be divided in two sub-classes, in the very short time domain called "Now-casting" (0–3 h), the forecast has to be based on extrapolations of real-time measurements [5]; in the Short-Term Forecasting (3–6h), Numerical Weather Prediction (NWP) models are coupled with post-processing modules in combination with real-time measurements or satellite data [11]. NWP models able to forecast up to two days ahead or beyond [12, 13] (up to 6 days ahead [13]). These NWP models are sometimes combined with post-processing modules and satellite information are often used. The NWP models predict the probability of local cloud formation and then predict indirectly the transmitted radiation using a dynamic atmosphere model. The extrapolation or statistical models analyze historical time series of global irradiation, from satellite remote sensing [15] or ground measurements by estimating the motion of clouds and project their impact in the future. Hybrid methods can improve some aspects of all of these methods [6, 14]. The statistical approach allows to forecast hourly solar irradiation (or at a lower time step) and NWP models use explanatory variables (mainly cloud motion and direction derived from atmosphere) to predict global irradiation N-steps ahead.

2.2.6.1 Machine Learning Methods

Machine learning is a sub-field of computer science, and it is classified as an artificial intelligence method. It can be used in several domains and the advantage of this method is that a model can solve problems which are impossible to be represented by explicit algorithms. In the reader can find a detailed review of some machine learning and deterministic methods for solar forecasting. The machine learning models find relations between inputs and outputs even if the representation is impossible; this characteristic allow the use of machine learning models in many cases, for example in pattern recognition, classification problems, spam filtering, and also in data mining and forecasting problems. The classification and the data mining are particularly interesting in this domain because one has to work with big datasets and the task of pre-processing and data preparation can be undertaken by the machine learning models. After this step, the machine learning models can be used in forecasting problems. In global horizontal irradiance forecasting the models can be used in three different ways: structural models which are based on other meteorological and geographical parameters; time-series models which only consider the historically observed data of solar irradiance as input features (endogenous forecasting); - hybrid models which consider both, solar irradiance and other variables as exogenous variables (exogenous forecasting). As already mentioned, machine learning is a branch of artificial intelligence. It concerns the construction and study of systems that can learn from data sets, giving computers the ability to learn without being explicitly programmed

1. **Discriminant analysis and Principal Component Analysis (PCA)** The principal component analysis (PCA) is a statistical method which uses an orthogonal transformation to transform a set of observations of probably correlated variables into a set of values of linearly uncorrelated variables which are called principal components. The number of principal components created in the process, is lower or equal to the number of original variables. Such transformation is defined in such a way so as the first principal component has the largest variance possible, i.e., to account for the maximum variability in the data, and each subsequent component to have the highest variance possible under the restriction that it is orthogonal to the previous components. As a result, the resulting vectors form an uncorrelated orthogonal basis set. It should be noted that the prin

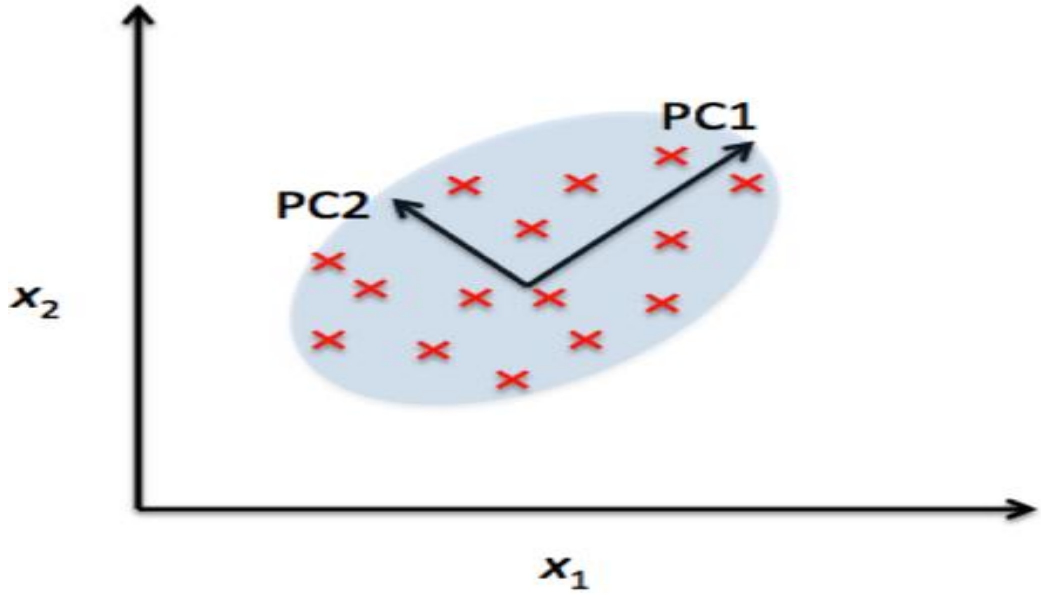


Figure 2.2: A Model of Discriminant analysis and Principal Component Analysis.

cipal components are orthogonal as they are the eigenvectors of the co-variance matrix, which is symmetric. Moreover, PCA is sensitive to the relative scaling of the original variables.

2. **Naive Bayes classification and Bayesian networks** in machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters proportional to the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form ex- pression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers [27]. A Bayesian network, also called Bayes network, Bayesian model, belief network or probabilistic directed a cyclic graphical model is a probabilistic graphical model, which is a type of statistical model that represents a set of random variables and their conditional dependencies via a directed a cyclic graph (DAG).

3. **Data mining approach** A data mining consists of the discovery of interesting, unexpected, or valuable structure in large data sets that can be called with the slogan Big Data. In other words, data mining consists of extracting the most important information from a very large data set. Indeed, the classical statistical inference has been developed for processing small samples. In the presence of very large databases, all the standard statistical indexes become significant and thus interesting (e.g. for 1 million of data, the significance threshold of correlation coefficient is very low reaching 0.002,...). Additionally, in data mining, data collected are analyzed for highlighting the main information before to use them in the forecasting models. Rather than opposing data mining and statistics, it is best to assume that data mining is the branch of statistics devoted to the exploitation of large databases. The techniques used are from different fields depending on classical statistics and artificial intelligence. This last notion was defined by "The construction of computer programs that engage in tasks that are, for now, more satisfactorily performed by humans because they require high-level mental processes such as perceptual learning organization memory and critical thinking ". There is not really a consensus of this definition, and many other similar ones are available.

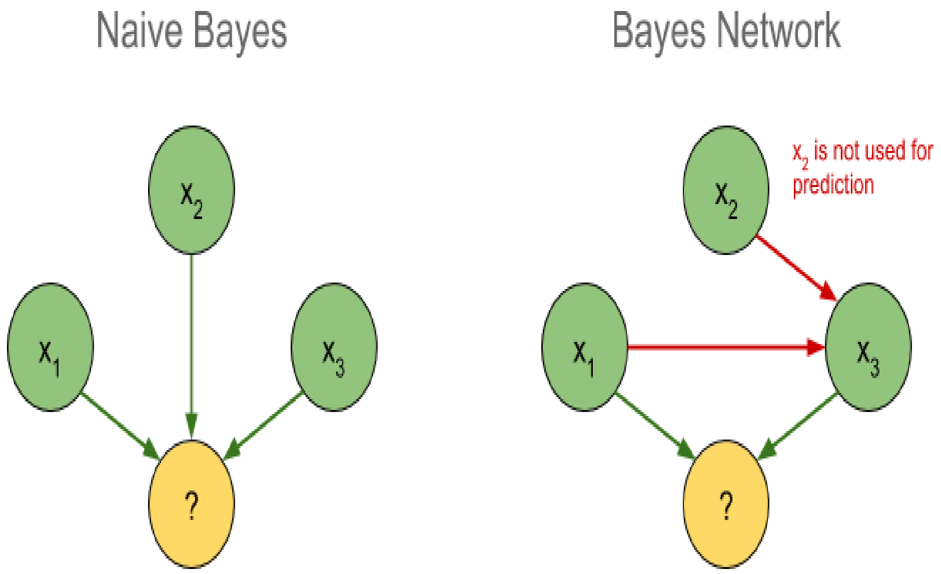


Figure 2.3: A Model of Naive and Bayesian Network.

2.2.6.2 Supervised learning

In supervised learning, the computer is presented with example inputs and their desired outputs, given by a" teacher", and the goal is to learn a general rule that maps inputs to outputs. These methods need an "expert " intervention. The training data comprise of a set of training examples. In supervised learning, each pattern is a pair which includes an input object and a desired output value. The function of the supervised learning algorithm is to analyze the training data and produce an inferred function.

1. **Linear Regression** Early attempts to study time series, particularly in the 19th century, were generally characterized by the idea of a deterministic world. It was the major contribution of Yule (1927) which launched the idea of stochasticity in time series by assuming that every time series can be regarded as the realization of a stochastic process. Based on this simple idea, a number of time series methods have been developed since that time. Workers such as Slutsky, Walker, Yaglom, and Yule first formulated the concept of autoregressive (AR) and moving average (MA) models. World's decomposition theorem led to the formulation and solution of the linear forecasting problem of Kolmogorov in 1941. Since then, a considerable amount of literature is published in the area of time series, dealing with parameter estimation, identification, model checking and forecasting; see, for example ref. For an early survey.

2. **Generalized Linear Models** Generalized linear model (GLM) in statistics, is a flexible generalization of ordinary linear regression which allows for response variables that have error distribution models other than a normal distribution. Generalizes linear regression by permitting the linear model to be related to the response variable through a link function and by considering the magnitude of the variance of each measurement to be a function of its predicted value [9]. Some studies improve the regression quality using a coupling with other predictors like Kalman filter.

3. **Nonlinear Regression** Artificial Neural Networks (ANN) are being increasingly used for nonlinear regression and classification problems in meteorology due to their usefulness in data analysis and prediction. The use of ANN is particularly predominant in the realm of time series forecasting with nonlinear methods. Actually, the availability of historical data on the meteorological utility

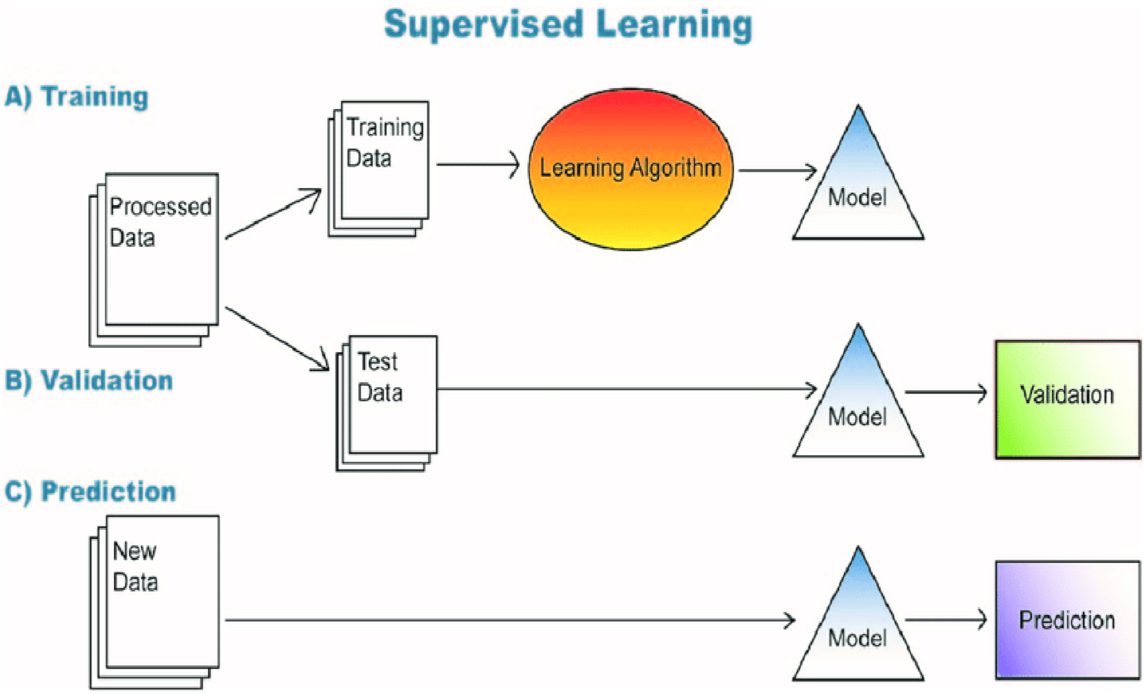


Figure 2.4: A Model of Supervised Learning.

databases and the fact that ANNs are data driven methods capable of performing a non-linear mapping between sets of input and output variables makes this modelling software tool very attractive.

4. **Support Vector Machines / Suppor Vector Regression** Support vector machine is another kernel-based machine learning technique used in classification tasks and regression problems introduced by Vapnik in 1986. Support vector regression (SVR) is based on the application of support vector machines to regression problems [18]. This method has been successfully applied to time series forecasting tasks.

5. **Decision tree learning (Breiman bagging)** The basic idea is very simple. A response or class Y from inputs X1, X2,...., Xp is required to be predicted. This is done by growing a binary tree. At each node in the tree, a test to one of the inputs, say Xi is applied. Depending on the outcome of the test, either the left or the right sub-branch of the tree is selected. Eventually a leaf node is reached, where a prediction is made. This prediction aggregates or averages all the training data points which reach that leaf. A model is obtained by using each of the independent variables. For each of the individual variables, mean squared error is used to determine the best split. The maximum number of features to be considered at each split is set to the total number of features [10–14].

6. **Nearest neighbor** Nearest neighbor neural network (k-NN) is a type of instance-based learning, where a function is only approximated locally, and all computation is delayed until classification . The k-NN algorithm is one of the simplest machine learning algorithms. For both classification and regression, it can be useful to assign a weight to the contributions of the neighbors, so that the nearest neighbors contribute more to the average than the distant ones. For example, in a common weighting arrangement, each neighbor is given a weight of 1/d, where d is the distance to the neighbor.

7. **Markov chain** in forecasting domain, some authors have tried to use the so-called Markov processes, specifically the Markov chains. A Markov process is a stochastic process with the Markov property, which means that given the present state, future states are independent of the past states. Expressed differently, the description of the present state fully captures all the information that

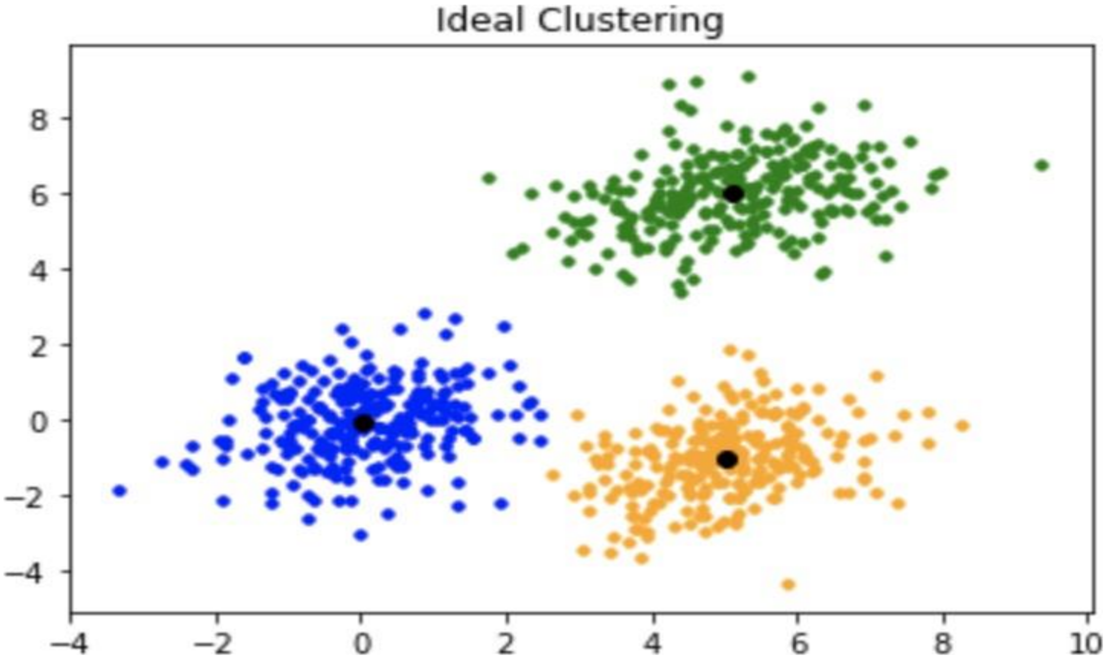


Figure 2.5: A Model of Unsupervised Learning.

could affect the future evolution of the process. In this, future states are reached through a probabilistic process instead of a deterministic one. The proper use of these processes needs to calculate initially the matrix of transition states. The transition probability of state i to the state j is defined by pi, j. The family of these numbers is called the transition matrix of the Markov chain R.

