Modeling and Forecasting of Power Plant Generation using Machine Learning Approach

By

Qazi Usman Najeeb
00000117193

Supervisor

Dr. Imran Mahmood

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Information Technology

Approval

It is certified that the contents and form of the thesis entitled "Modeling and Forecasting of Power Plant Generation using Machine Learning Approach" submitted by Qazi Usman Najeeb has been found satisfactory for the requirement of the degree.

Advisor: Dr. Imran Mahmood

Signature: _____

Date: _____

Committee Member 1:

Dr. Fahad Javed

Signature: _____

Date: _____

Committee Member 2:

Dr. Sidra Sultana

Signature: _____

Date: _____

Committee Member 3:

Dr. Hasan Arshad Nasir

Signature: _____

Date: _____

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr. Qazi Usman Najeeb, (Registration No 117193), of School of Electrical Engineering and Computer Science (SEECS) has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: Dr. Imran Mahmood

Date: _____

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: Qazi Usman Najeeb

Signature: _____

Acknowledgment

I thank Allah for giving me the strength to make this little effort reach completion and for constant reminder that His plans are better than my dreams.

I am grateful to my Supervisor, Dr. Imran Mahmood for his patient guidance, encouragement and advice to complete my thesis and for giving me numerous opportunities to learn and grow.

I appreciate contribution of Chief Engineer Arfan Miana for providing me with required data of Tarbela Powerhouse towards development and analysis of this study.

Lastly, I thank all my class fellows for their support where and when ever needed.

Table of Contents

Modeling and Forecasting of Power Plant Generation using Machine Learning Approach 1
Certificate of Originality 1-4
Acknowledgment 1-5
Table of Contents 1-6
List of Abbreviation 1-1
List of Tables 1-2
List of Figures 1-3
Abstract 1-5
Chapter 1 Introduction 1-1
1.1 Challenges 1-2
1.2 Problem Statement 1-3
1.3 Solution Statement 1-3
1.4 Key Contributions 1-4
1.5 Research Impact 1-4
1.6 Related Industry 1-4
1.7 Model of Study 1-6
1.8 Thesis Organization 1-7
1.8.1 Chapter 2: Background 1-7
1.8.2 Chapter 3: Literature 1-7
1.8.3 Chapter 4: Case Study 1-7
1.8.4 Chapter 5: Methodology and Results 1-7
1.8.5 Chapter 6: Conclusion and Future work 1-7
Chapter 2 Background 2-8
2.1 Renewable Energy 2-8
2.1.1 Types of Renewable Energy 2-8
2.2 Potential of Hydro Power in Pakistan 2-12
2.2.1 Types of Hydropower Sources: 2-13
2.3 Hydro Power Generation 2-13
2.4 Machine Learning 2-14
2.5 Machine Learning Algorithms Selected for this Study 2-16
2.5.1 Multiple Linear Regression 2-16
2.5.2 Support Vector Regression 2-16
2.5.3 Decision Tree 2-19
2.5.4 Random Forest 2-21
Chapter 3 Literature Review 3-22
3.1 Area of Research 3-22
3.1 Research on Hydro Power Plants Around the World 3-23
Chapter 4 Methodology and Results 4-29
4.1 Data Analysis 4-29
4.1.1 Case Study: Tarbela Power Plant 4-29
4.1.2 Data Set 4-31
4.1.3 Pre-processing 4-34
4.1.4 Training & Test Sets 4-36
4.2 Model Development 4-36
4.2.1 Feature Extraction 4-36
4.2.2 Model Development Process 4-40
4.2.3 Multiple Linear Regression 4-41
4.2.4 Support Vector Regression 4-43
4.2.5 Decision Trees 4-44
4.2.6 Random Forest 4-46
4.2.7 Artificial Neural Networks 4-48
4.2.8 Comparison of Results 4-50

Chapter 5 Conclusion and Future Work 5-51
5.1 CONCLUSION 5-51
5.2 FUTURE WORK 5-51
References 5-52

List of Abbreviation

ABM	Machine Learning
MLR	Multiple Linear Regression
SVR	Support Vector Regression
DT	Decision Trees
RF	Random Forest
ANN	Artificial Neural Networks
WAPDA	Water and Power Development Authority
PPIB	Private Power Infrastructure Board
IPPs	Individual Power Plants
PCRET	Pakistan Council of Renewable Energy Technologies
SHP	Small Hydro Power Plant
ABC	Artificial Bee Colony

List of Tables

Table 1: Literature Review in the Area of Machine Learning 3-24
Table 2: Literature Review in the Area of Modeling and Simulation 3-27
Table 3: Output Power of Installed Turbines 4-31
Table 4: Parameters 4-35
Table 5: Comparision of Model Parameters vs Hyperparameters 4-40
Table 6: RMSE and MAPE of MLR on Training and Test Set 4-42
Table 7: Cross Validation Results and Accuracy of MLR Model on Training and Test Set 4-42
Table 8: RMSE and MAPE of NU-SVR on Training and Test Set 4-44
Table 9: Cross Validation Results and Accuracy of NU-SVR Model on Training and Test Set 4-44
Table 10: RMSE and MAPE of Decision Tree on Training and Test Set 4-46
Table 11: Cross Validation Results and Accuracy of Decision Tree Model on Training and Test Set 4-46
Table 12: RMSE and MAPE of Random Forest on Training and Test Set 4-47
Table 13: Cross Validation Results and Accuracy of Random Forest Model on Training and Test Set 4-47
Table 14: RMSE and MAPE of ANN on Training and Test Set 4-50
Table 15: Cross Validation Results and Accuracy of ANN Model on Training and Test Set 4-50
Table 16: RMSE and MAPE of all Models on Training and Test Set 4-50
Table 17: Cross Validation Results and Accuracy of all Models on Training and Test Set 4-50

List of Figures

Figure 1: Primary Energy Consumption by Source 2011 [11] 1-3
Figure 2: Stakholders of Energy Planning and Forecasting 1-5
Figure 3: Model of Study 1-6
Figure 4: Thesis Organization 1-7
Figure 5: Types of Renewable Energy Sources 2-8
Figure 6: Hydropower Generation 2-9
Figure 7: Solar Energy Farm 2-10
Figure 8: Geothermal Energy Plant 2-10
Figure 9: Bio Energy 2-11
Figure 10: Solar Energy Farm 2-11
Figure 11: Forecasted Installed Capacity of Different Energy Sources by the End of 2030 [13] 2-12
Figure 12: Cost of Renewable Energy Sources 2-12
Figure 13: Hydropower Sources in Pakistan 2-13
Figure 14: Overview of Hydropower Station 2-14
Figure 21: Types of Machine Learning 2-15
Figure 15: A Linear SVM Classification [OL-2] 2-17
Figure 16: Example showing linear SVM boundry is not obtained on given data set [OL-3]. 2-18
Figure 17: A Decision boundry gained after applying Kernal [OL-3] 2-18
Figure 18: A) No Kernel Applied B) New features gained after applying Polynomial Kernal 2-19
Figure 19: Working of Decision Tree [OL-5] 2-20
Figure 20: Areas of Research 3-22
Figure 22: Location of Tarbela Hydro Power Plant 4-29
Figure 23: Capacity of Hydropower Projects in Pakistan 4-30
Figure 25: Structure of Energy Generation Data Set 4-31
Figure 24: Daily Energy Generation of Tarbela Power Plant 1993-2016 4-31
Figure 26: Evaporation Rate in Tarbela 4-32
Figure 27: Daily Inflow of River Indus at Tarbela 4-32
Figure 28: Daily Precipitation of River Indus at Tarbela 4-33
Figure 29: Daily inflow of River Indus at Tarbela 4-33
Figure 30: Structure of Energy Generation Data Set 4-34
Figure 31: Features of Tarbela Power Plant 4-35
Figure 32: Features Extracted for Energy Forecasting Model 4-36
Figure 33: Types of Feature Selection Approaches 4-37
Figure 34: Steps of Filter Method 4-37
Figure 35: Steps of Wrapper Method 4-38