2.2.6.3 Unsupervised learning

In contrary with supervised learning model, an unsupervised learning model does not need an "expert" intervention and the model is able to find hidden structure in its inputs without knowledge of outputs. Unsupervised learning is similar to the problem of density estimation in statistics. Unsupervised learning, however, also incorporates many other techniques that seek to summarize and explain the key features of the data. Many methods normally employed in unsupervised learning are based on data mining methods used to pre-process data.

1. **k-Means and k-Methods Clustering** k-means clustering is a method of vector quantization, originally derived from signal processing, which is popular for cluster analysis in data mining. k-means clustering aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. k-Means algorithms are focused on extracting useful information from the data with the purpose of modelling the time series behavior and find patterns of the input space by clustering the data. Furthermore, nonlinear autoregressive (NAR) neural networks are powerful computational models for modelling and forecasting nonlinear time series. A lot of methods of clustering are available; the interested reader can see for more information.

2. **Hierarchical Clustering** In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis) is a method of cluster analysis which seeks to build a hierarchy of clusters. Hierarchical clustering creates a hierarchy of clusters which can be represented in a tree structure called "dendrogram" which includes both roots and leaves. The root of the tree consists of a single cluster which contains all observations, whereas the leaves correspond to individual observations. Algorithms for hierarchical clustering are generally either agglomerative, in which the process

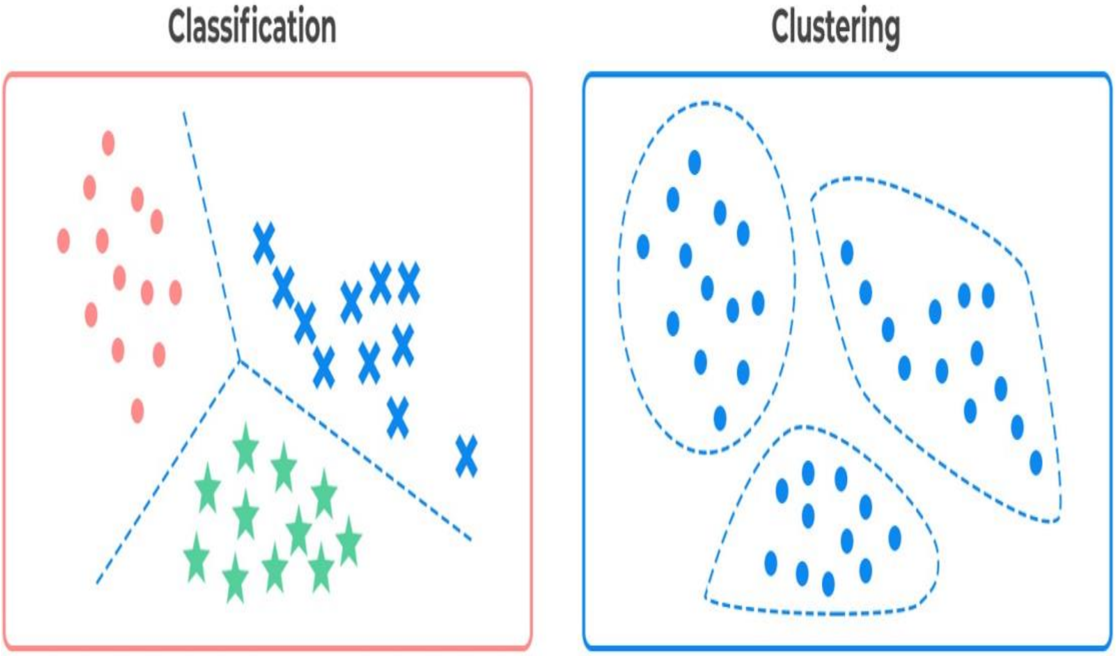


Figure 2.6: A Model of Classification and Clustering.

starts from the leaves and successively merges clusters together; or divisive, in which the process starts from the root and recursively splits the clusters. Any function which does not have a negative value can be used as a measure of similarity between pairs of observations. The choice of which clusters to merge or split, is determined by a linkage criterion that is a function of the pairwise distances between observations. It should be noted that cutting the tree at a given height will give a clustering at a selected precision.

3. **Cluster Evaluation** Typical objective functions in clustering formalize the goal of attaining high intra-cluster similarity and low inter-cluster similarity. This is an internal criterion for the quality of a clustering. Good scores on an internal criterion do not necessarily mean a good effectiveness in an application. An alternative to internal criteria is the direct evaluation of the application of interest. For search result clustering, the amount of the time a user is required to find an answer with different clustering algorithms may be required. This is the most direct evaluation, but it is time consuming, especially if large number of studies are necessary.

2.2.6.4 Ensemble learning

The basic concept of ensemble learning is to train multiple base learners as ensemble members and combine their predictions into a single output that should have better performance on average than any other ensemble member with uncorrelated error on the target data sets. Supervised learning algorithms are usually described as performing the task of searching through a hypothesis space to find a suitable hypothesis that can perform good predictions for a particular problem. Even if the hypothesis space contains hypotheses that are very well-matched for a particular problem, it may be very difficult to find which one is the best. Ensembles combine multiple hypotheses to create a better hypothesis. The term ensemble is usually used for methods that generate multiple hypotheses using the same base learner. Fast algorithms such as decision trees are usually used with ensembles, although slower algorithms can also benefit from ensemble techniques. Evaluating the prediction accuracy of an ensemble typically requires more computation time than evaluating the prediction accuracy of a single model, so ensembles may be considered as a way to compensate for poor learning algorithms by performing much more computation.

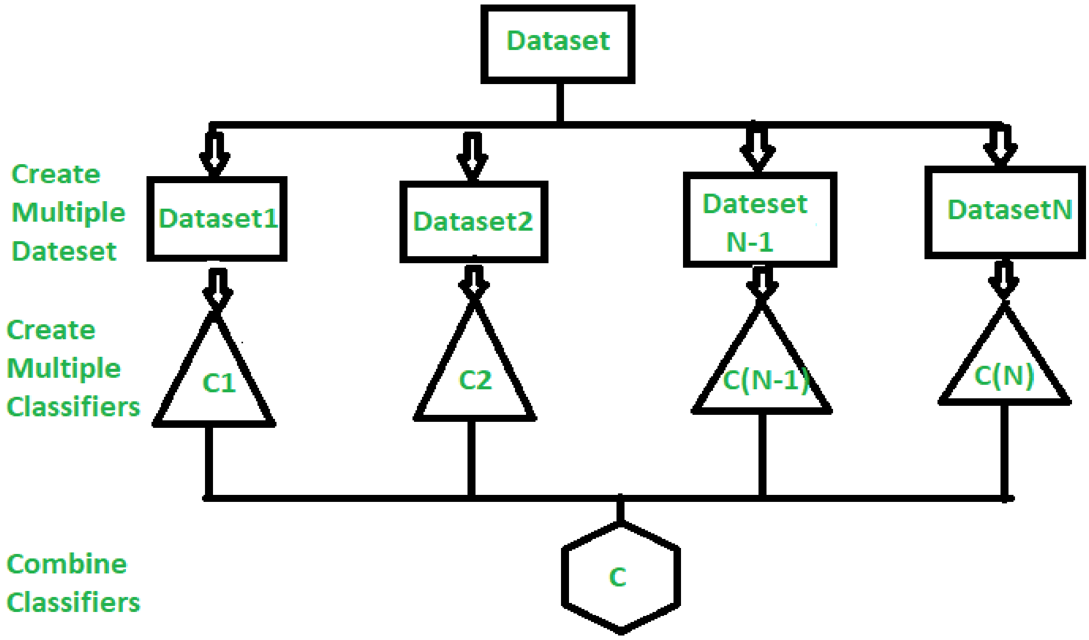


Figure 2.7: A Model of Ensemble Learning.

The general term of multiple classifier systems covers also hybridization of hypotheses that are not induced by the same base learner.

1. **Boosting** An ensemble model uses decision trees as weak learners and builds the model in a stage-wise manner by optimizing a loss function. Boosting emerged as a way of combining many weak classifiers to produce a powerful committee. It is an iterative process that gives more and more importance to bad classification. Simple strategy results in dramatic improvements in classification performance. To do so, a boosting auto regression procedure is applied at each horizon on the residuals from the recursive linear forecasts using a so-called weak learner, which is a learner with large bias relative to variance.

2. **Bagging** Bootstrap aggregating, also called bagging used in statistical classification and regression, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms. The algorithm also reduces variance and helps to prevent over fitting. Although it is generally applied to decision tree methods, it can be used with any type of learning method. Bagging is a special case of the model averaging approach. Bagging predictors is generally used to generate multiple versions of a predictor and using them to get an aggregated predictor. The aggregation averages all the versions when predicting a numerical result and does a plurality vote to predict a class. The multiple versions are formed by making bootstrap replicates of the learning set and using them as new learning sets.

3. **Random Subspace** The machine learning tool that is used in the proposed methodology is based on Random Forests, which consists of a collection, or ensemble of a multitude of decision trees, each one built from a sample drawn with replacement (a bootstrap sample) from a training set, is the group of outputs. Furthermore, only a random subset of variables is used when splitting a node during the construction of a tree. As a consequence, the final nodes (or leaf's), may contain one or several observations. For regression problems, each tree can produce a response when presented with a set of predictors, being the conditional mean of the observations present on the resulting leaf. The conditional mean is typically approximated by a weighted mean. As a result of the random construction of the trees, the bias of the forest generally slightly increases with respect

to the bias of a single non-random tree but, due to the averaging its variance decreases, frequently more than compensating for the increase in bias, hence yielding an overall better model. Finally, the responses of all trees are also averaged to obtain a single response variable for the model, and here as well a weighted mean is used. Substantial improvements in classification accuracy were obtained from growing an ensemble of trees and letting them vote for the most popular class. To grow these ensembles, often random vectors are generated which govern the growth of each tree in the ensemble. One of the first examples used is bagging, in which to grow each tree a random selection (without replacement) is made from the examples contained in the training set.

4. **Predictors ensemble** Current practice suggests that forecasts should be composed either by a number of sample say conventional forecasts- or produce a simple forecast from other simple forecasts (not only point forecasts, but also probabilistic). This leads to gains in performance, relative to the contributing forecasts. In the case of statistical models, realizations coming from the same technology (for example the same neural network architecture) trained multiple times, or using different samples of the data-set, or different technologies. Once first stage forecasts are available, different combination approaches are possible. The simplest approach is averaging of results given by different methods. A more general approach assigns a weight to each of the contributing methods, for each time horizon, depending on different criteria and with different weighting policies. Simple forecasts can be seen as different perceptions of the same true state. In this way, approaches of imperfect sensor data fusion should also be valid to perform a combination of forecasts. Ensemble-based artificial neural networks and other machine learning technics have been used in a number of studies in global radiation modeling and provided better performance and generalization capability compared to conventional regression models.

2.2.6.5 Evaluation of model accuracy

Evaluation, generally, measures how good something is. This evaluation is used at various steps of the model development as for example during the evaluation of the forecasting model itself (during the training of a statistical model for example), for judging the improvement of the model after some modifications and for comparing various models. As previously mentioned, this performance comparison is not easy for various reasons such as different forecasted time horizons, various time scale of the predicted data and variability of the meteorological conditions from one site to another one. It works by comparing the forecasted outputs (or predicted time series) with observed data y (or observed or measured time series) which are also measured data themselves linked to an error (or precision) of a measure.

2.2.6.6 Jupyter Notebook

Solar energy is one of the leading renewable energy sources in the world and it continues to grow. However, it depends on sunlight which is an intermittent natural resource. This makes power output predictability critical for the integration of solar photovoltaics into traditional electrical grid systems. While irradiance is a strong predictor of solar power output, collecting this information about a location is often tedious and its estimation may have significant errors. Hence, the ability to predict power output without irradiance data needs to be further explored to save time, effort, and cost with no significant loss of accuracy.

Anaconda Navigator is a desktop graphical user interface (GUI) included in anaconda distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Linear Regression is used in this case, as it is known as the most basic and widely used technique for regression [4]. It models the relationship between the input and output variables using linear predictor functions whose unknown model parameters are estimated from the data using a least square approach. The parameter values can be estimated either by solving a set of linear equations or using an iterative method such as gradient descent [5].

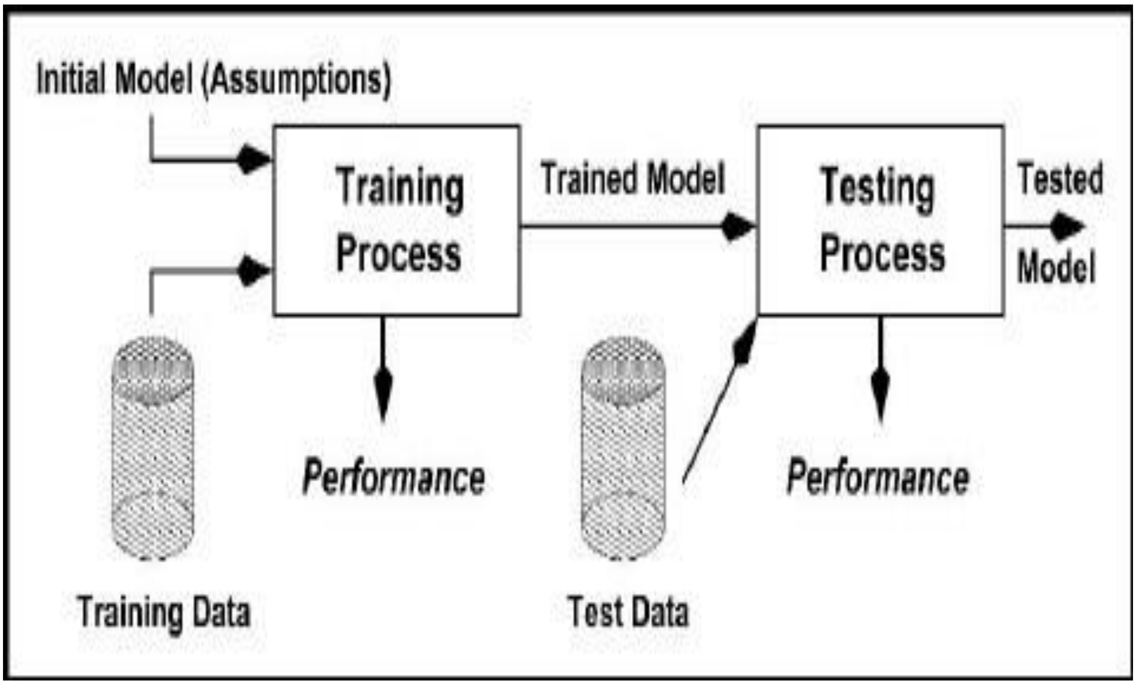


Figure 2.8: A Model of Supervised Learning.

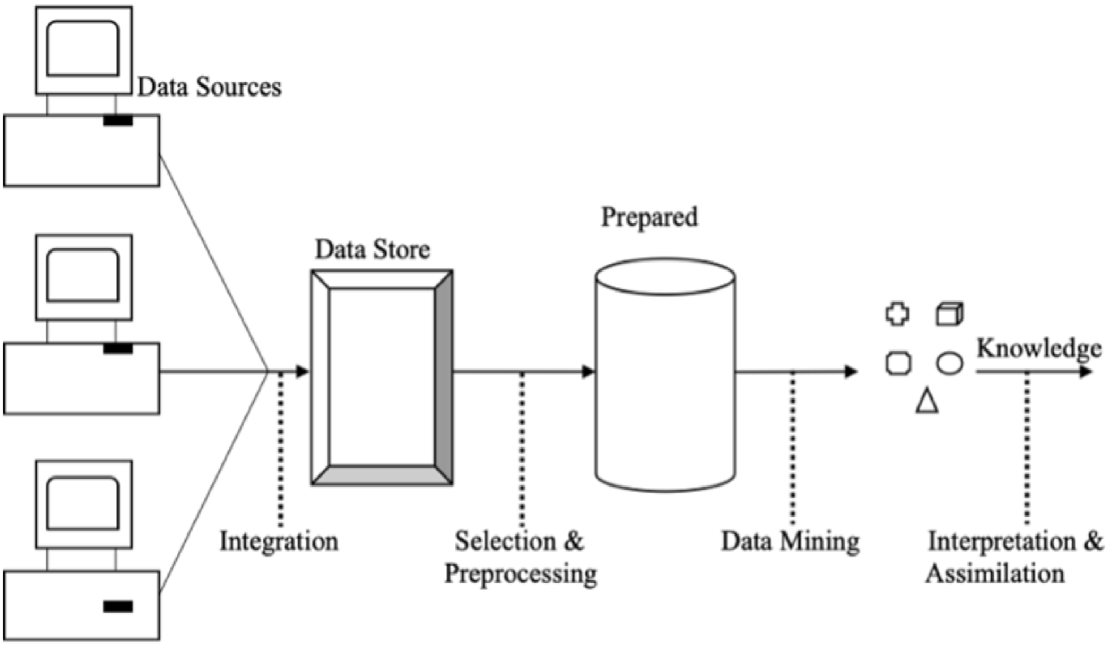


Figure 2.9: A Complete Data Processing Unit.

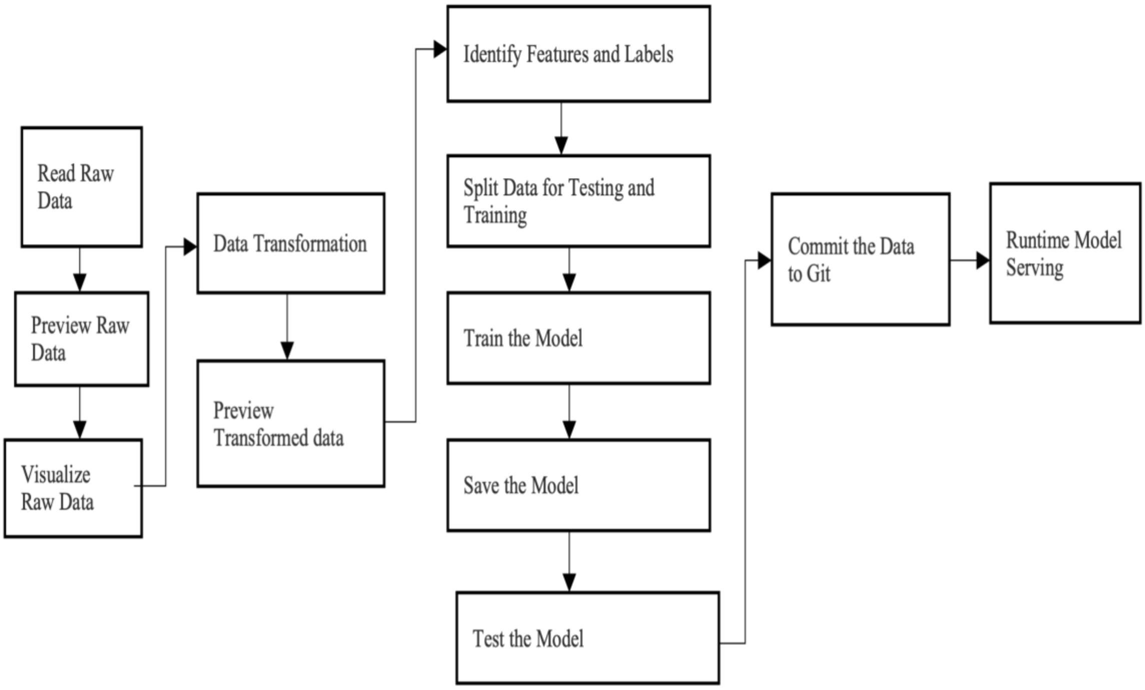


Figure 2.10: Shows the Notebook workflow for the Solar outcome.

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS.

For this research the Jupyter notebook is used for the data learning and predictions to forecast data to be used. Data comes in, possibly from many sources. It is integrated and placed in some common data store. Part of it is then taken and pre-processed into a standard format. This prepared data is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of patterns. These are then interpreted to give the foretasted value which is also term useful knowledge as seen in Fig.2 and Fig.3 which explains the process of data transformation from its raw state and then previewed, identified, split, trained, saved and tested. The data is further committed to Git and finally to runtime model serving.

Jupyter is an electronic lab notebook to document procedures, data, calculations, and findings. It provides an interactive computational environment for developing data science applications. Jupyter notebooks combine software code, computational output, explanatory text, and rich content in a single document. It allows in-browser editing and execution of code and display computation results which is saved in .ipynb extension.

2.3 Summary

This chapter summaries the methods to predict solar radiation or solar production One of the most important challenge for the near future global energy supply will be the large integration of renewable energy sources .The various methods of machine learning briefly explained, while pointing out the main method carried out for data analyses in this research.

Bibliography

[1] Blake, C. L., Merz, C. J. (1998). UCI repository of machine learning databases. Irvine: University of California, Department of In- formation and Computer Science. http://www.ics.uci.edu/ mlearn/ MLRepository.html.

[2] P. A. G. M. Amarasinghe and S. K. Abeygunawardane, ”Application of Machine Learning Algorithms for Solar Power Forecasting in Sri Lanka” (2nd International Conference On Electrical Engineering (EECon), Colombo, Sri Lanka, 87 2018).”. for the Energy Information Administration. Retrieved April 13, 2015.

[3] hangsong Chen, Shanxu Duan, Tao Cai, and Bangyin Liu. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Solar Energy, 85(11):2856 – 2870, 2011.

[4] .Z.Hassan,M.E.K.Ali,A.B.M.S.AliandJ. Kumar, ”Forecasting Day-Ahead Solar Radiation Using Machine Learning Approach” (4th Asia- Pacific World Congress on Computer Science and Engineering (APWC on CSE), Mana Island, Fiji, 252 2017).

[5] A. Khan, R. Bhatnagar, V. Masrani and V. B. Lobo, ”A Comparative Study on Solar Power Forecasting using Ensemble Learning,” (4th International Conference on Trends in Electronics and Informatics (ICOEI), 224 2020).

[6] Jawaid F, NazirJunejo K. Predicting daily mean solar power using machine learning regression techniques. (Sixth International Conference on Innovative Computing Technology (INTECH) 355 2016).

[7] Batcha RR, Geetha MK. A survey on IOT based on renewable energy for efficient energy conservation using machine learning approaches. (3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE) 123 2020). *Last access on Mar* 02, 2017.

[8] RENA. Renewable power generation costs in 2017. Technical report, Interna- tional Renewable Energy Agency, Abu Dhabi, January 2018.

[9] ose R. Andrade and Ricardo J. Bessa. Improving renewable energy forecasting with a grid of numerical weather predictions. IEEE Transactions on Sustainable Energy, 8(4):1571–1580, October 2017.

[10] ich H. Inman, Hugo T.C. Pedro, and Carlos F.M. Coimbra. Solar forecasting methods for renewable energy integration. Progress in Energy and Combustion Science, 39(6):535 – 576, 2013.

[11] . Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. Solar Energy, 136:78–111, October 2016.

[12] Kostylev, A Pavlovski, et al. Solar power forecasting performance–towards industry standards. In 1st international workshop on the integration of solar power into power systems, Aarhus, Denmark, 2011.

[13] Ao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. International Journal of Forecasting, 32(3):896 – 913, 2016.

[14] Ordon Reikard. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Solar Energy, 83(3):342 – 349, 2009.

[15] Eder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online short-term solar power forecasting. Solar Energy, 83(10):1772 – 1783, 2009.

[16] Ugo T.C. Pedro and Carlos F.M. Coimbra. Assessment of forecasting tech- niques for solar power production with no exogenous inputs. Solar Energy, 86(7):2017 – 2028, 2012.

[17] Federica Davò, Stefano Alessandrini, Simone Sperati, Luca Delle Monache, Da- vide Airoldi, and Maria T. Vespucci. Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. Solar Energy, 134:327 – 338, 2016.

Chapter 3

Data Analysis: Predicting Solar Energy Generation Using the Jupyter Notebook

3.1 Introduction

The sun delivers solar energy in the form of solar radiation, which is produced by the photovoltaic effect. Sunlight intensity is the most important factor influencing the output of photovoltaic (PV) solar panels. A PV system output can be affected by a variety of different environmental variables among others. Identifying which parts of PV are valuable and which aspects are not also essential so that a suitable feature subset may be selected as an input to the model. We propose a hybrid method for feature variable selection that comprises two basic processes, namely, the filter stage and the wrapper stage, as seen below.

3.1.1 Importance of Feature Selection in Machine Learning

Machine learning works on a simple rule if you put garbage in, you will only get garbage to come out. By garbage here, I mean noise in data. This becomes even more important when the number of features are very large. The need not to use every feature at your disposal for creating an algorithm. You can assist your algorithm by feeding in only those features that are really important.

Feature selection can be very useful in industrial applications. You not only reduce the training time and the evaluation time, but you also have less things to worry about! Here are top reasons to use feature selection are:

1. **It enables the machine learning algorithm to train faster**
2. **It reduces the complexity of a model and makes it easier to interpret.**
3. **It improves the accuracy of a model if the right subset is chosen.**
4. **It reduces over-fitting.**

Prior to begin the learning process, the filter technique analyzes features based on the inherent attributes of each one of the features. Filter criteria are used to select a subset of features from a dataset based on their relevance. Because of the characteristics of PV data, the Pearson correlation coefficient (PCC) is employed to assess the relationship between input factors and the target variable. A PCC is a statistical metric that is used to determine the linear correlation between two variables, X and Y, in a dataset. Data from time series analysis captures the degree to which a target variable Y correlates with an input variable X over the course of an observation period. When calculating the correlation between two variables, the time series data at the points t and t1 of the variable are not used. In our example, the

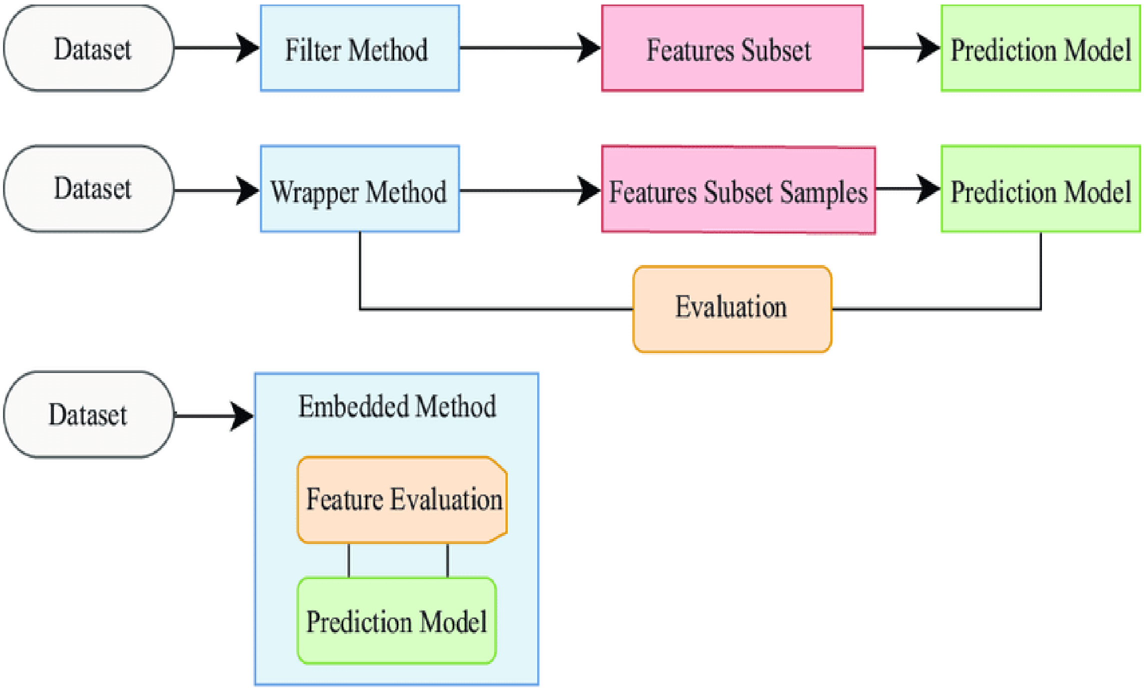


Figure 3.1: Prediction Interface of Solar Outcome Using Jupyter Notebook.

meteorological factors that have an impact on PVPG are represented by the letters Y and X, respectively. As a result, the PCC may be expressed mathematically as,

$$\rho(X, Y) = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}} \tag{3.1}$$

where PCC is the value that lies between +1 and -1. PCC is one of the most commonly used criteria for describing the relationship between variables in practice, and it is also one of the most widely studied. The wrapper approach is utilized to analyze each subset that has been selected. The learning algorithm is integrated into the feature selection process, which in turn makes use of the error of a given model to determine which feature subset is most important to the user. It was decided to employ the traditional LSTM model for the analysis of feature subsets in this study because of its capacity to address time series forecasting issues. As a result, the optimal subset of training characteristics may be determined from among all of the subsets that have been investigated.

Through the use of a hybrid approach, the proposed feature selection attempts to integrate the best aspects of wrapper and filter methods into a single method. After examining the correlations between variables using filter criteria, appropriate thresholds are determined in order to reduce the number of feature variables that can be evaluated. The filter technique, in contrast to other learning algorithms, is univariate in nature. This results in it being significantly more efficient and faster to compute than the wrapper technique, and it is capable of dealing with massive datasets with ease. No consideration is currently given to how features interact with one another or with the learning algorithms, which is a problem. In this case, the wrapper technique is required because coupled features in the single feature evaluation. In order to effectively use an individual wrapper strategy, a significant amount of computer power is required. This is owing to the learning methods used and the enormous number of feature subsets that must be analyzed. Despite this, when the correlation results of the filter approach are utilized as a guide, fewer feature subsets are generated and analyzed than would otherwise be the case. As a result, the hybrid method that has been proposed has the potential to improve the effectiveness of the feature selection.

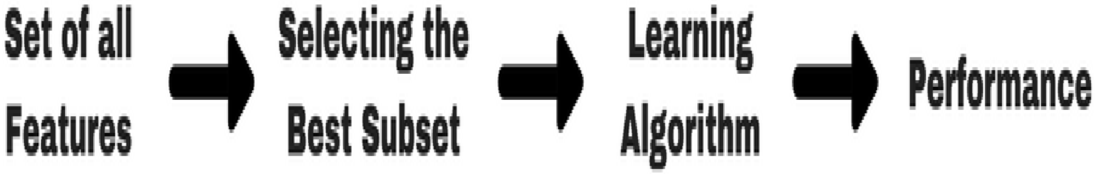


Figure 3.2: Filter Method

3.1.2 Filter Method

Filter methods are generally used as a pre-processing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. The correlation is a subjective term here. For basic guidance, you can refer to the following table for defining correlation coefficients.

1. **Pearson's Correlation** It is used as a measure for quantifying linear dependence between two continuous variables X and Y. Its value varies from -1 to +1. Pearson's correlation is given as:

$$\rho X, Y = cov(X, Y)/\delta_x \delta_y \tag{3.2}$$

2. **LDA** Linear discriminant analysis is used to find a linear combination of features that characterizes or separates two or more classes (or levels) of a categorical variable.

3. **ANOVA:** ANOVA stands for Analysis of variance. It is similar to LDA except for the fact that it is operated using one or more categorical independent features and one continuous dependent feature. It provides a statistical test of whether the means of several groups are equal or not.

4. **Chi-Square** It is a is a statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution. One thing that should be kept in mind is that filter methods do not remove multicollinearity.

3.1.3 Wrapper Method

In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model; we decide to add or remove features from your subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive. Some common examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

1. **Forward Selection** Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

2. **Backward Elimination** In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

3. Feature elimination Recursive Feature elimination: It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left

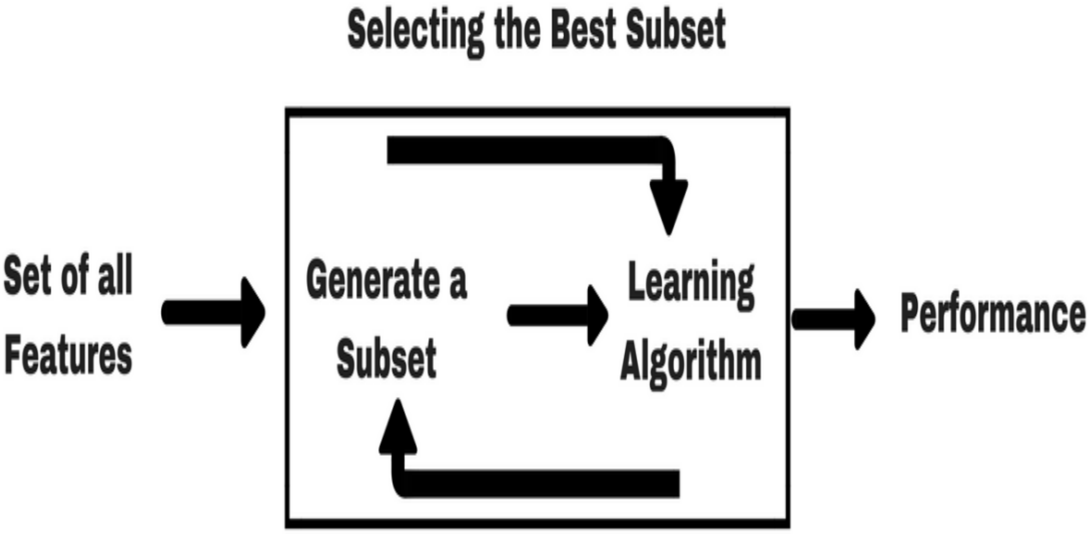


Figure 3.3: Wrapper Method

features until all the features are exhausted. It then ranks the features based on the order of their elimination.