Figure 36: Steps of Embedded Method 4-38
Figure 37: Overview of Backward Elimination Method..... 4-39
Figure 38: Backward Elimination Results 4-39
Figure 39: Actual vs Predicted Result of MLR on Train Data..... 4-41
Figure 40: Actual vs Predicted Result of MLR on Test Data 4-42
Figure 41: Actual vs Predicted Result of SVR on Train Data..... 4-43
Figure 42: Actual vs Predicted Result of SVR on Test Data 4-43
Figure 43: Graphical Representation of Decision Tree 4-44
Figure 44: Actual vs Predicted Result of Decision Tree on Train Data..... 4-45
Figure 45: Actual vs Predicted Result of Linear Regression on Test Data..... 4-45
Figure 46: Actual vs Predicted Result of Random Forest on Train Data 4-46
Figure 47: Actual vs Predicted Result of Random Forest on Train Data 4-47
Figure 48: Graphical Representation of a 2 Layer ANN..... 4-48
Figure 49: Actual vs Predicted Result of ANN on Train Data..... 4-49
Figure 50: : Actual vs Predicted Result of ANN on Test Data 4-49

Abstract

Electricity usage planning is a main concern for electricity stakeholders in a country. The exhaustible resources are not enough to address energy demand in our country. It comes with varied problems such as price, environmental hazards and availability of resources. Renewable energy resources, mainly hydropower energy is an ideal solution to energy deficiency in Pakistan. However, to meet the compelling demand for electricity and to deal with different uncertainties involved in this process, development of sustainable policies through proper planning is becoming increasingly challenging.

To overcome the above-mentioned problems, we propose to study electricity generation trend of Tarbela Power Plant. Historical data of last 5 years on a daily resolution of generation is used to develop regression models and predict energy generation. This study shall help in analysis of future supply of electricity produced by the Powerhouse.

A prediction-based model for electricity production will be useful in forecasting future energy generation and therefore will play a significant role in electric energy planning. It will allow stakeholders to visualize operational excellence of the Plant and incorporate many influencing factors such as decision making on tariffs, policy regulations, investments, available resources and the environmental factors in the country. Once the proposed model is validated using historical data, it will be accredited as a decision support tool for planning future generation needs.

Keywords: *Hydro Power Plant, Tarbela Power Plant, Machine Learning, Multiple Linear Regression, Decision Tree, Random Forest, Artificial Neural Network*

Chapter 1

Introduction

This chapter provides the opening and general information of the research to provide a clear understanding about this thesis. It also covers the problem statement along with solution statement.

Pakistan Energy plays an important role in socio-economic development of a country [1]. It is also an essential part of mankind that offers various daily life functionalities such as lightning houses, lighting offices, charging electronic appliances, providing heating and cooling etc. In addition to that, it acts as a catalyst for production of different industrial items as majority of the heavy machinery works using electricity (a form of energy). Furthermore, energy is also facilitating modern modes of communication (electric vehicles), entertainment (Media, cinema, camera) and medication (X-ray, ECG) [2] [3].

For the past few decades, Pakistan is facing severe energy crises resulting in power outages and load shedding [4]. In the rural areas, conditions are even worst where the duration of load shedding reaches to 20 hours during the extreme weather conditions. Such outages are heavily affecting the industrial productivity of the country. The impact is worst for small and medium scale industries where many such companies are already shutting down or considering of doing so. Energy crises also pose a negative effect on the economy of the county, thus, declining the Gross Domestic Product (GDP) of the country. Similarly, it is a leading cause of poverty, unemployment, exports and competitiveness of Pakistan with other countries [5].

A few causes of energy crises include old and unreliable infrastructure that results in easting energy, energy theft, lack of bill payment, less use of hydro energy and circular debt [6]. Having said that, energy crises do not emerge suddenly, and it takes years of negligence. The poor policies of energy department and government is majorly responsible for such crises. The above-mentioned problems can be resolved by generating more and cheap electricity. Some frequently used energy generating resources are:

- Fossils Energy
- Water (Hydroelectric) Energy
- Wind Energy
- Solar Energy

Currently, the major source of generating energy in Pakistan is through fossils [7]. However, it is resulting in the depletion of fossils due to its finite nature. Also, the generation of energy through fossils involves toxic methods that are not good for environment. Therefore, fossils energy is not a good choice to overcome the energy crises. Solar energy is eco-friendly, but it is only applicable in a few places in Pakistan during a certain time of the year. Likewise, the wind energy, in majority of cases, can

be produced near the coastline. Among all the available choices, water or hydroelectric energy is the appropriate choice due to its eco-friendly nature, cheapness and availability.

Therefore, in this paper, we aim to target the hydropower energy generation and its forecasting. We study the Power generation trend of Tarbela Power Plant. A data set of 23 years on a daily generation will be used to model and forecast the energy generation of the Plant. This study shall help in analysis of future supply of electricity produced by the Powerhouse. Furthermore, various influencing factors like the environment will be coupled with the model to study its sensitivity. This model will be used by analysts to answer different energy related research questions which further will lead the decision makers to adopt optimal choices of future electricity energy planning in the country.

1.1 Challenges

Energy is a key parameter in development of country because it is used in almost every aspect of daily life starting from gadgets, household items to large industries. The modern culture is called the "Energy-intensive culture" due to the overdependence of humans to the energy [8]. In modern era, the Gross Domestic Product (GDP) is directly associated with the energy consumption.

The energy occurs in two major form; exhaustible and renewable. The problem of exhaustible resources is the depletion with the time. It results in the scarcity of resources, hence, causing high prices of electricity. The major challenge of Pakistan is its over dependence on the exhaustible fossil fuel energy generation. It offers following drawbacks:

- The first is the price of such energy is going to increase due to scarcity of the resources.
- Fossils energy also resulting in environmental hazards.
- It is unable to meet the growing demand of energy in Pakistan.
- It is also analyzed the excessive use of conventional fuels such as fossil fuels make the country unstable and limited energy generation harms the stability of national as well as international interests [9].