3.1.4 Difference between Filter and Wrapper methods

The main differences between the filter and wrapper methods for feature selection are:

- Filter methods measure the relevance of features by their correlation with dependent variable while wrapper methods measure the usefulness of a subset of feature by actually training a model on it.
- Filter methods are much faster compared to wrapper methods as they do not involve training the models. On the other hand, wrapper methods are computationally very expensive as well.
- Filter methods use statistical methods for evaluation of a subset of features while wrapper methods use cross validation.
- Filter methods might fail to find the best subset of features in many occasions but wrapper methods can always provide the best subset of features.
- Using the subset of features from the wrapper methods make the model more prone to over fitting as compared to using subset of features from the filter methods.

The current dataset is based on monthly weather parameter values over a long period of 10 years. Various weather characteristics were gathered in order to investigate the relationship between mean solar irradiance and meteorological data in order to accurately estimate mean solar irradiance. The average daily values of air temperature, humidity, wind speed, wind direction, visibility, average pressure, average wind speed, and electricity generated are among the data collected. The direction of the wind, on the other hand, indicates how high the sun is. It's also expressed in degrees. Machine Learning (ML) Models are used for forecasting the solar power generation weather analysis. The Regression techniques here proposed are Support Vector Machine, Random Forest, Linear Regression are various ML Models used. For this research the Japan Meteorological Agency data on weather was used for analysis. The supervised learning approach is implemented here, as it deals with labels and features. Support vector regression (SVR) using a radial basis function as the kernel and random forest (RF) approaches are used to create the models. Because of the non-linearity of the dataset, we used the models indicated above instead of linear models. The most basic and widely used regression method is linear regression. It uses linear predictor functions to represent the relationship between the input and output variables, and a least squares approach is used to estimate the unknown model parameters from the data. A set of linear equations or an iterative method like gradient descent can be used to estimate parameter values. Nonlinear relationships can be mapped using these methods. In data science challenges of various kinds, methods including decision trees, RF, and gradient boosting are commonly utilized. The RF method is a tree-based machine learning approach that can be used for regression and classification. It also performs dimensional reduction, controls missing and outlier values, and performs a variety of additional data exploration activities. The bagging approach is used to train RFs. This method allows for the usage of numerous instances for the training stage because the dataset is sampled with a replacement. Linear regression is a method for demonstrating the link between a dependent variable and one or more independent variables by using the best-fit linear curve. It is concerned with determining the best-fit line with the data by attaining a perfect slope and intercept value. The best model for forecasting solar power system output based on numerous weather parameters was then created. The models that gave the greatest results on the dataset were support vector regression, random forests, and linear regression, and these models were then utilized to anticipate PV system performance.

3.2 Selection Process for Multiple Regression

The basis of a multiple linear regression is to assess whether one continuous dependent variable can be predicted from a set of independent or predictor variables, in simple terms, how much variance a continuous dependent variable is explained by a set of predictors. Certain regression selection approaches are helpful in testing predictors, thereby increasing the efficiency of analysis.

3.2.1 Entry Method

The standard method of entry is simultaneous; all independent variables are entered into the equation at the same time. This is an appropriate analysis when dealing with a small set of predictors and when it’s hard to determine which independent variables will create the best prediction equation. Each predictor is assessed as though it were entered after all the other independent variables were entered and assessed by what it offers to the prediction of the dependent variable that is different from the predictions offered by the other variables entered into the model.

3.2.2 P-values and statistical significance

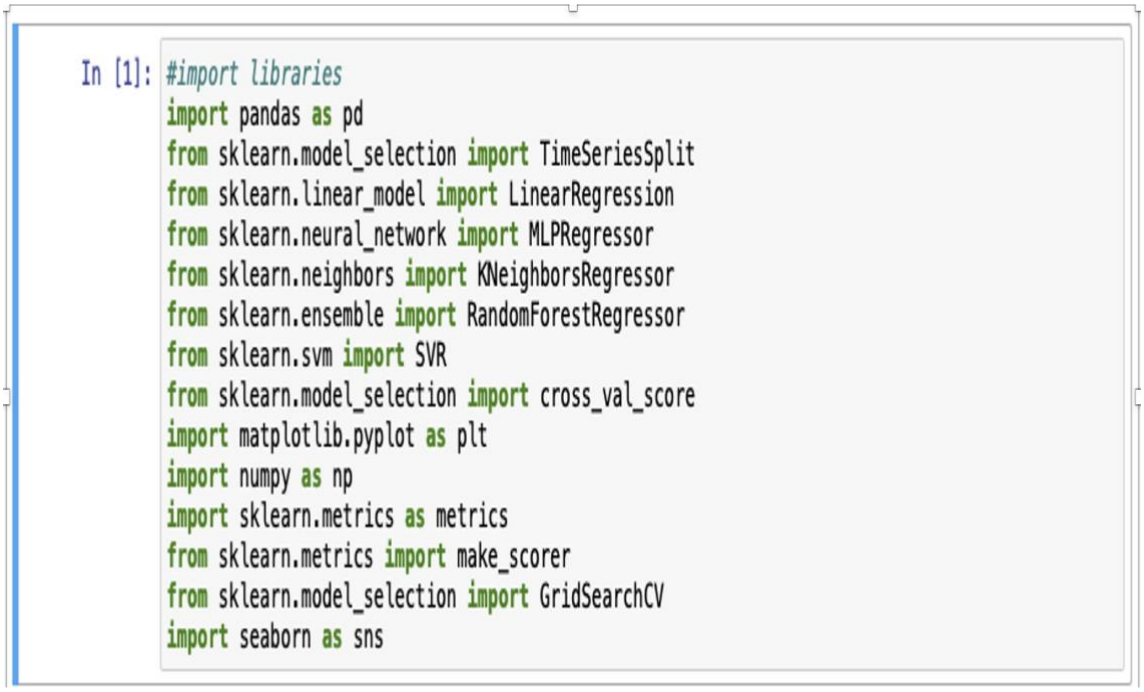
P-values are most often used by researchers to say whether a certain pattern they have measured is statistically significant. Statistical significance is another way of saying that the p-value of a statistical test is small enough to reject the null hypothesis of the test. How small is small enough? The most common threshold is p ¡ 0.05; that is, when you would expect to find a test statistic as extreme as the one calculated by your test only 0.05 of the time. But the threshold depends on your field of study. Some fields prefer thresholds of 0.01, or even 0.001. The threshold value for determining statistical significance is also known as the alpha value.

3.2.3 Exploring the Raw data

Exploring the Raw data: A code cell was used to import the required python libraries and the raw files converted from .csv to a Data frame with a time series as seen in Fig. 4 Libraries used for the analysis of the raw data includes:

Table 3.1: The Feature Description of Dataset.

Source-Dependency	Feature	Description
Averagely dependent	Total Precipitation	The sum of rainfall and the assumed water equivalent of snowfall for a given year.
Averagely dependent	Mean relative humidity	The ratio of the amount of water vapor actually presents in the air to the greatest amount possible at the same temperature.
Averagely dependent	Mean air temperature	The average temperature of the air as indicated by a properly exposed thermometer during a given time.
Averagely dependent	Mean wind speed	Time-averaged wind speed, averaged over a specified time interval
Highly dependent	Total sunshine duration	The length of time that the ground surface is irradiated by direct solar radiation.
Highly dependent	Percentage possible sunshine	The total amount of possible sunshine this represents the average annual amount. This measurement is the total time that sunshine reaches the earth expressed as the percent of the possible maximum amount of sunshine from sunrise to sunset (with clear sky conditions.)
Highly dependent	Solar radiation	The energy radiated from the sun in the form of electromagnetic waves, including visible and ultraviolet light and infrared.



```
In [1]: #import libraries
        import pandas as pd
        from sklearn.model_selection import TimeSeriesSplit
        from sklearn.linear_model import LinearRegression
        from sklearn.neural_network import MLPRegressor
        from sklearn.neighbors import KNeighborsRegressor
        from sklearn.ensemble import RandomForestRegressor
        from sklearn.svm import SVR
        from sklearn.model_selection import cross_val_score
        import matplotlib.pyplot as plt
        import numpy as np
        import sklearn.metrics as metrics
        from sklearn.metrics import make_scorer
        from sklearn.model_selection import GridSearchCV
        import seaborn as sns
```

Figure 3.4: Prediction Interface of Solar Outcome Using Jupyter.

1. **Numpy** is for manipulating and creating vectors and matrices.

2. **Pandas** For analyzing, wrangling, and munging data.

3. **Matplotlib** Is for data visualization.

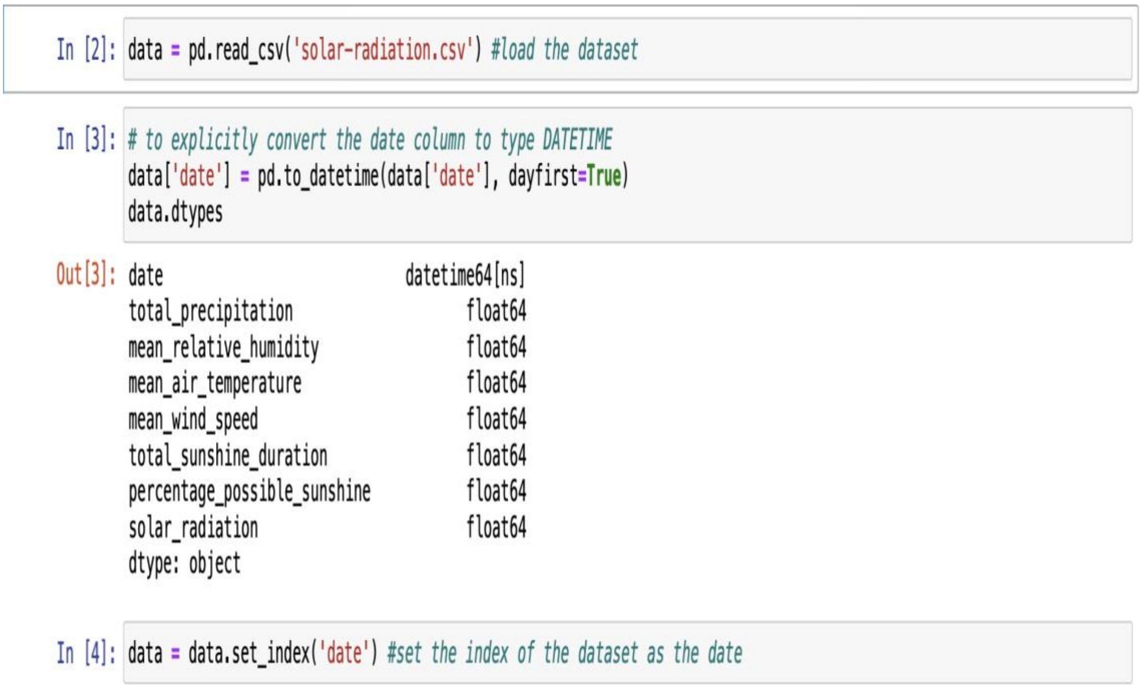
4. **Sklern** For supervised and unsupervised learning. This library provides various tools for model fitting, data preprocessing, model selection, and model evaluation. It has built-in machine learning algorithms and models called estimators.

5. **Seaborn** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

For the forecast and prediction of the solar radiation the Japan meteorological Agency (JMA) data set was used as a sample study for over 10 years. Different weather features were used for the analysis as seen in table 1. Importation of different libraries to help with the analysis as the functions of imported library has been stated above. The raw data is converted to csv format and extracting of independent and dependent variable is done, followed by the splitting of the dataset into training and test set, then proceed to Fitting the Multi-Linear-Regression (MLR) model to the training set and finally Predicting the Test set result and checking the score.

Why Backward Elimination? Backward elimination is a feature selection technique while building a machine learning model. It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output. There are various ways to build a model in Machine Learning, which are: All-in, Backward Elimination, Forward Selection, Bidirectional Elimination, Score Comparison, Above are possible methods for building the model in Machine learning, but we only used the Backward Elimination process as it is the fastest method. Backward Elimination is Important for Multiple linear Regression Model. Unnecessary features increase the complexity of the model. Hence it is good to have only the most significant features and keep our model simple to get the better result. So, in order to optimize the performance of the model, we used the Backward Elimination method. This process is used to optimize the performance of the multi linear regression (MLR) model as it only includes the most effective feature and removes the least effective feature.

Below are some main steps which are used to apply backward elimination process:



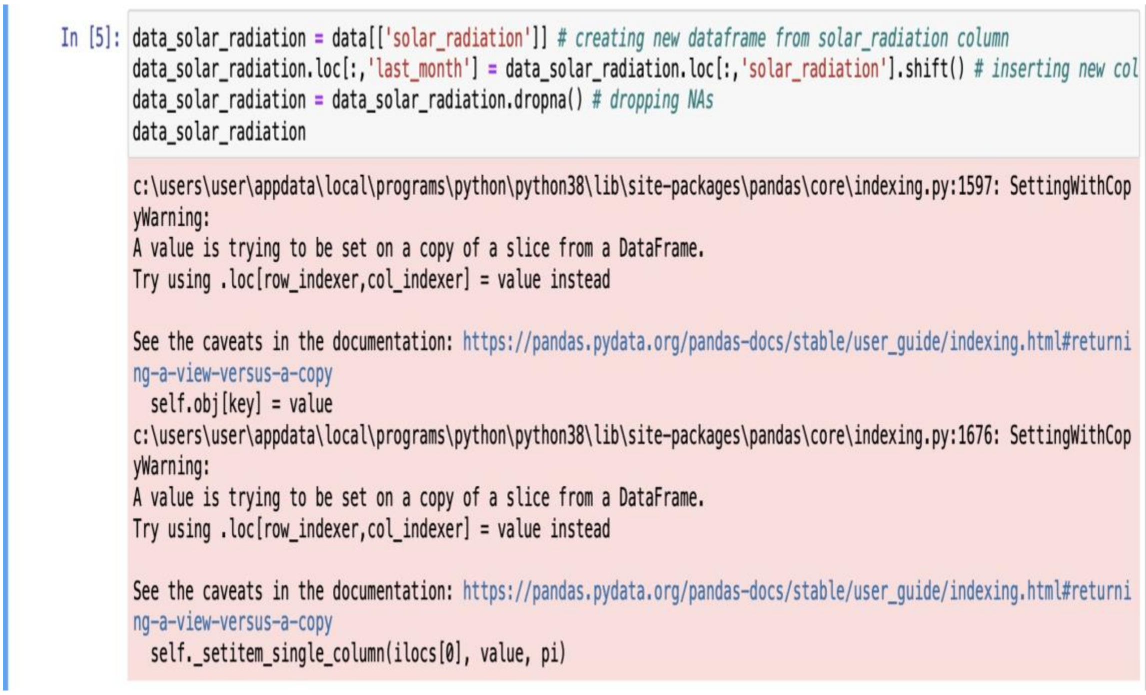
```
In [2]: data = pd.read_csv('solar-radiation.csv') #load the dataset

In [3]: # to explicitly convert the date column to type DATETIME
        data['date'] = pd.to_datetime(data['date'], dayfirst=True)
        data.dtypes

Out[3]: date                            datetime64[ns]
        total_precipitation                    float64
        mean_relative_humidity                 float64
        mean_air_temperature                   float64
        mean_wind_speed                        float64
        total_sunshine_duration                float64
        percentage_possible_sunshine           float64
        solar_radiation                        float64
        dtype: object

In [4]: data = data.set_index('date') #set the index of the dataset as the date
```

Figure 3.5: Prediction Interface of Solar Outcome Using Jupyter.