Such problems can be resolved by using the renewable energy resources. Such resources are cheap in nature and environment friendly. Renewable energy resources are a way forward to overcome the global warming challenge in the world [10]. Currently, very minor proportion of word is using renewable energy as shown in Figure. 1. The energy consumption stats for Organization for Economic Co-operation and Development (OECD) and non-OCED countries is shown in the figure. From analysis, it is observed that major source of energy generation in the world is oil and coal that generates 20000 GW and 25550 GW electricity in OCED countries. However, the use and scope of renewable energies is considerably low.

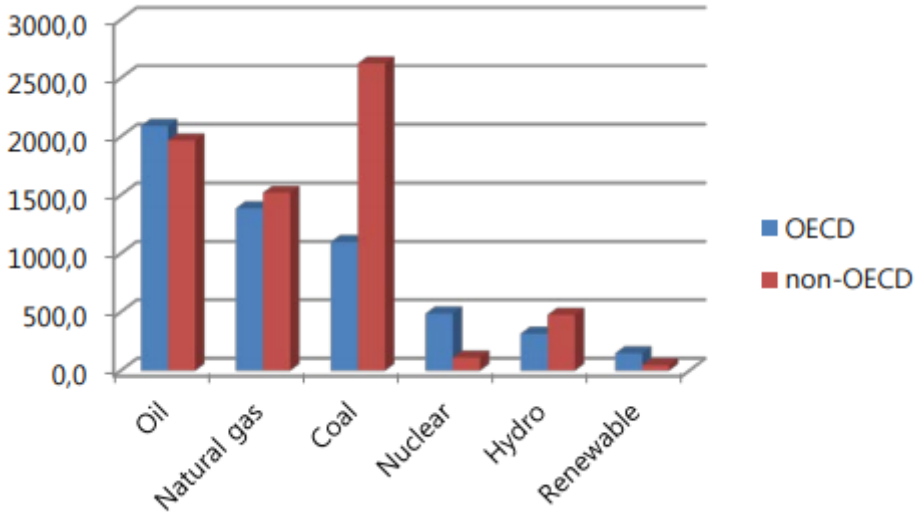

Figure 1: Primary Energy Consumption by Source 2011 [11]

Second problem is large renewable energy resources require very large infrastructures. Therefore, the modelling and forecasting of such infrastructures is very important to avoid any potential losses.

1.2 Problem Statement

Managing electricity demands is a key challenge for authorities considering the limited economy. Pakistan's electricity demand has always been in deficit in its supply. Energy usage planning is a problem for electricity stakeholders in this era of enthralling demand for electric power. Hydroelectricity is one of primary sources of energy of Pakistan while Tarbela Dam has the largest Power Plant generating about 29% of electricity. Electricity planning requires a modelled framework in order to study future demand and supply of electricity. There is a dire need to study the supply capacity of Tarbela Power Plant and understand the effects of related uncertainties to better plan for future energy and assist authorities in opting an approach to forecast the gap between supply and demand.

1.3 Solution Statement

Goal is to develop a prediction model using historic data,

- To study and analyze various regression algorithms in forecasting energy supply of Tarbela Powerhouse.
- Analyze effects of various factors to study sensitivity in estimating short-term future energy.
- Conclude a framework and tool for modeling and visualization of the future supply of electricity energy in Pakistan based on the demand using Tarbela hydro power plant.

1.4 Key Contributions

Pakistan is rich with resources for generating energy from various renewable resources. Country's energy demands are increasing more than its generation capacity with the growing economy in recent years. Electricity production from Hydro sources contributes for a third to the national grid still mass majority is without access of electricity. By virtue of geography, the country has an advantage of utilizing this source for energy production from small, medium and micro level hydel projects. The discussion and efforts continue to increase or at most equalize supply-demand ratio. Stakeholders needs to keep economy and existing resources into account for new ventures. The need for development of sustainable policies through proper planning is becoming increasingly challenging which this study can contribute and act as a starting point for growth in the sector of hydro power technology.

1.5 Research Impact

Our proposed framework analyzes the role of Hydro Power Plant for electricity generation. It is the country's largest source of electricity production using Hydropower technology. Water is used as fuel for generation of clean and cheap energy. This study shall introduce a strategy to foresee future energy demands that'll support in energy planning resulting in meeting future demands. An early warning can be communicated so that decision makers and other stakeholders may take appropriate steps in effective management of resources.

Initially, in this thesis, we are studying and analyzing the roles and advantages of hydropower Energy for electricity production in Pakistan. In our proposed framework, we are modelling and forecasting the hydropower electricity production by Tarbela dam using machine learning algorithms.

A few advantages of it but not limited to is, comparison and analysis of a regression models that fits and validates past data in selecting the best approach to benefit from goal of forecasting. It shall help stakeholders take timely decisions and steps to alter generation parameters in order to meet the demands. This study shall also help improve generation, maintain uniformity and minimize water loss that can be utilized for other purposes in need.

1.6 Related Industry

This study provides appropriate support to both private and governmental organizations that involve in energy planning, supplying and policymaking. Our proposed work and research apply in some of the following areas:

- The reference study could be beneficial in future proof reading of new systems developed and installed to optimize energy generation.
- It could help benefit in energy planning and its demand - supply ratio.
- It is also useful in evaluating the correct timing to perform regular maintenance activities to keep turbines operational without hampering demand.
- Water being the source of energy produce from Hydro Power technology also demands to study water resources and it can also be linked to operation of

water supply utilities, optimal reservoir operation involving multiple objectives of irrigation, and sustainable development of water resources.

A few key stakeholders of this study include:

- Ministry of Water and Power
- Planning Commission of Pakistan
- Water and Power Development Authority (WAPDA)
- Private Power and Infrastructure Board (PPIB)
- Individual Power Plants (IPPs)
- Pakistan Council of Renewable Energy Technologies (PCRET)
- Power Distribution Companies (Tarbela Power Plant, IESCO)

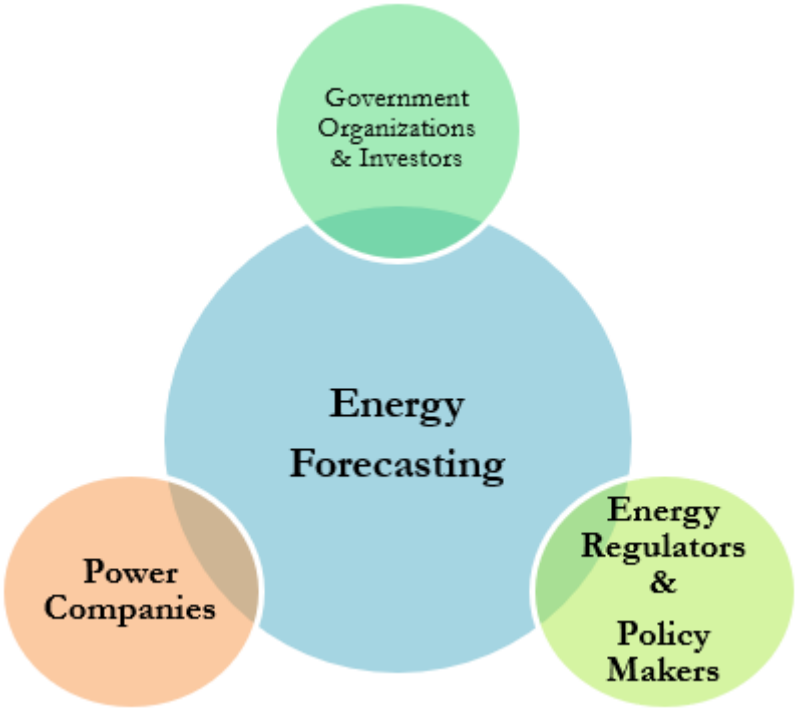

Figure 2: Stakholders of Energy Planning and Forecasting

1.7 Model of Study

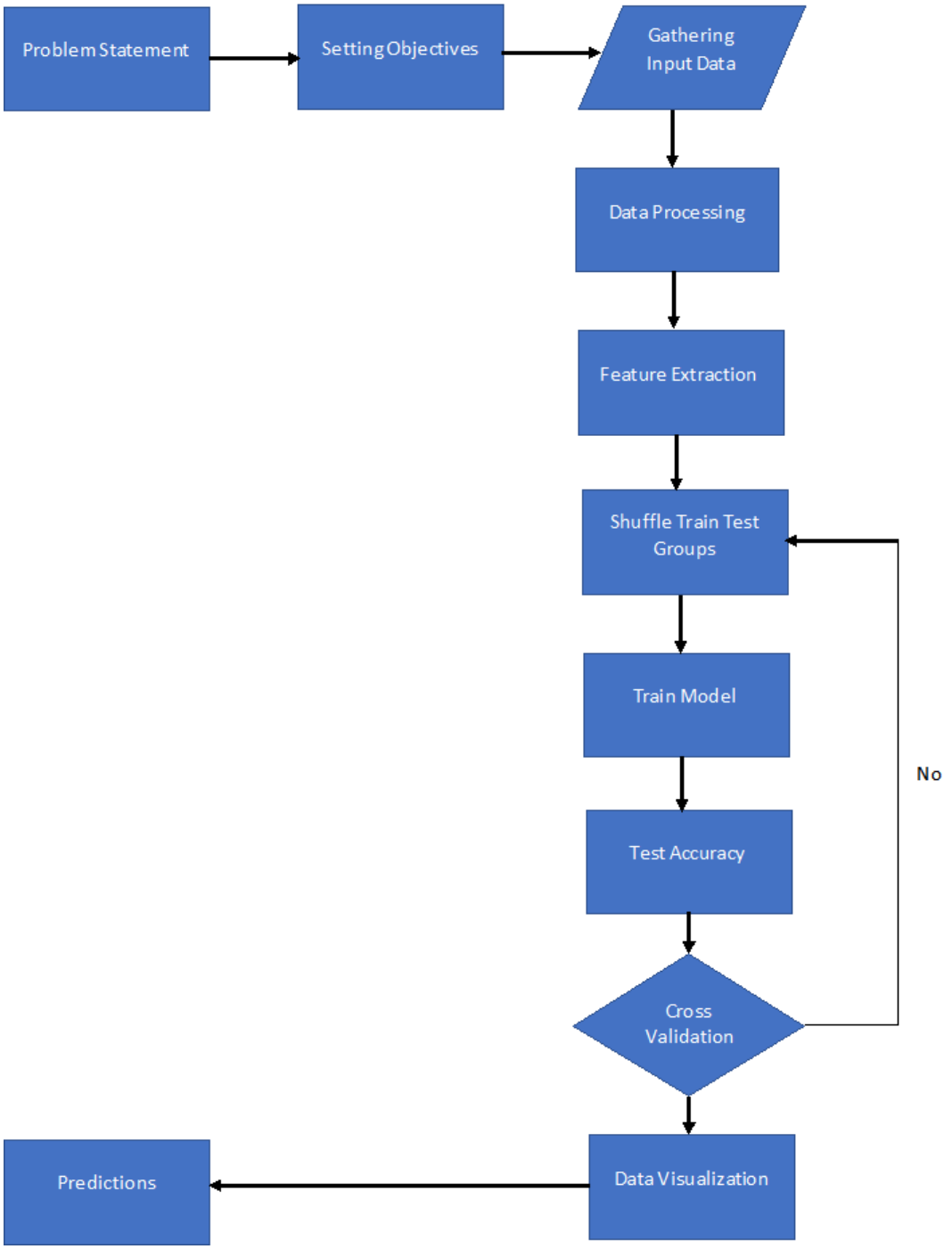

Figure 3: Model of Study

1.8 Thesis Organization

Rest of the thesis is organized in following chapters

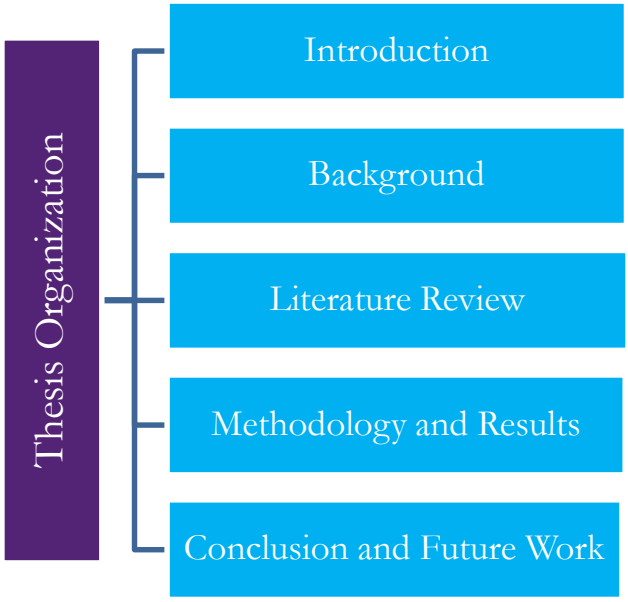

Figure 4: Thesis Organization

1.8.1 Chapter 2: Background

Chapter 2 provides brief overview of Renewable Energy Resources, its types and advantages. This chapter also explains Hydro power potential in Pakistan. Moreover, few preliminary concepts, used in methodology chapter are discussed as well.

1.8.2 Chapter 3: Literature

This chapter explains the work done so far related to hydro power plants, its component and verification of model. It summarizes the hierarchal approaches of various approaches used for different use cases and formulates research directions for this dissertation.

1.8.3 Chapter 4: Case Study

This chapter discusses location and operational capacity of Tarbela Power Plant selected for analysis of this study. It explains the role and importance of the asset in energy generation and its history electricity produced.

1.8.4 Chapter 5: Methodology and Results

Modeling of proposed Hydro Power Plant is carried out in R Programming Software. Brief introduction to the software is provided in this section along with detailed explanation of proposed machine learning algorithms. Each model is validated on the historical generation of Tarbela Hydro Power Plant built on Indus River near Topi, KPK.. The chapter is concluded by results and their detailed discussion.

1.8.5 Chapter 6: Conclusion and Future work

Brief description of my thesis research work is presented in this section provided with tasks that can be carried out later for further research studies.

Chapter 2

Background

This chapter provides a brief overview of Renewable Energy Resources, its types and Hydro power potential in Pakistan.