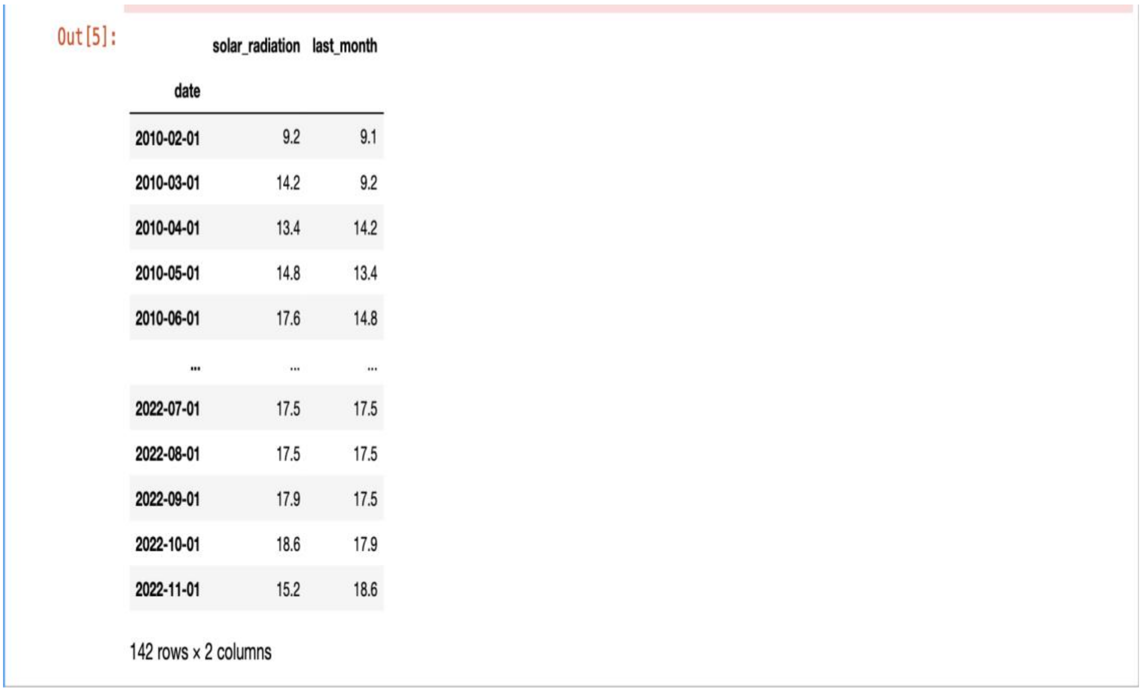
```
In [5]: data_solar_radiation = data[['solar_radiation']] # creating new dataframe from solar_radiation column
        data_solar_radiation.loc[:,'last_month'] = data_solar_radiation.loc[:,'solar_radiation'].shift() # inserting new col
        data_solar_radiation = data_solar_radiation.dropna() # dropping NAs
        data_solar_radiation

        c:\users\user\appdata\local\programs\python\python38\lib\site-packages\pandas\core\indexing.py:1597: SettingWithCop
        yWarning:
        A value is trying to be set on a copy of a slice from a DataFrame.
        Try using .loc[row_indexer,col_indexer] = value instead

        See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returni
        ng-a-view-versus-a-copy
          self.obj[key] = value
        c:\users\user\appdata\local\programs\python\python38\lib\site-packages\pandas\core\indexing.py:1676: SettingWithCop
        yWarning:
        A value is trying to be set on a copy of a slice from a DataFrame.
        Try using .loc[row_indexer,col_indexer] = value instead

        See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returni
        ng-a-view-versus-a-copy
          self._setitem_single_column(ilocs[0], value, pi)
```

Figure 3.6: Prediction Interface of Solar Outcome Using Jupyter.

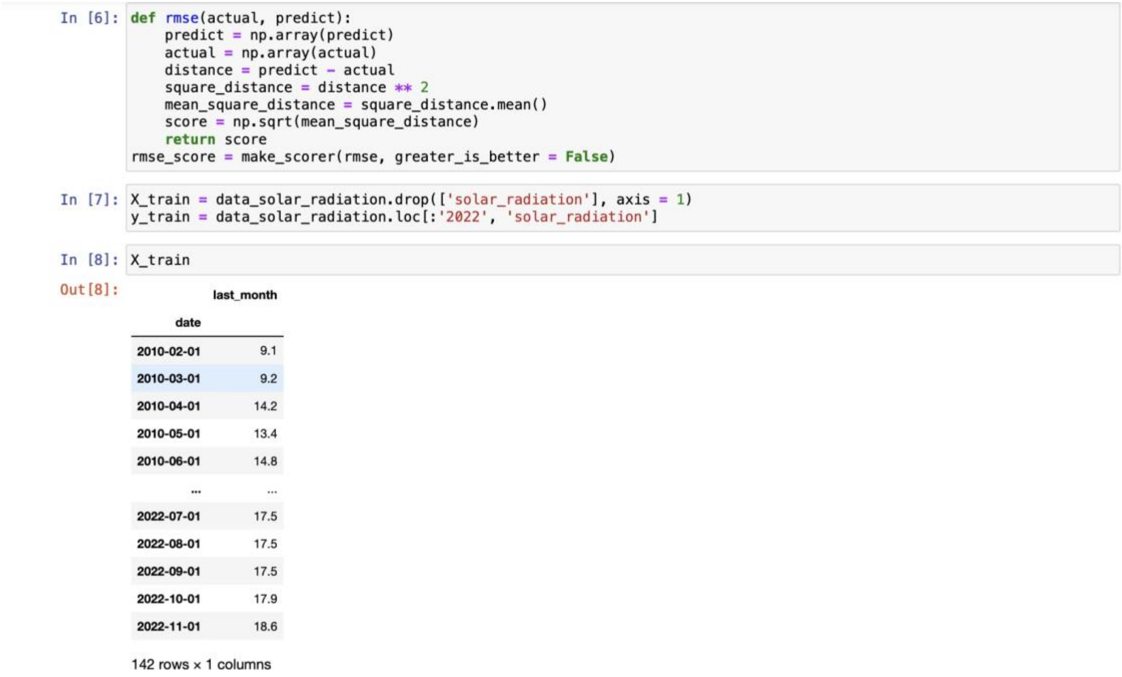


Out[5]:

date	solar_radiation	last_month
2010-02-01	9.2	9.1
2010-03-01	14.2	9.2
2010-04-01	13.4	14.2
2010-05-01	14.8	13.4
2010-06-01	17.6	14.8
...	...	...
2022-07-01	17.5	17.5
2022-08-01	17.5	17.5
2022-09-01	17.9	17.5
2022-10-01	18.6	17.9
2022-11-01	15.2	18.6

142 rows x 2 columns

Figure 3.7: Prediction Interface of Solar Outcome Using Jupyter.



In [6]:

```
def rmse(actual, predict):
    predict = np.array(predict)
    actual = np.array(actual)
    distance = predict - actual
    square_distance = distance ** 2
    mean_square_distance = square_distance.mean()
    score = np.sqrt(mean_square_distance)
    return score
rmse_score = make_scorer(rmse, greater_is_better = False)
```

In [7]:

```
X_train = data_solar_radiation.drop(['solar_radiation'], axis = 1)
y_train = data_solar_radiation.loc[:'2022', 'solar_radiation']
```

In [8]:

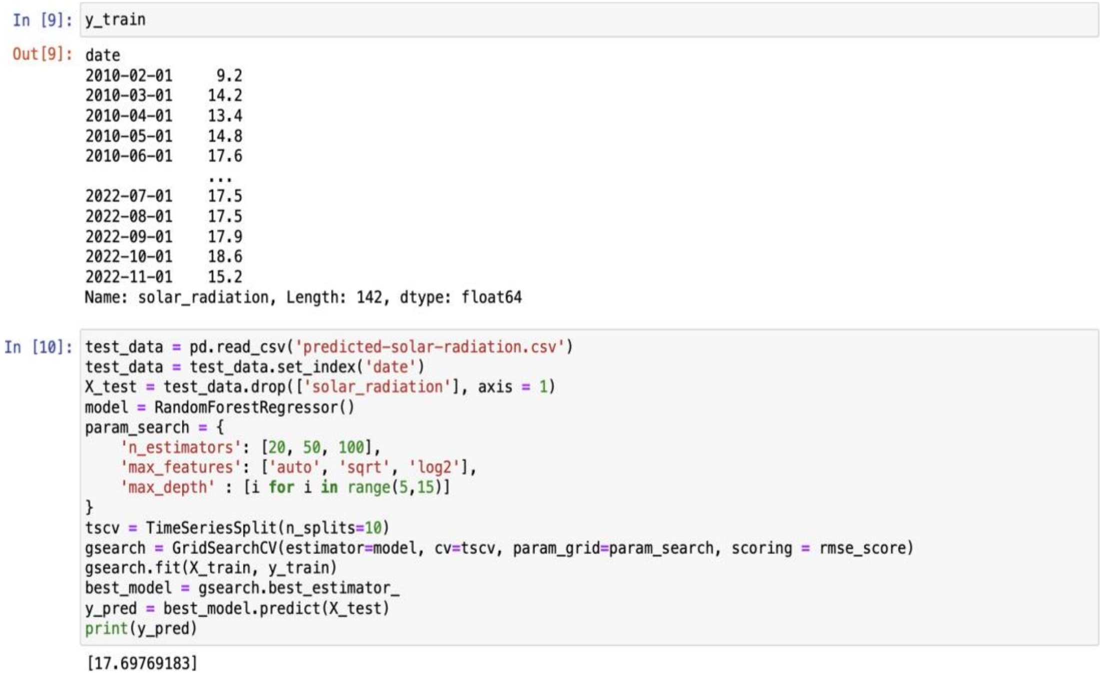
```
X_train
```

Out[8]:

date	last_month
2010-02-01	9.1
2010-03-01	9.2
2010-04-01	14.2
2010-05-01	13.4
2010-06-01	14.8
...	...
2022-07-01	17.5
2022-08-01	17.5
2022-09-01	17.5
2022-10-01	17.9
2022-11-01	18.6

142 rows × 1 columns

Figure 3.8: Prediction Interface of Solar Outcome Using Jupyter.



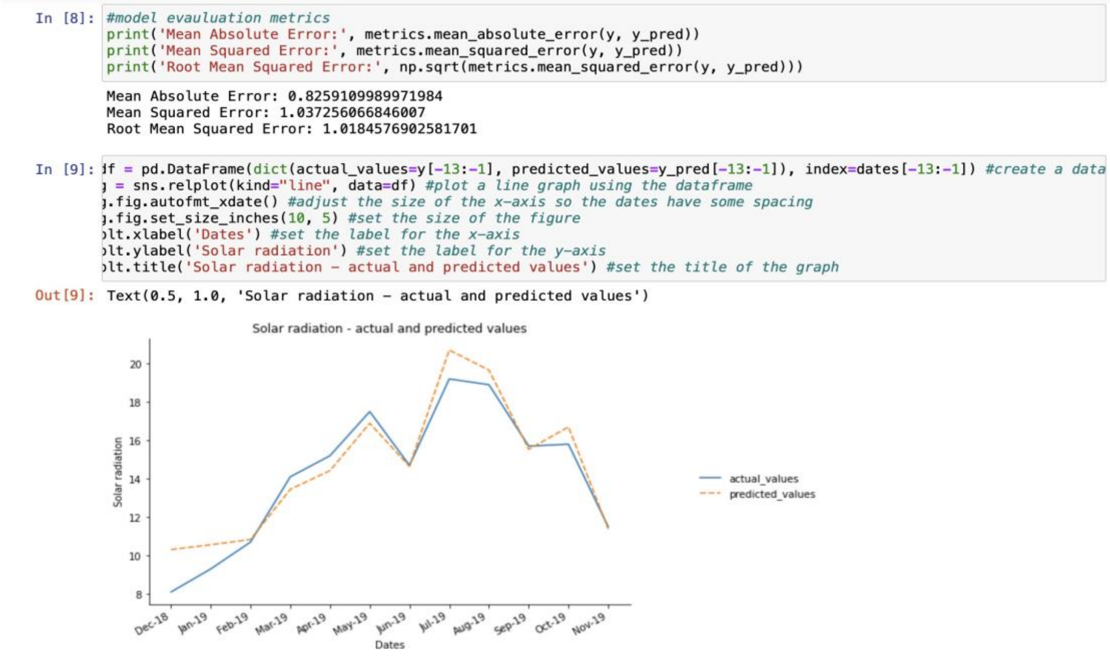
```
In [9]: y_train

Out[9]: date
        2010-02-01     9.2
        2010-03-01    14.2
        2010-04-01    13.4
        2010-05-01    14.8
        2010-06-01    17.6
                      ...
        2022-07-01    17.5
        2022-08-01    17.5
        2022-09-01    17.9
        2022-10-01    18.6
        2022-11-01    15.2
        Name: solar_radiation, Length: 142, dtype: float64

In [10]: test_data = pd.read_csv('predicted-solar-radiation.csv')
         test_data = test_data.set_index('date')
         X_test = test_data.drop(['solar_radiation'], axis = 1)
         model = RandomForestRegressor()
         param_search = {
             'n_estimators': [20, 50, 100],
             'max_features': ['auto', 'sqrt', 'log2'],
             'max_depth' : [i for i in range(5,15)]
         }
         tscv = TimeSeriesSplit(n_splits=10)
         gsearch = GridSearchCV(estimator=model, cv=tscv, param_grid=param_search, scoring = rmse_score)
         gsearch.fit(X_train, y_train)
         best_model = gsearch.best_estimator_
         y_pred = best_model.predict(X_test)
         print(y_pred)

         [17.69769183]
```

Figure 3.9: Prediction Interface of Solar Outcome Using Jupyter.



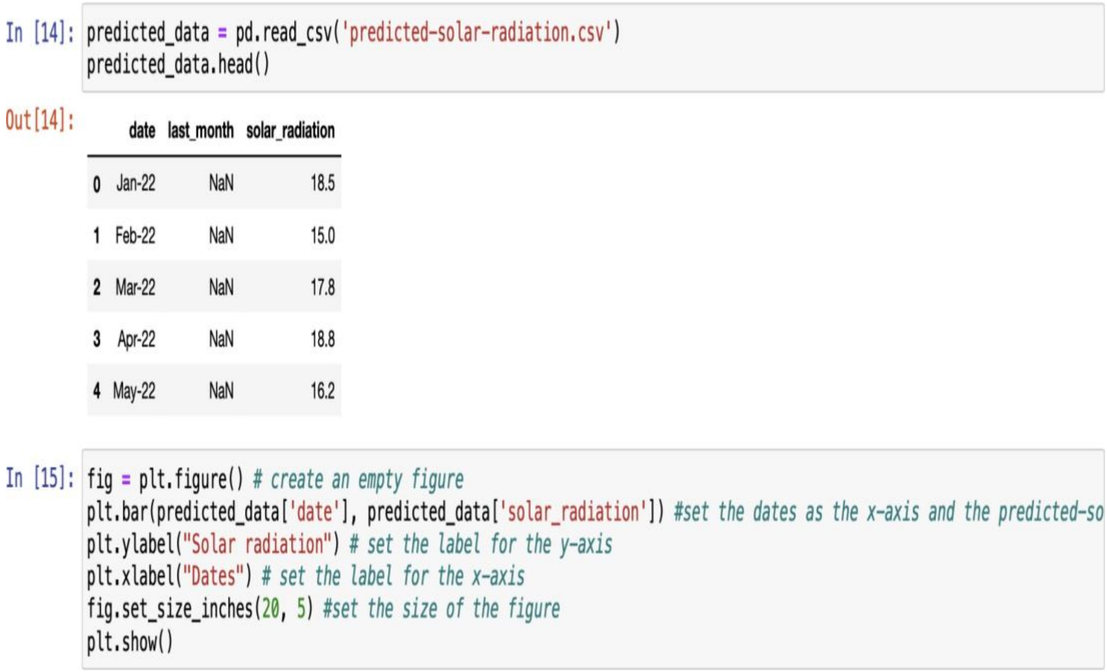
```
In [8]: #model evauluation metrics
        print('Mean Absolute Error:', metrics.mean_absolute_error(y, y_pred))
        print('Mean Squared Error:', metrics.mean_squared_error(y, y_pred))
        print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y, y_pred)))

        Mean Absolute Error: 0.8259109989971984
        Mean Squared Error: 1.037256066846007
        Root Mean Squared Error: 1.0184576902581701

In [9]: df = pd.DataFrame(dict(actual_values=y[-13:-1], predicted_values=y_pred[-13:-1]), index=dates[-13:-1]) #create a data
        g = sns.relplot(kind="line", data=df) #plot a line graph using the dataframe
        g.fig.autofmt_xdate() #adjust the size of the x-axis so the dates have some spacing
        g.fig.set_size_inches(10, 5) #set the size of the figure
        plt.xlabel('Dates') #set the label for the x-axis
        plt.ylabel('Solar radiation') #set the label for the y-axis
        plt.title('Solar radiation - actual and predicted values') #set the title of the graph

Out[9]: Text(0.5, 1.0, 'Solar radiation - actual and predicted values')
```

Figure 3.10: Prediction Interface of Solar Outcome Using Jupyter.



```
In [14]: predicted_data = pd.read_csv('predicted-solar-radiation.csv')
         predicted_data.head()
```

Out[14]:

	date	last_month	solar_radiation
0	Jan-22	NaN	18.5
1	Feb-22	NaN	15.0
2	Mar-22	NaN	17.8
3	Apr-22	NaN	18.8
4	May-22	NaN	16.2

```
In [15]: fig = plt.figure() # create an empty figure
         plt.bar(predicted_data['date'], predicted_data['solar_radiation']) #set the dates as the x-axis and the predicted-so
         plt.ylabel("Solar radiation") # set the label for the y-axis
         plt.xlabel("Dates") # set the label for the x-axis
         fig.set_size_inches(20, 5) #set the size of the figure
         plt.show()
```

Figure 3.11: Prediction Interface of Solar Outcome Using Jupyter.