2.1 Renewable Energy

Renewable resource is a resource that has the capability to reused repeatedly. In addition to that, it also has the capability to be replaced naturally. Renewable energy is produced using renewable natural resources such as sunlight, water, rain, tides and heat. In contrast to finite resources such as coal, oil and gas, renewable resources are easily available on every part of the land. It has the potential to address the energy crises in the world. In addition to that, it results in cheap and environmentally friendly energy generation. The use of renewable energy is growing day by day. Currently, there are at least 30 countries that are using renewable energy to generate at least 20% of the daily used electricity [12]. There are various renewable energy sources available such as hydropower, solar energy, geothermal energy and bio energy.

2.1.1 Types of Renewable Energy

The renewable energy sources and types are shown in the figure below:

Figure 5: Types of Renewable Energy Sources

2.1.1.1 Hydro Power

Hydropower plant uses water to make electricity. When flowing water is captured and turned into electricity, it is called hydroelectric power or hydropower. There are several types of hydroelectric facilities; they are powered by the kinetic energy which produces by flowing water to the downstream. Turbines convert the energy into electricity, which is then fed into the electrical grid to be used in homes, businesses, and by industry.

Figure 6: Hydropower Generation

2.1.1.2 Solar Energy

A solar power plant converts the sunlight into electricity, either using photovoltaics (PV) or concentrated solar power (CSP). Concentrated solar power systems contains mirrors, lenses and tracking systems which focuses on a large area of sunlight and convert it into a small beam. Photovoltaics converts light into electric current using the photoelectric effect.

These kinds of power plants have the following basic components:

- Solar panels that convert sunlight into useful electricity. It generates DC current with voltages up to 1500V.
- They are directly connected to an external power grid system and have some form of a monitoring system to manage the plant.

Figure 7: Solar Energy Farm

2.1.1.3 Geothermal Energy

The heat which produces by the earth is known as geothermal energy. Geothermal energy is widely used in two ways: to warn homes and other buildings or to generate electricity. The heat found below the earth's surface is used by the geothermal heat pumps which is then forwarded to the home and building using its pipe systems.

Figure 8: Geothermal Energy Plant

2.1.1.4 Bioenergy

Bio energy can be produced by the living organisms such as timber, agriculture wastes. Several methods have been proposed to generate the electricity from the biomass. First method is to simply burn biomass directly, heat water to steam, and sending it through a steam turbine, which then generates electricity. The second method requires gasification process. The gasifier takes dry biomass, such as agriculture waste, and with the absence of oxygen and high temperatures and processes it which is then transferred to gas turbine to generate the electricity.

Figure 9: Bio Energy

2.1.1.5 Wind Energy

Electricity can also be generated through wind power plants. Wind turbines usually comprise of three blades mounted to a tower. At 100 feet or more tower allows the turbine to take advantage of faster wind speeds. When the wind blows, it causes the rotor to turn which shafts in the nacelle – known as the box-like structure at the top of a wind turbine. A generator then converts the kinetic energy of the turning shaft into electrical energy.

Figure 10: Solar Energy Farm

2.2 Potential of Hydro Power in Pakistan

The world is moving towards the renewable energy resources due to its benefits. It has the potential to overcome the crises in Pakistan as well as in the world. The Pakistani government is trying to use all the effective renewable energy resources to fill the deficiency of the electricity. Pakistan is a water rich country and has a lot of potential to generate the hydropower. The hydropower potential of Pakistan was basically started with the Indus treaty however, no significance progress has been achieved so far due to the confusions and contradictions in investors. According to analysis, the forecasted installed capacity of different energy sources by the end of 2030 is shown in the figure below.

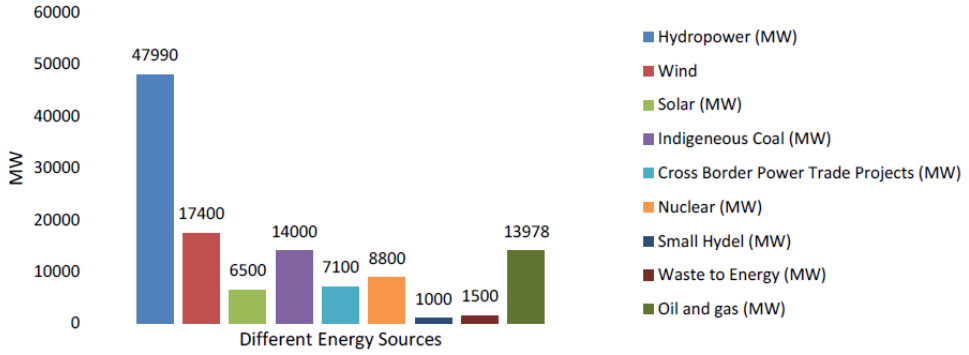

Figure 11: Forecasted Installed Capacity of Different Energy Sources by the End of 2030 [13]

From figure, the major percentage of electricity production will come from hydropower. Hydropower will be contributed 47,990 MW of energy as compared to wind (17,400 MW), coal (14,000 MW), oil and gas (13,978 MW).

The second advantage of hydropower is its pecuniary value as compared to other energy generating sources as shown in the figure below.

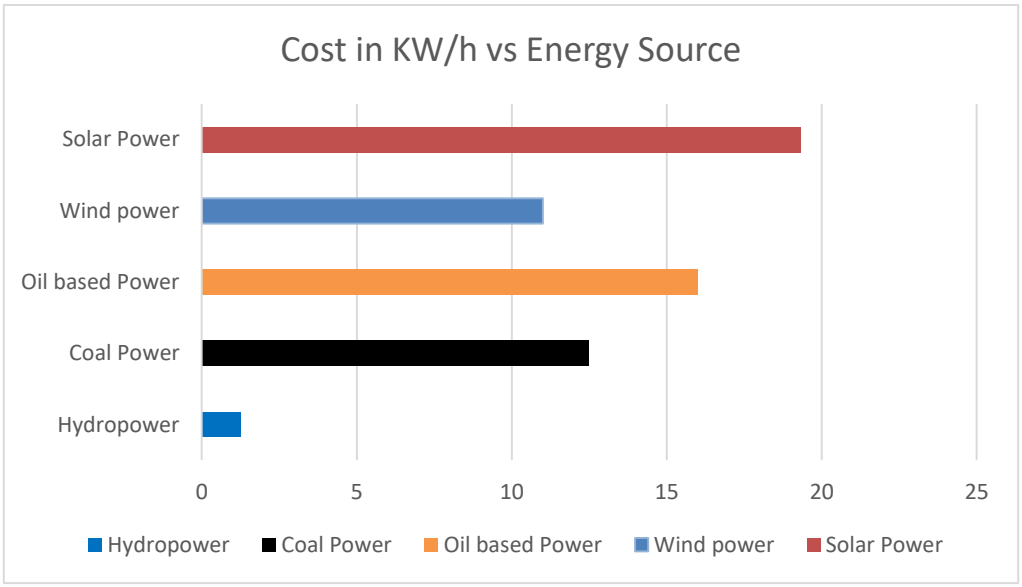

Figure 12: Cost of Renewable Energy Sources

2.2.1 Types of Hydropower Sources:

There are three major sources of hydropower generation Pakistan.

1) Large dams with the capacity greater than 50 MW.
2) Medium or small dams with the capacity lesser than 50 MW.
3) The large canal system with covering the area of nearly 60,000 km.

The available hydropower resources and projects in different provinces are shown in the figure below:

Summary of Hydropower Resources in Pakistan

Province/ Territory	Projects in Operation (MW)	Projects Under Implementation			Solicited Sites (Projects with Feasibility Study Completed) (MW)	Projects with Raw Sites (MW)	Total Hydropower Resources (MW)
		Public Sector (MW)	Private Sector (MW)				
			Province Level	Federal Level			
Khyber Pakhtunkhwa	3849	9482	28	2370	77	8930	**24736**
Gilgit-Baltistan	133	11876	40	-	534	8542	**21125**
Punjab	1699	720	308	720	3606	238	**7291**
Azad Jammu and Kashmir	1039	1231	92	3172	1	915	**6450**
Sindh	-	-	-	-	67	126	**193**
Balochistan	-	-	-	-	1		**1**
TOTAL	6720	23309	468	6262	4286	18751	59796

Figure 13: Hydropower Sources in Pakistan

2.3 Hydro Power Generation

The actual law of energy is; energy is neither created nor destroyed. In fact, it changes from one from to another. To produce hydropower, water in flowing form runs the turbine. It results in the generation of kinetic energy. The blades of turbine turn on the generator to produce the electric energy. There are major resources of generating electric power such as rivers, streams etc. but they are not reliable enough. For reliable supply of water, dams are highly needed. A general architecture of dam is shown in the figure below.

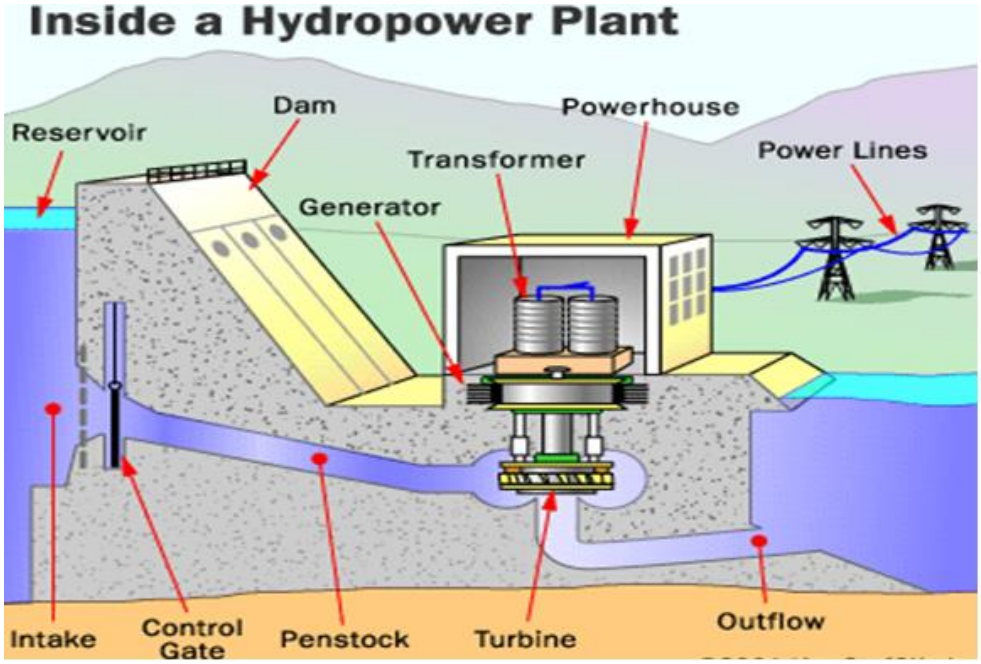

Figure 14: Overview of Hydropower Station

The dam is used to store water that serves various purposes such as domestic use, agricultural use and electricity usage. It consists of larger reservoirs of water that acts as battery with water to generate electricity when needed.

The dam consists of following parts:

- **Headwater:** The head of dam or reservoir from which the water flows.
- **Penstock:** It is also called the pipe that carries the water from head towards the generator.
- **Generator:** The fast-moving water from penstock rotates turbine blades enabling the motor of generator resulting in the generation of electric energy.

Tail Water and After bay: After the generation of electricity, water is called the tailed water that is released to after bay. It is then used for other domestic and agriculture purposes.

2.4 Machine Learning

Machine learning (ML) is a domain of artificial intelligence (AI) which compose available programs and enabling a computer to learn without being programmed [14]. These computer programs can be updated accordingly when new data needs to be incorporated. ML algorithms are generally categorized into three divisions namely supervised learning, unsupervised learning and reinforcement learning. The progression of machine learning and data mining is comparable as both domains explore or predicts the patterns from the data. On the other hand, in choice to extracting data for human knowledge as is the case in data mining applications; machine learning generates use of the data to identify patterns in data and fine-tune program actions.

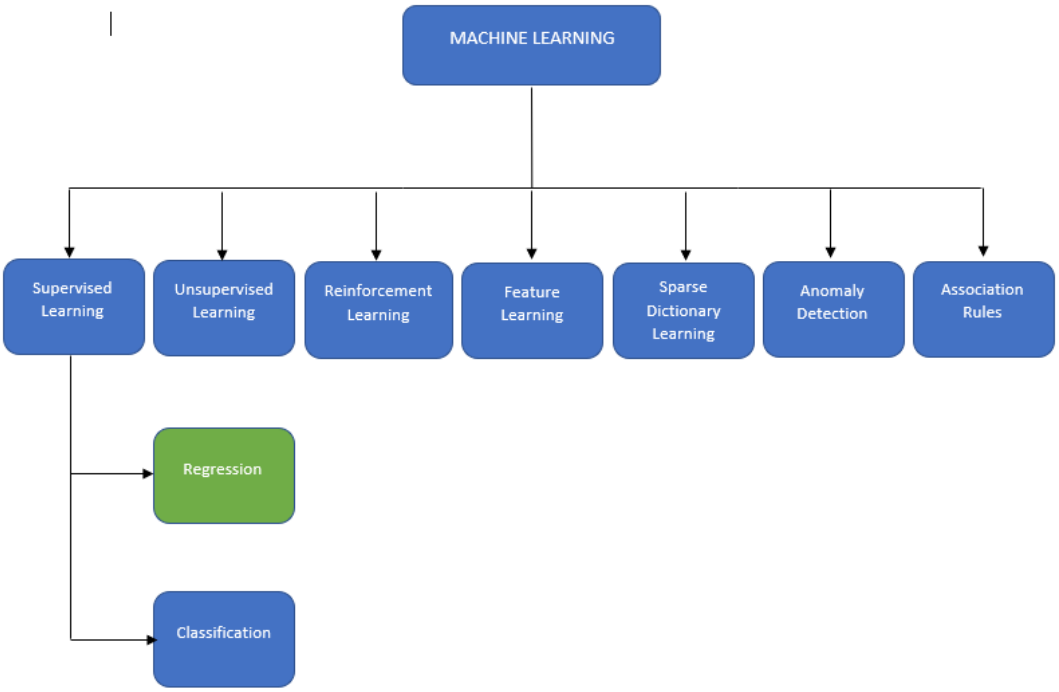

Figure 15: Types of Machine Learning

Supervised machine learning can be defined as, making a program to learn from labelled training data which has a set of training examples. In this learning approach, the training or input data which is defined as a vector quantity and having an output value or class. At first, a supervised learning algorithm learns from the practice or training data and generates a model, in order to map new examples. The supervised methods cover broad range of application areas such as finance, marketing, testing, manufacturing, stock market prediction etc.