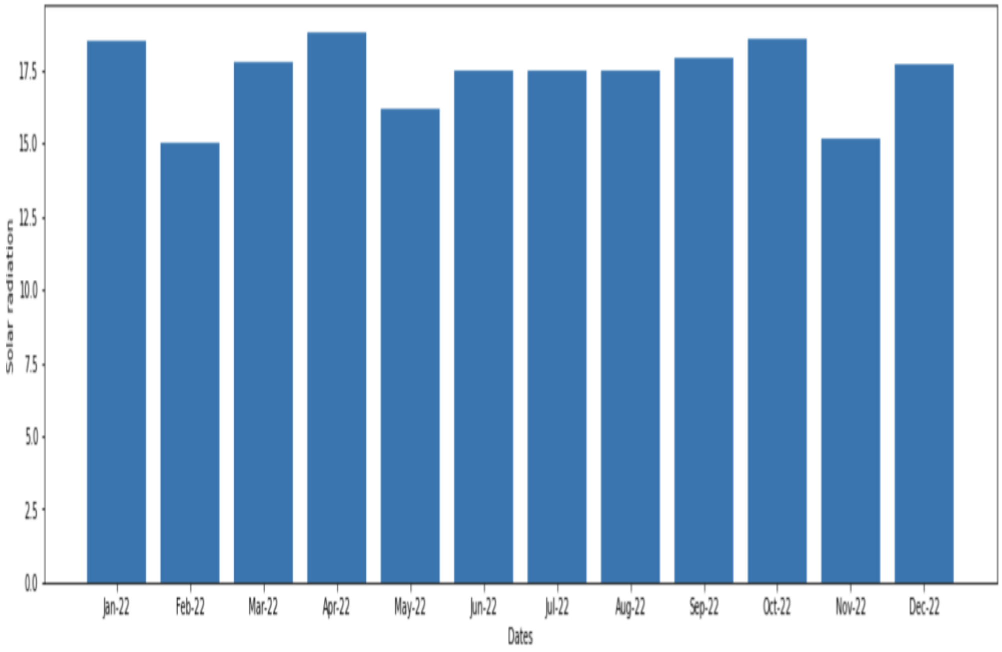


Figure 3.12: Prediction Interface of Solar Outcome Using Jupyter.

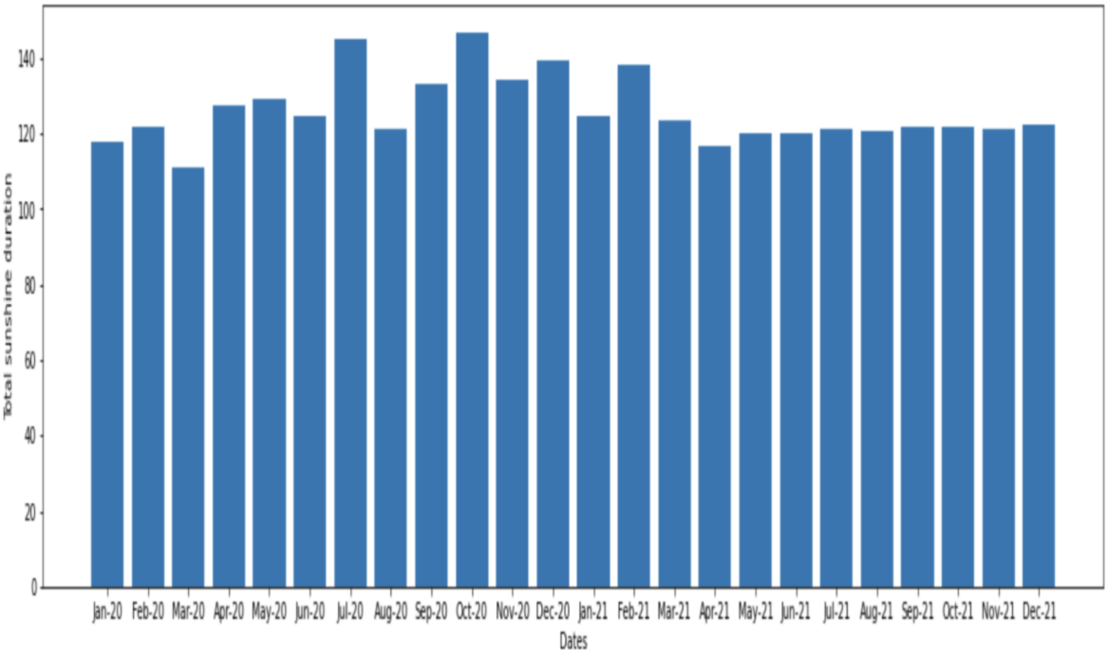


Figure 3.13: Prediction Interface of Total Sunshine Duration.

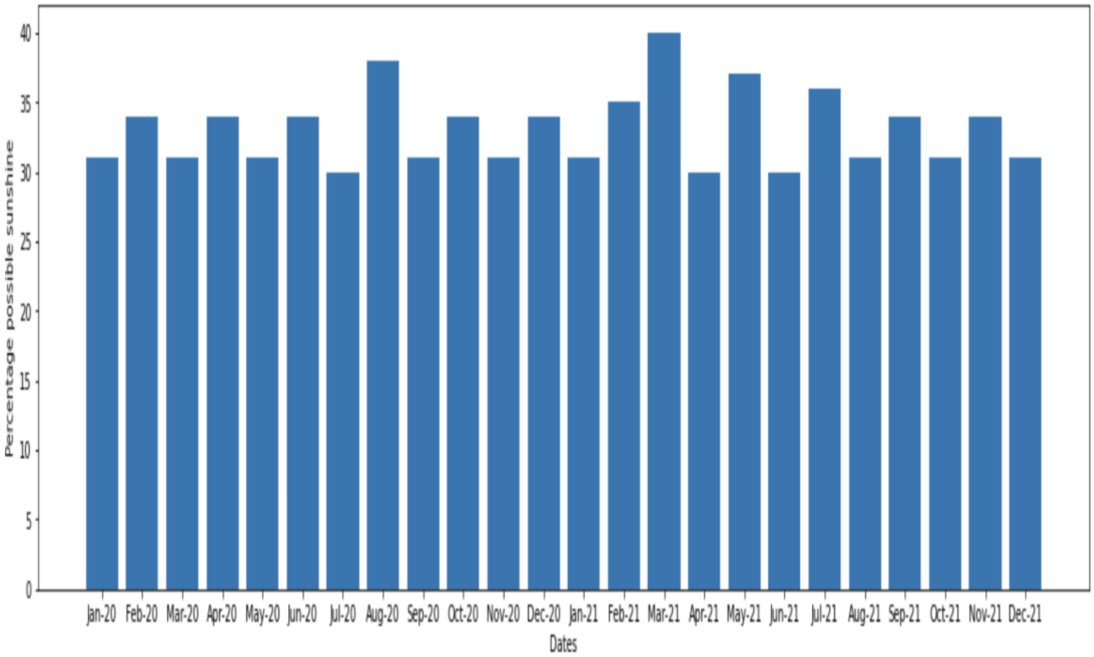


Figure 3.14: Prediction Interface of Percentage Possible Sunshine

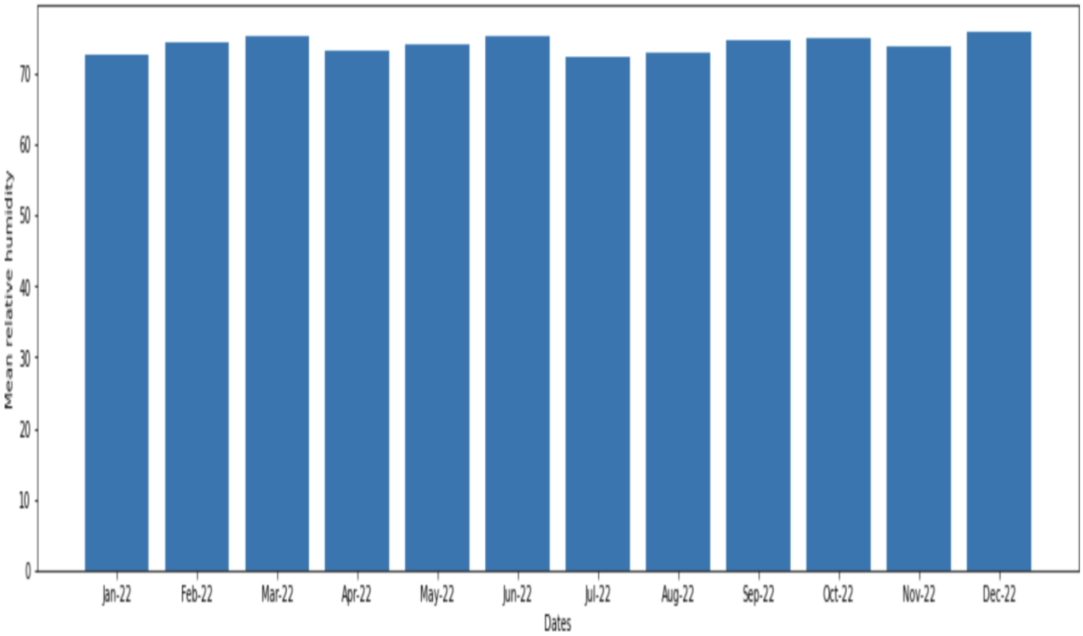


Figure 3.15: Prediction Interface of Mean Relative Humidity.

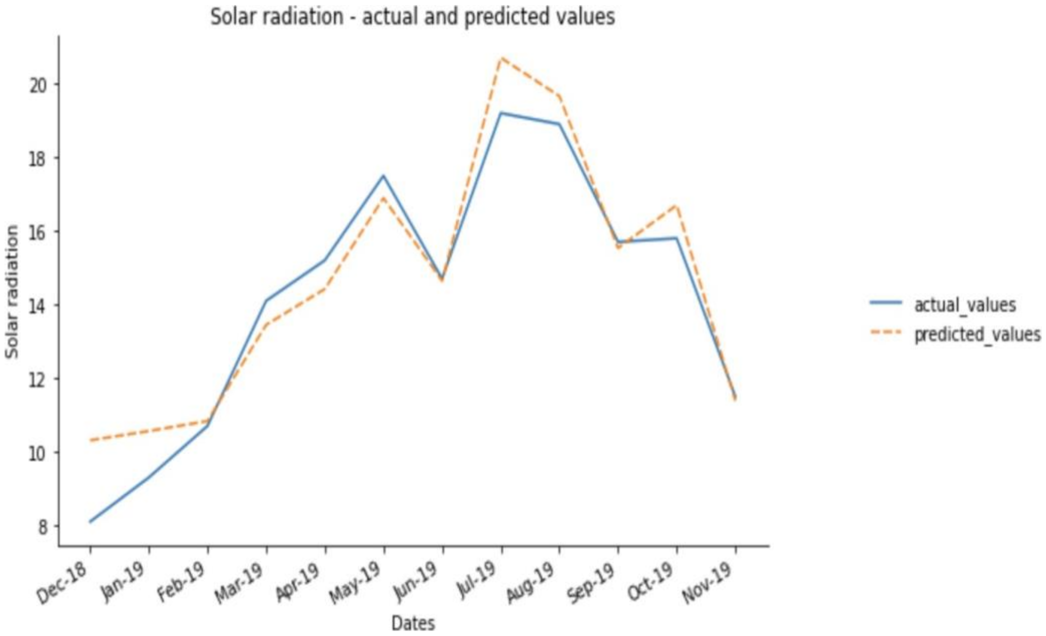
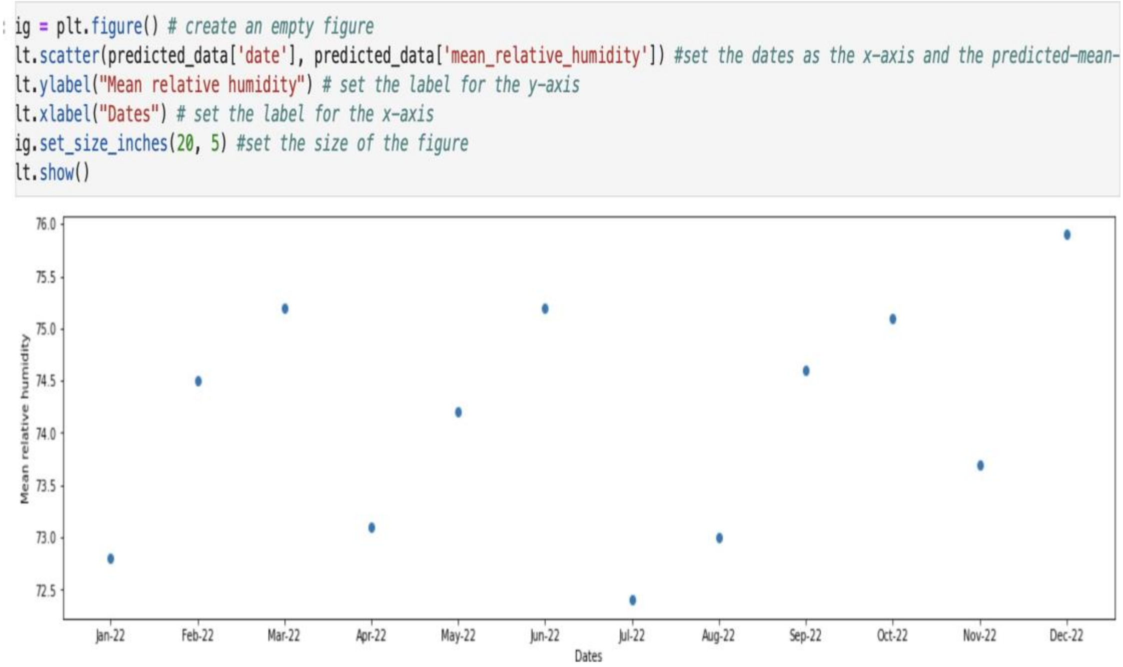


Figure 3.16: Prediction Interface of Solar Outcome Using Jupyter.



```
: ig = plt.figure() # create an empty figure
lt.scatter(predicted_data['date'], predicted_data['mean_relative_humidity']) #set the dates as the x-axis and the predicted-mean-
lt.ylabel("Mean relative humidity") # set the label for the y-axis
lt.xlabel("Dates") # set the label for the x-axis
ig.set_size_inches(20, 5) #set the size of the figure
lt.show()
```

Figure 3.17: Prediction Interface of Solar Outcome Scatter Plot.

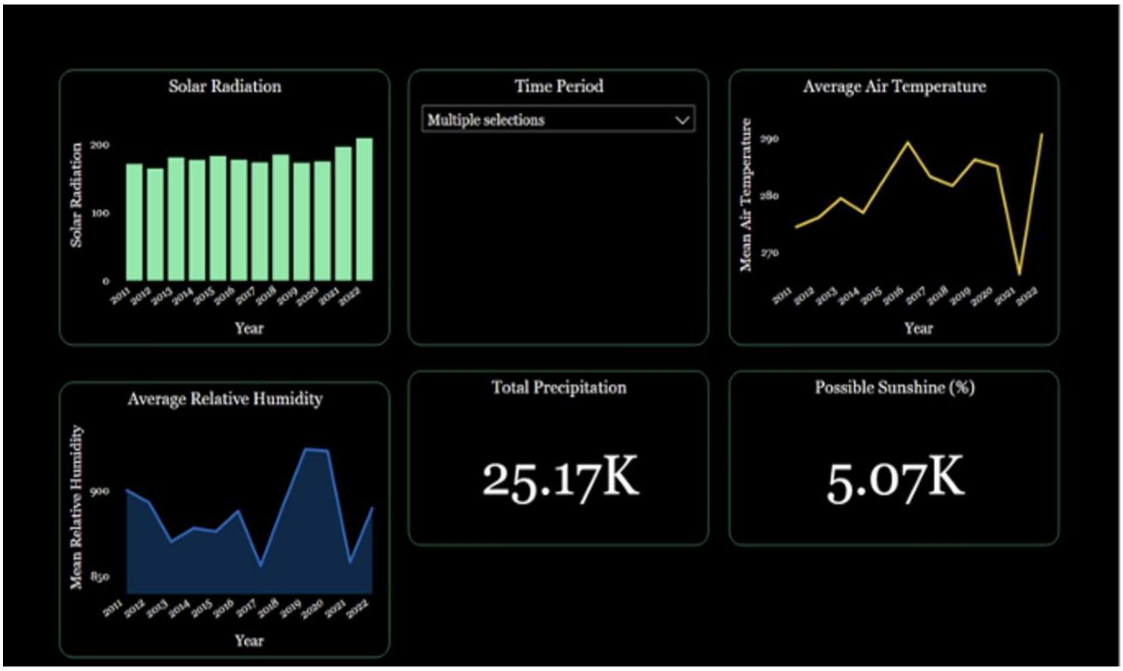


Figure 3.18: Prediction Interface of Solar Outcome Using Jupyter.

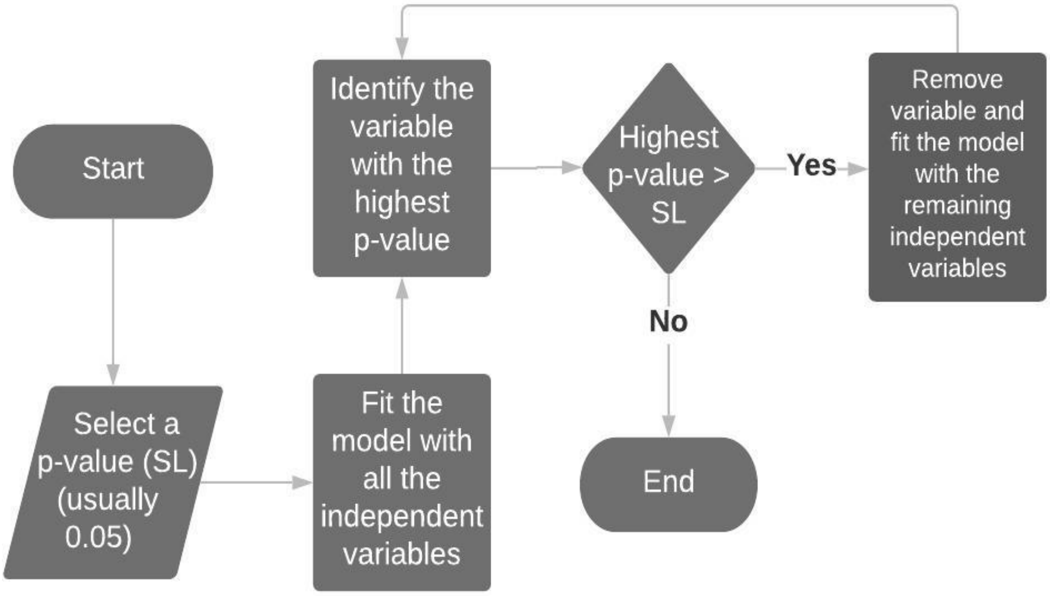

Start
Select a p-value (SL) (usually 0.05)
Fit the model with all the independent variables
Identify the variable with the highest p-value
Highest p-value > SL
Yes
No
Remove variable and fit the model with the remaining independent variables
End

Figure 3.19: Backward Elimination.

- Step-1: Firstly, we need to select a significance level to stay in the model. (SL=0.05)
- Step-2: Fit the complete model with all possible predictors/independent variables.
- Step-3: Choose the predictor which has the highest P-value, such that.
  If P-value ¿ SL, go to step 4. Else Finish, and Our model is ready.
- Step-4: Remove that predictor.
- Step-5: Rebuild and fit the model with the remaining variable

Below is the complete code for the forecast and prediction analysis:

- importing libraries import numpy as nm import matplotlib.pyplot as mtp import pandas as pd

- importing datasets data$_s$*et* = *pd.read*$_c$*sv*('*yourfile*')

- Extracting Independent and dependent Variable x= data$_s$*et.iloc*[:, : −1]*.valuesy* = *data*$_s$*et.iloc*[:, 4]*.values*

train$_t$*est*$_s$*plit*(*x*, *y*, *test*$_s$*ize* = 0.2, *random*$_s$*tate* = 0) • Splitting the dataset into training and test set: from sklearn.model$_s$*electionimporttrain*$_t$*est*$_s$*plitx*$_t$*rain*, *x*$_t$*est*, *y*$_t$*rain*, *y*

- Fitting the MLR model to the training set: from sklearn.linear$_m$*odelimportLinearRegressionregressor* = *LinearRegression*()*regressor.fit*(*x*$_t$*rain*, *y*$_t$*rain*)

- Predicting the Test set result; y$_p$*red* = *regressor.predict*(*x*$_t$*est*)

- Checking the score print ('Train Score: ', regressor.score(x$_t$*rain*, *y*$_t$*rain*))*print*('*TestScore* :', *regressor.score*(*x*$_t$*est*,

3.3 Summary

This chapter highlights the use of the Jupyter interface to analyze the data set, learn the data set with different features inclusive of mean relative humidity, mean air temperature, mean wind speed, total sunshine , solar radiation and more to predict the expected outcome of the solar radiation.

Bibliography

[1] Jae-Gon Kim, Dong-Hyuk Kim, Woo-Sik Yoo, Joung-Yun Lee, Youg Bae Kim "Daily prediction of solar power generation based on weather forecast information in Korea." IET. Predicting the Power output of Solar Panel based on Weather and Air pollution features using machine learning.

[2] Tserenpurev Chuluunsaikhan, Aziz Nasridinov, Woo Seok Choi, Da Bin Choi, Sang Hyun Choi, Young Myoung Kim "Predicting the Power output of Solar Panel based on Weather and Air pollution features using machine learning". Journal of Korea, Multimedia society Vol 24, February 24 pp 222-232

[3] .C.Montgomery,E.A.Peck,andG.G.Vining,Introductiontolinear regression analysis. John Wiley Sons, 2015.

[4] Faizan Jawaid, Khurum NazirJunejo "Predicting Daily Mean Solar Power Using Machine Learning Regression Techniques"

[5] Agada Ihuoma Nkechi, Abdul Motin Howlader and Atsushi Yona "Integration of photovoltaic Energy to the grid, using the virtual synchronous generator control technique". Journal of Energy and Power Engineering 12 (2018) 329-339, doi: 10.17265/1934-8975/2018.07.001

[6] M. E. Lotfy, T. Senjyu, M. A.-F. Farahat, A. F. Abdel-Gawad, and A. Yona, "A frequency control approach for hybrid power system using multi-objective optimization," *Energies*, vol. 10, 2017.

[7] M. E. Lotfy, T. Senjyu, M. A. Farahat, A. F. Abdel-Gawad, and A. Yona, "Enhancement of a small power system performance using multi-objective optimization," *IEEE Access*, vol. 5, pp. 6212-6224, 2017.

[8] ebli, F. Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by Pearson correlation using machine learning," Energy, vol. 224, article 120109, 2021.

[9] . Narvaez, L. F. Giraldo, M. Bressan, and A. Pantoja, "Machine learning for site-adaptation and solar radiation forecasting," Renewable Energy, vol. 167, pp. 333–342, 2021.

[10] . Park, Y. Kim, N. J. Ferrier, S. M. Collis, R. Sankaran, and P. H. Beckman, "Prediction of solar irradiance and photovoltaic solar energy product based on cloud coverage estimation using machine learning methods," Atmosphere, vol. 12, no. 3, p. 395, 2021.

[11] . Mahmud, S. Azam, A. Karim, S. Zobaed, B. Shanmugam, and D. Mathur, "Machine learning based PV power generation forecasting in Alice Springs," IEEE Access, vol. 9, pp. 46117–46128, 2021.

[12] . Nespoli, A. Niccolai, E. Ogliari, G. Perego, E. Collino, and D. Ronzio, "Machine learning techniques for solar irradiation nowcasting: cloud type classification forecast through satellite data and imagery," Applied Energy, vol. 305, article 117834, 2022.

[13] . Rodríguez, F. Martín, L. Fontán, and A. Galarza, “Ensemble of machine learning and spatiotemporal parameters to forecast very short-term solar irradiation to compute photovoltaic generators output power,” Energy, vol. 229, article 120647, 2021.

[14] . Musbah, H. H. Aly, and T. A. Little, “Energy management of hybrid energy system sources based on machine learning classification algorithms,” Electric Power Systems Research, vol. 199, article 107436, 2021.

[15] . Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, “A machine learning-based gradient boosting regression approach for wind power production forecasting: a step towards smart grid environments,” Energies, vol. 14, no. 16, article 5196, 2021.

[16] . Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, “A review of deep learning for renewable energy forecasting,” Energy Conversion and Management, vol. 198, article 111799, 2019.

[17] . Ferrero Bermejo, J. F. Gomez Fernandez, F. Olivencia Polo, and A. Crespo Márquez, “A review of the use of artificial neural network models for energy and reliability prediction A study of the solar PV hydraulic and wind energy sources,” Applied Sciences, vol. 9, no. 9, article 1844, 2019.

[18] . Mosavi, M. Salimi, S. Faizollahzadeh Ardabili, T. Rabczuk, S. Shamshirband, and A. R. Varkonyi-Koczy, “State of the art of machine learning models in energy systems, a systematic review,” Energies, vol. 12, no. 7, article 1301, 2019.

[19] . Ahmed and M. Khalid, “A review on the selected applications of forecasting models in renewable power systems,” Renewable and Sustainable Energy Reviews, vol. 100, pp. 9–21, 2019.

[20] . Zendehboudi, M. A. Baseer, and R. Saidur, “Application of support vector machine models for forecasting solar and wind energy resources: a review,” Journal of Cleaner Production, vol. 199, pp. 272–285, 2018.

[21] . K. Das, K. S. Tey, M. Seyedmahmoudian et al., “Forecasting of photovoltaic power generation and model optimization: a review,” Renewable and Sustainable Energy Reviews, vol. 81, pp. 912–928, 2018.

[22] . Voyant, G. Notton, S. Kalogirou et al., “Machine learning methods for solar radiation forecasting: a review,” Renewable Energy, vol. 105, pp. 569–582, 2017.

[23] . Pérez-Ortiz, S. Jiménez-Fernández, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martínez, and S. Salcedo-Sanz, “A review of classification problems and algorithms in renewable energy applications,” Energies, vol. 9, no. 8, article 607, 2016.

[24] . F. Stefenon, M. H. D. M. Ribeiro, A. Nied et al., “Time series forecasting using ensemble learning methods for emergency prevention in hydroelectric power plants with dam,” Electric Power Systems Research, vol. 202, article 107584, 2022.

[25] . Lingelbach, Y. Lingelbach, S. Otte, M. Bui, T. Künzell, and M. Peissner, “Demand forecasting using ensemble learning for effective scheduling of logistic orders,” In International Conference on Applied Human Factors and Ergonomics, Springer, Cham, pp. 313–321, 2021.

[26] . Zhao, Y. Zhang, H. Xu, and T. Han, “Ensemble classification based on feature selection for environmental sound recognition,” Mathematical Problems in Engineering, vol. 2019, 7 pages, 2019.

Chapter 4

Conclusion and Future Work

4.1 Background

It’s not new that power systems are going through a paradigm change and the use of machine learning models to forecast and predict outcome will indeed give the renewable sector an uplift. Renewable energy has recently received a great deal of attention as a result of its long-term viability and minimal impact on the surrounding environment. In the near future, the provision of renewable energy will be one of the most critical challenges to be addressed. Alternative terminology: the inclusion of renewable energy sources into current or future electric power generation systems.

Existing energy challenges, such as increasing supply stability and alleviating regional power shortages, can be solved through the evolution of renewable energy technologies. This creation of diverse energy sources, on the other hand, is interrupted and chaotic as a result of the volatility of the energy market as well as the unpredictable and intermittent renewable energy. Dealing with renewable energy fluctuation in an accurate way prevails as a challenge. The energy system efficiency is improved via energy monitoring with high precision. The application of energy forecasting technologies can assist in the creation, management, and formulation of energy policy at all levels of government. As renewable energy sources become more generally available, it is vital to create cutting-edge technologies for storing this energy [1]–[7].

Several studies have discovered that a variety of machine learning algorithms have been employed to estimate the output of renewable energy resources. With the help of data-driven models, it is possible to make more accurate predictions about renewable energy. With the use of hybrid machine learning algorithms, projections for renewable energy sources have also been enhanced. In order to effectively predict the availability of renewable energy sources, it was required to use a number of time intervals. When it comes to renewable energy forecasting, these criteria have been widely used to evaluate the accuracy and efficiency of machine learning algorithms [8].

There have been a variety of theories and implementations presented that are based on the three fundamental learning principles [9]. As a result, deep learning is capable of achieving characteristic nonlinear features and invariant high-level data configurations, and as a result, it has been applied in a variety of diverse fields with good results.

According to some studies, a single machine learning model has also been used to anticipate the availability of renewable energy sources [10]. Because of the large range of datasets and time steps, prediction ranges, settings, and performance measurements, a single machine learning model cannot improve forecasting performance on a single dataset or time step. There have been a number of studies in renewable energy forecasting that have resulted in hybrid machine learning models or overall prediction methodologies that are intended to improve prediction performance. Significant attention has lately been drawn to support vector machines (SVMs) and deep learning algorithms [11]– [15].

The most important feature were total sunshine duration and solar radiation, although various temperature related variables contributed towards solar irradiance prediction accuracy. The prediction of solar radiation is very important to determine the amount of energy that can be generated in a day. Linear regression tells us exactly the outcome expected. After importing all libraries needed, the data is cleaned and trained. Linear regression models have predictive power but also many shortcomings like low values were overestimated while high values were underestimated, also residuals violate homoscedasticity and normality. Further work like the use of distance-weighted average of weather forecasts (at multiple grid points) as the weather forecast variable features.

4.1.1 Conclusion

In this work, we have compared time series techniques and machine learning techniques for solar energy forecasting in Japan. We find that employing time series models is a complex procedure due to the non-stationary energy time series. In contrast, machine learning techniques were more straight forward to implement. In particular, we find that the Artificial Neural Networks and Grad/ient Boosting Regression Trees performs best on average across all sites.

Firstly, different libraries from python are imported, and the csv file is read, followed by a guided sequence created for the solar radiation data frame. ilog access the csv file to give an output on the terminal and arrange the excel to read rows and column. Also using the root mean squared error (RMSE) to assess how well a regression model fits a dataset which is to calculate the root mean square error, which tells us about the average distance between the predicted values from the model and the actual values in the dataset. The lower the RMSE, the better a given model can “fit” a dataset. However, the range of the dataset is important in determining whether a given RMSE value is “low” or not. For this case the Mean absolute error, the mean squared error, and the root mean squared error gives low values which indicates precise analysis of the model. The random forest Library was important to determine the mean absolute error, mean squared error and root mean squared error.

This study has compared the different models on a general level. For further research, we suggest continuing comparing different machine learning techniques in depth while using feature engineering approaches of numerical weather predictions.

It is also important to note that an integrated machine learning model and the statistical approach are used to anticipate future solar power generation from renewable energy plants. Hybrid models improves accuracy by integrating machine learning methods and the statistical method. In order to improve the accuracy of the suggested model, an ensemble of machine learning models can be used. When comparing the performance of an ensemble model that integrates all of the combination strategies to standard individual models, the suggested ensemble model outperformed the conventional individual models. According to findings, a hybrid model that made use of both machine learning and statistics outperformed a model that made sole use of machine learning in its performance. In future work, the proposed method can improvise the performance, accuracy, and the other metrics using several deep learning mechanisms.

4.1.2 Future Research

Additional input variables can be included in the forecasting process such as weather data, customers’ classes and event day, instead of only the load data. Besides, other methods may be implemented such as Neural Network, Fuzzy Logic as well as hybrid method.

4.2 Summary

The basis of a multiple linear regression is to assess whether one continuous dependent variable can be predicted from a set of independent (or predictor) variables. Or in other words, how much variance in a continuous dependent variable is explained by a set of predictors. Certain regression selection approaches are helpful in testing predictors, thereby increasing the efficiency of analysis.