Working of Supervised Machine Learning Algorithms consists of the following steps:

Step – 1: Create the training examples and define the data type according to the requirements

Step – 2: Train the computer to learn from data.

Step – 3: Evaluate the accuracy of the learned model by feeding the test data set that is break up from the training set.

2.5 Machine Learning Algorithms Selected for this Study

2.5.1 Multiple Linear Regression

Multiple linear regression is a technique that uses multiple explanatory variables to predict the results of the response variables. The main objective of multiple linear regression is to model the linear relationship between the independent (explanatory) variables and dependent (response) variable.

Multiple regression is based on assumption that there is no linear relation between independent and dependent variables. An example of multiple linear regression is explained below:

Consider a model with n predictor variables x1, x2,… xn and a response Y is written as,

$Y = \beta0 + \beta1*x1 + \beta2*x2 + \ldots\ldots\beta1*xn + €$ [15]

the € is considered as the residual terms of the model considered and the distribution assumption, we place on the residuals will allow us later to do inference on the remaining model parameters.

Examples:

Multiple regression can be used to predict exam performance of the students based on multiple explanatory variables such as lecture attendance, amount of study, anxiety and time taken to prepare for the exam. Similarly, you can predict the consumption of the cigarette daily based on variables such as age, type of smoker, smoking duration, gender etc.

Multiple regression lets us to identify the overall fit of the model and identify the involvement of each of the predictor to the total variance. For example, you need to identify how much the variation in daily consumption of the cigarette can be explained by factors age, type of smoker, smoking duration, gender as a whole but we can also find the relative contribution of each explanatory (independent) variable in explaining the variance [OL-1].

2.5.2 Support Vector Regression

Support Vector Machine is a type of supervised learning algorithm used for both classification and regression data problems. In SVM each data item is plotted in n-dimensional space (n being the number of features available), where value of each coordinate on the n-dimensional space represents the features of data set. After plotting the classification is performed on the data by discovering the hyperplane that splits the classes that have different features. Following figure shows the results of the SVM.

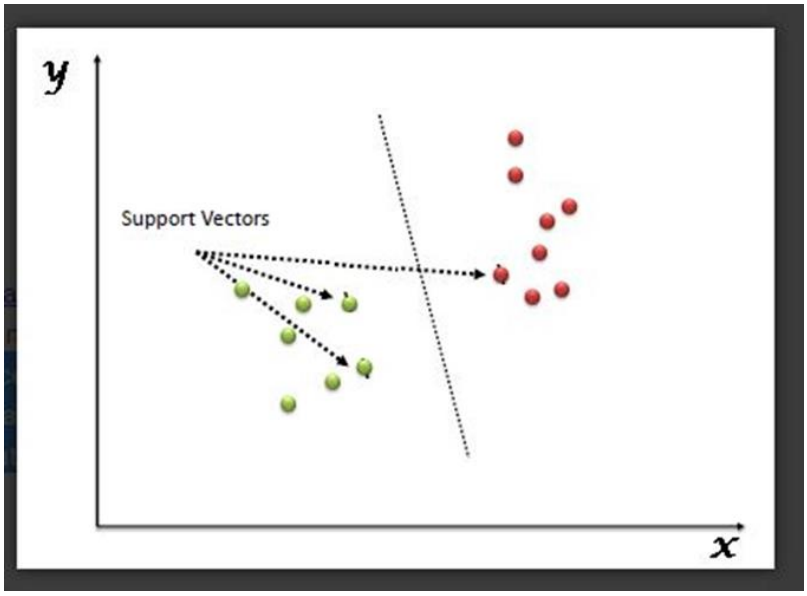

Figure 16: A Linear SVM Classification [OL-2]

SVM is effective when there are high dimensional spaces. It gives better results when the number of dimensions is greater than the number of samples. SVM is memory efficient because it uses a subset of training data when making decisions.

If the dataset is large it doesn't perform well because a lot of training time is required. If the data set has noise, the performance of SVM is not good.

2.5.2.1 Types of SVM:

Support vector machines can be used as classification machine, regression machine or for novelty discovery. The default setting of SVM is C-Classification or EPS-Regression but these may be overwritten.

EPS-Regression and NU-Regression are used when the data to be predicted is continuous.

If the data is not separable in linear space than the linear SVM boundary will not work properly. For example, in the data set given below the linear SVM cannot be used to divide the data n dimensional space.

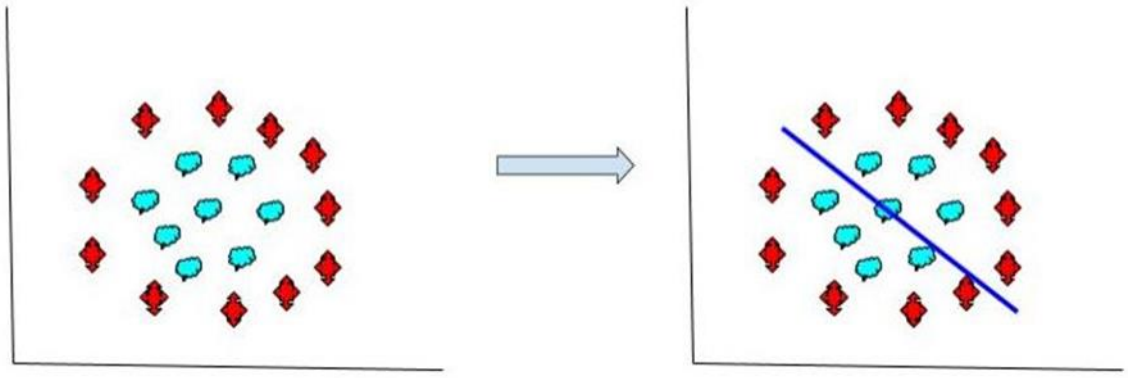

Figure 17: Example showing linear SVM boundry is not obtained on given data set [OL-3].

So, to overcome such problems a concept of kernel is used.

2.5.2.2 Kernel Function:

A kernel function is used to transform the training data from non-linear space to linear equation when there are higher number of dimensions such that the data is separable [16]. So, after getting the hyper plane in n+k dimensional space, we can get the boundary lines using kernels that separates the data set to gain the decision boundary.