Bibliography

[1] R. Kyoho, T. Goya, W. Mengyan, T. Senjyu, A. Yona, T. Funabashi, C. Kim, “Optimal operation of thermal generating units and smart houses considering transmission constraints” *Power Electronics and Drive Systems (PEDS)*, 2013 IEEE 10th International Conference, 22-25 April, 2013, Kitakyushu, Japan. 1225 - 1230. DOI- 10.1109/PEDS.2013.6527206.

[2] A. Gholami, J. Ansari, M. Jamei, A. Kazemi, “Environmental/economic dispatch incorporating renewable energy sources and plug-in vehicles”, *IET Generation, Transmission & Distribution.*, vol. 8, pp.2183-2198, 2014

[3] M. A. Bou-Rabeea, D. H. Hanb, G. H. Choeb, “A photovoltaic power-generation system with peak power cut function”, *International Journal of Sustainable Energy.*, vol. 32 pp.506-515, 2015.

[4] M. Akmal, B. Fox, J. D. Morrow, T. Littler, “Impact of heat pump load on distribution networks”, *IET Generation, Transmission & Distribution.*, vol. 8, pp.2065-2073, 2014.

[5] Z. Lu, C. Lu, T. Feng, H. Zhao, “Carbon dioxide capture and storage planning considering emission trading system for a generation corporation under the emission reduction policy in China”, *IET Generation, Transmission & Distribution.*, pp.10, DOI:10.1049/iet-gtd.2014.0060., 2014

[6] S. G. Malla, C. N. Bhende, “Enhanced operation of stand-alone “Photovoltaic-Diesel Generator-Battery” system”, *Electric Power Systems Research.*, vol. 107, pp.250-257, 2015

[7] A. M. Howlader, N. Urasaki, A. Yona, T. Senjyu, A. Y. Saber, “A review of power smoothing methods for wind energy conversion systems”, *Renewable and Sustainable Energy Reviews.*, vol. 26, pp.135-46, 2013.

[8] R. Kyoho, T. Goya, W. Mengyan, T. Senjyu, A. Yona, T. Funabashi, C. Kim, “Thermal units commitment with demand response to optimize battery storage capacity” *Power Electronics and Drive Systems (PEDS)*, 2013 IEEE 10th International Conference, 22-25 April, 2013, Kitakyushu, Japan. 1207 – 1212. DOI- 10.1109/PEDS.2013.6527203.

[9] M. S. Khalid, M. A. Abido, “A novel and accurate photovoltaic simulator based on seven-parameter model”, *Electric Power Systems Research.*, vol. 116, pp.243-251, 2104.

[10] J. Saebi, M. H. Javidi, M. O. Buygi, “Toward mitigating wind-uncertainty costs in power system operation: A demand response exchange market framework”, *Electric Power Systems Research.*, vol. 119, pp.157-67,2015.

[11] A. M. Howlader, N. Urasaki, A. Yona, T. Senjyu, A. Y. Saber, “Design and implement a digital H_∞ robust controller for a MW-Class PMSG-based grid-interactive wind energy conversion system”, *Energies.*, vol. 6, pp.2084-2109, 2013.

[12] S. Mohammadi, B. Mozafari, S. Solymani, T. Niknam, “Stochastic scenario-based model and investigating size of energy storages for PEM-fuel cell unit commitment of micro-grid considering profitable strategies”, *IET Generation, Transmission & Distribution.*, vol. 8, pp. 1228-1243, 2014.

[13] R. Kyoho, T. Goya, W. Mengyan, T. Senjyu, A. Yona, T. Funabashi, C. H. Kim, “Optimal operation of thermal generating units and smart houses considering transmission constraints”, *Power Electronics and Drive Systems (PEDS),* 2013 IEEE 10th International Conference, 22-25 April, 2013, Kitakyushu, Japan. 1225-1230. DOI- 10.1109/PEDS.2013.6527206.

[14] M. R. Ansari, N. Amjady, B. Vatani, “Stochastic security-constrained hydrothermal unitcommitment considering uncertainty of load forecast, inflows to reservoirs and unavailability of units by a new hybrid decomposition strategy,” *IET Generation, Transmission & Distribution.*, vol. 8, pp.1900-1915, 2014.

[15] D. Choling, P. Yu, B. Venkatesh, “Effects of security constraints on unit commitment with wind generators”, *Power & Energy Society General Meeting.*, pp. 1-6, 2009.

[16] P. Denholm, M. Hand, “Grid flexibility and storage required to achieve very high penetration of variable renewable electricity”, *Energy Policy.*, vol. 39, pp.1817-30, 2011.

[17] M. Lehtonen, M. Z. Degefa, M. Humayun, A. Safdarian, M. Koivisto, R. J. Millar, “Unlocking distribution network capacity through real-time thermal rating for high penetration of DGs”, *Electric Power Systems Research.*, vol. 117, pp.36-46, 2014.

[18] J. M. Morales, C. L. Mancha, C. Real, A. J. Conejo, J., Perez-Ruiz, “Economic valuation of reserves in power systems with high penetration of wind power”, *IEEE Trans. on Power Systems.*, vol. 24, pp. 900-10, 2009.

[19] E. Nasrolahpour, H. Ghasemi, “A stochastic security constrained unit commitment model for reconfigurable networks with high wind power penetration”, *Electric Power Systems Research.*, (Early access.)

[20] J. Wang, M. Shahidehpour, Z. Li, “Security-constrained unit commitment with volatile wind power generation”, *IEEE Trans. Power Systems.*, vol. 23, pp.1319-27, 2008.

[21] H. Yamin, S. Al-Agtash, M. Shahidehpour, “Security-constrained optimal generation scheduling for GENCOs”, *IEEE Trans. on Power Systems.*, vol. 19, pp.1365-1372, 2004.

[22] M. A. Ortega-Vazquez, “Optimal scheduling of electric vehicle charging and vehicle-to-grid services at household level including battery degradation and price uncertainty”, *IET Generation, Transmission & Distribution.*, vol. 6, pp.1007-1016, 2014.

[23] C. Liu, J. Wang, A. Botterud, Y. Zhou, A. Vyas, “Assessment of impacts of PHEV charging patterns on wind-thermal scheduling by stochastic unit commitment”, *IEEE Trans. on Smart Grid.*, vol. 3, pp. 675-83, 2012.

[24] J. R. Pillai, B. Bak-Jensen,“Integration of vehicle-to-grid in the Western Danish Power system”, *IEEE Trans. Smart Grid.*, vol. 2, pp.12-19, 2012.

[25] C. Fernandes, P. Frias, J. M. Latorre, “Impact of vehicle-to-grid on power system operation costs: The Spanish case study”, *Applied Energy.*, vol. 96, pp.194-202, 2012.

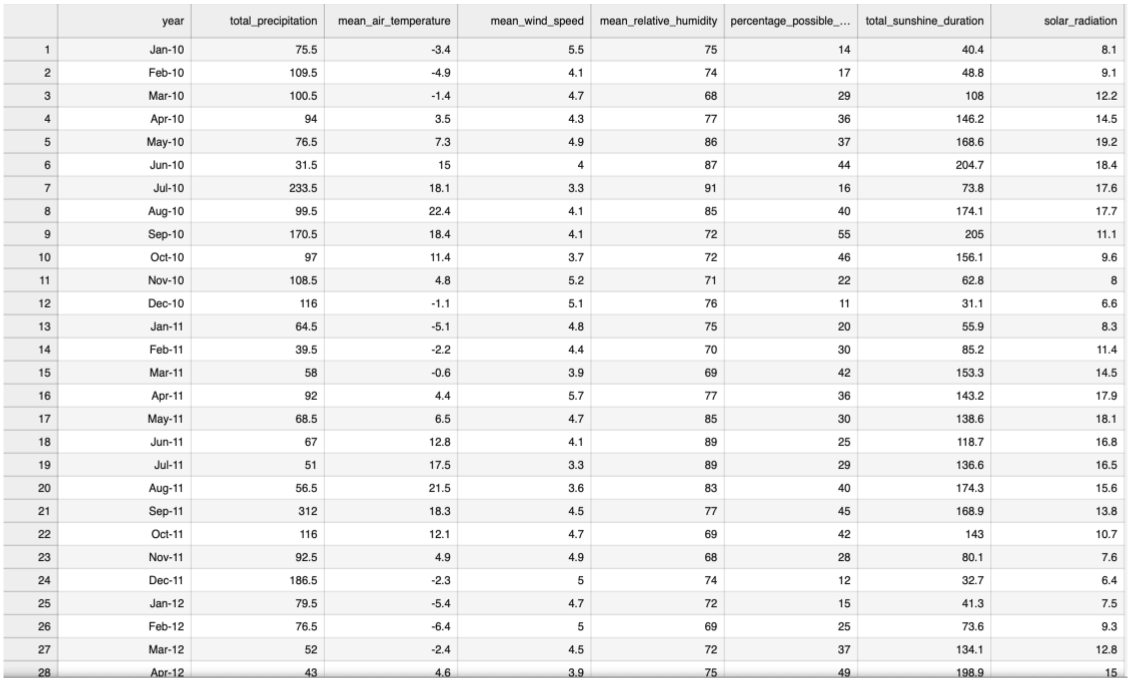
[26] M. Behrangrad, H. Sugihara, T. Funaki, “Optimal spinning reserve procurement using demand response in simultaneous clearing process”, IEEE PES/IAS conference on Sustainable Alternative Energy (SAE)., Valensia, Spain, 28-30 Sept. 2009.

[27] A. Khodaei, M. Shahidehpour, S. Bahramirad, “SCUC with hourly demand response considering intertemporal load characteristics”, *IEEE Trans. on Smart Grid.*, vol. 2, pp.564-571,2011.

Chapter A

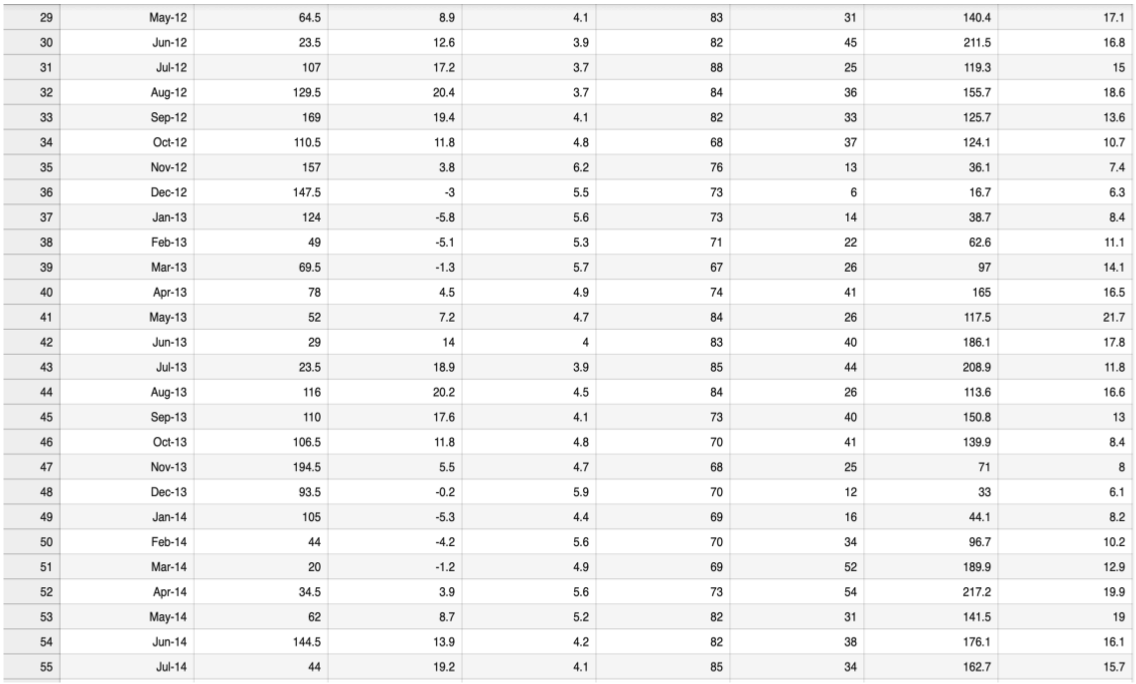
Appendix

A .1 Simulation Data for Parameters Identification



	year	total_precipitation	mean_air_temperature	mean_wind_speed	mean_relative_humidity	percentage_possible_...	total_sunshine_duration	solar_radiation
1	Jan-10	75.5	-3.4	5.5	75	14	40.4	8.1
2	Feb-10	109.5	-4.9	4.1	74	17	48.8	9.1
3	Mar-10	100.5	-1.4	4.7	68	29	108	12.2
4	Apr-10	94	3.5	4.3	77	36	146.2	14.5
5	May-10	76.5	7.3	4.9	86	37	168.6	19.2
6	Jun-10	31.5	15	4	87	44	204.7	18.4
7	Jul-10	233.5	18.1	3.3	91	16	73.8	17.6
8	Aug-10	99.5	22.4	4.1	85	40	174.1	17.7
9	Sep-10	170.5	18.4	4.1	72	55	205	11.1
10	Oct-10	97	11.4	3.7	72	46	156.1	9.6
11	Nov-10	108.5	4.8	5.2	71	22	62.8	8
12	Dec-10	116	-1.1	5.1	76	11	31.1	6.6
13	Jan-11	64.5	-5.1	4.8	75	20	55.9	8.3
14	Feb-11	39.5	-2.2	4.4	70	30	85.2	11.4
15	Mar-11	58	-0.6	3.9	69	42	153.3	14.5
16	Apr-11	92	4.4	5.7	77	36	143.2	17.9
17	May-11	68.5	6.5	4.7	85	30	138.6	18.1
18	Jun-11	67	12.8	4.1	89	25	118.7	16.8
19	Jul-11	51	17.5	3.3	89	29	136.6	16.5
20	Aug-11	56.5	21.5	3.6	83	40	174.3	15.6
21	Sep-11	312	18.3	4.5	77	45	168.9	13.8
22	Oct-11	116	12.1	4.7	69	42	143	10.7
23	Nov-11	92.5	4.9	4.9	68	28	80.1	7.6
24	Dec-11	186.5	-2.3	5	74	12	32.7	6.4
25	Jan-12	79.5	-5.4	4.7	72	15	41.3	7.5
26	Feb-12	76.5	-6.4	5	69	25	73.6	9.3
27	Mar-12	52	-2.4	4.5	72	37	134.1	12.8
28	Apr-12	43	4.6	3.9	75	49	198.9	15

Figure A .1: CSV Data for Weather Features used for Predictions.



29	May-12	64.5	8.9	4.1	83	31	140.4	17.1
30	Jun-12	23.5	12.6	3.9	82	45	211.5	16.8
31	Jul-12	107	17.2	3.7	88	25	119.3	15
32	Aug-12	129.5	20.4	3.7	84	36	155.7	18.6
33	Sep-12	169	19.4	4.1	82	33	125.7	13.6
34	Oct-12	110.5	11.8	4.8	68	37	124.1	10.7
35	Nov-12	157	3.8	6.2	76	13	36.1	7.4
36	Dec-12	147.5	-3	5.5	73	6	16.7	6.3
37	Jan-13	124	-5.8	5.6	73	14	38.7	8.4
38	Feb-13	49	-5.1	5.3	71	22	62.6	11.1
39	Mar-13	69.5	-1.3	5.7	67	26	97	14.1
40	Apr-13	78	4.5	4.9	74	41	165	16.5
41	May-13	52	7.2	4.7	84	26	117.5	21.7
42	Jun-13	29	14	4	83	40	186.1	17.8
43	Jul-13	23.5	18.9	3.9	85	44	208.9	11.8
44	Aug-13	116	20.2	4.5	84	26	113.6	16.6
45	Sep-13	110	17.6	4.1	73	40	150.8	13
46	Oct-13	106.5	11.8	4.8	70	41	139.9	8.4
47	Nov-13	194.5	5.5	4.7	68	25	71	8
48	Dec-13	93.5	-0.2	5.9	70	12	33	6.1
49	Jan-14	105	-5.3	4.4	69	16	44.1	8.2
50	Feb-14	44	-4.2	5.6	70	34	96.7	10.2
51	Mar-14	20	-1.2	4.9	69	52	189.9	12.9
52	Apr-14	34.5	3.9	5.6	73	54	217.2	19.9
53	May-14	62	8.7	5.2	82	31	141.5	19
54	Jun-14	144.5	13.9	4.2	82	38	176.1	16.1
55	Jul-14	44	19.2	4.1	85	34	162.7	15.7

Figure A .2: CSV Data for Weather Features used for Predictions.