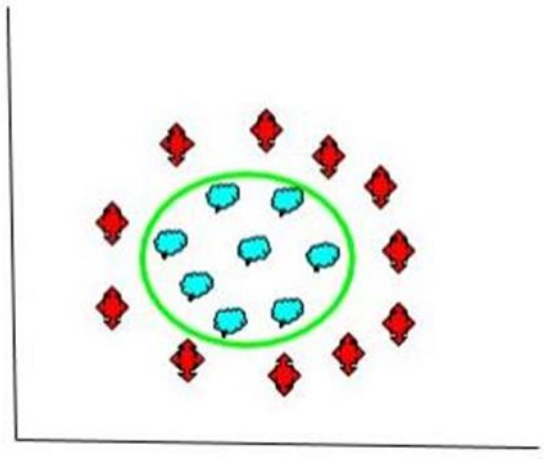

Figure 18: A Decision boundry gained after applying Kernal [OL-3]

Types of Kernel Functions

Linear Kernel:

Linear discriminant is used to provide 2 class classifiers, such that the available features can be divided by linear surface

Polynomial Kernel:

The Polynomial kernel is defined as a processor for generating new features from the existing features by applying the polynomial combination [OL-4].

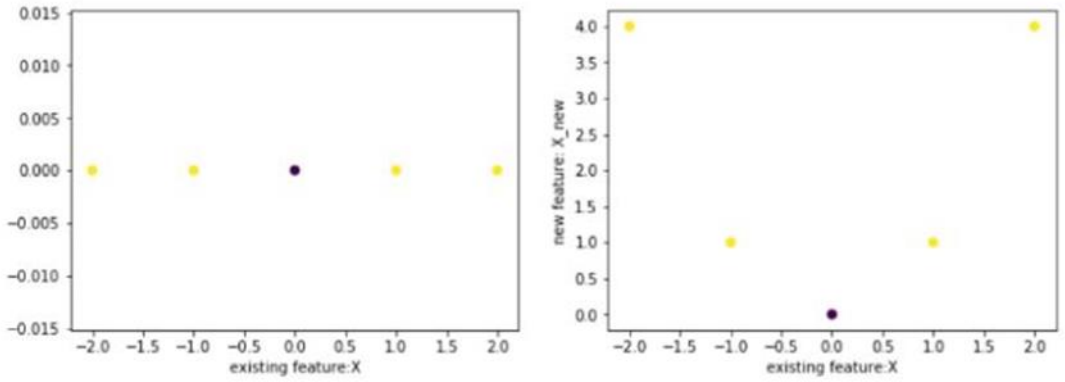

Figure 19: A) No Kernel Applied B) New features gained after applying Polynomial Kernal

Radial Basis Function (RBF) Kernel:

The radial basis function (RBF) kernel can be considered as processor that uses the distance between all other data points (dots) to some specific data points (dots) centers for generating new features. The widely used RBF kernel is known as Gaussian Radial Basis Function [OL-4].

So, in SVM if data cannot be separated linearly also known as soft margin, kernel trick can be used to identify the decision boundary for data sets that are not linearly separable.

2.5.3 Decision Tree

A Decision Tree is a supervised learning technique in which the target is already defined. It works for both continuous and discrete data. It is most widely used because they are easier to train and produce better results.

A decision tree is based on rules. It is used to classify the labelled training data and make set of rules. When we provide training set to the decision tree classifier some targets are set to be achieved. The Decision Tree identifies some rules to gain the

targets. These rules can be used on random test data to perform prediction and gain results [17].

A Decision Tree is a tree made by choosing a root variable based on Gini index, mutual information or gain ratio. The data is then divided into a tree and the process is repeated for every child. After the tree is build, pruning is done. Pruning is done to reduce the size of the decision tree.

A decision tree looks like a flow chart, in which each internal node represents a test done on an attribute, the branches represent the outcome of the test and the leaf nodes represents the classifier. A classifier is a decision made after calculating all the attributes of the tree. A simple example of decision tree is shown in Figure 19.

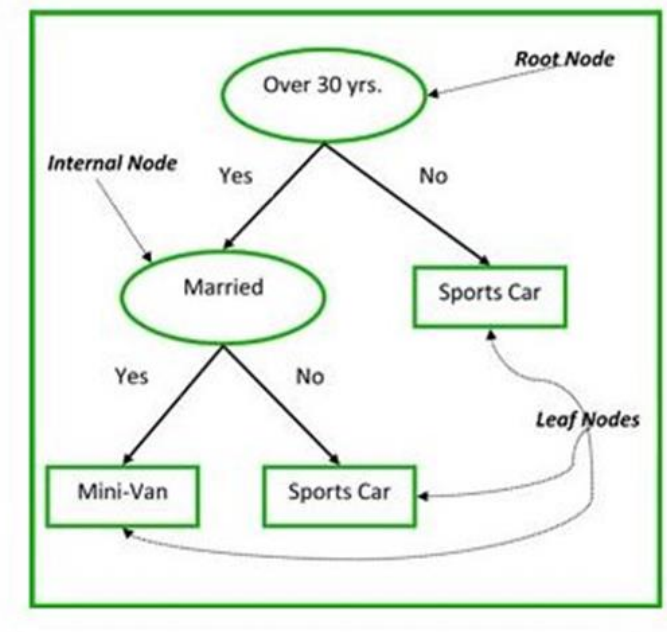

Figure 20: Working of Decision Tree [OL-5]

After the decision tree is made pruning is done to check overfitting of data and if there is any noise in the data. The major advantage of decision tree is that it is easy to understand, and the results can be interpreted easily. It works on both categorical and numerical values and is robust to outliers.

Uses:

Decision trees are considered popular because they achieve reasonable accuracy and their computation is cheaper. Most current classification algorithms such as C4.5 are based on ID3 classification decision tree algorithms [18].

2.5.4 Random Forest

The random forest classifier consists of multiple tree classifiers where each classifier is created using a random sample taken from the input vector independently, and every tree has a unit vote to identify the most popular class of an input vector. The output is determined by the majority votes of the trees [19].

The main idea behind the random tree forest is to combine the results of multiple decision trees into one single model. The predictions made by the individual decision tree may not provide accurate results but combining multiple trees will produce results closer to the real values. Higher the number of trees in the forest, higher chances of gaining more accurate results [OL-6].

Working of Random Forest:

Each tree in the random forest is grown following the steps mentioned below:

If there are N number of records in the training data, the records are sampled at random with replacement this is called bootstrap sample. If the variables available are M, then m<<M variables are selected at random, and the best split node is found from those m variables. The value of m is kept same during forest formation. There is no pruning in the trees and trees are formed to the largest size possible. The design factors of the random forest are the features available in the data set that are selected randomly, the number of trees to be formed and the number of samples to be considered as stopping criteria for the leaf node [20].

Random forest can work on large datasets, many input variables can be handled by random forest classifiers without removing any input variable. It can handle missing data by applying effective techniques and accuracy is maintained if large data set is missing.

The major difference between random forest and decision tree is that random forest doesn't only rely on predictions of decision trees, but the training samples used are also random when building trees and nodes are split on the base of random subsets of input features.

Chapter 3

Literature Review

This chapter explains the research performed in the field of hydro power systems around the World. Moreover, it contributes to understand development in the area of research.

3.1 Area of Research

This review has been conducted to identify studies related to hydro Power Plants and its components. Many water rich countries have already shifted their focus to hydro power production. Literature Review is divided into following sections,

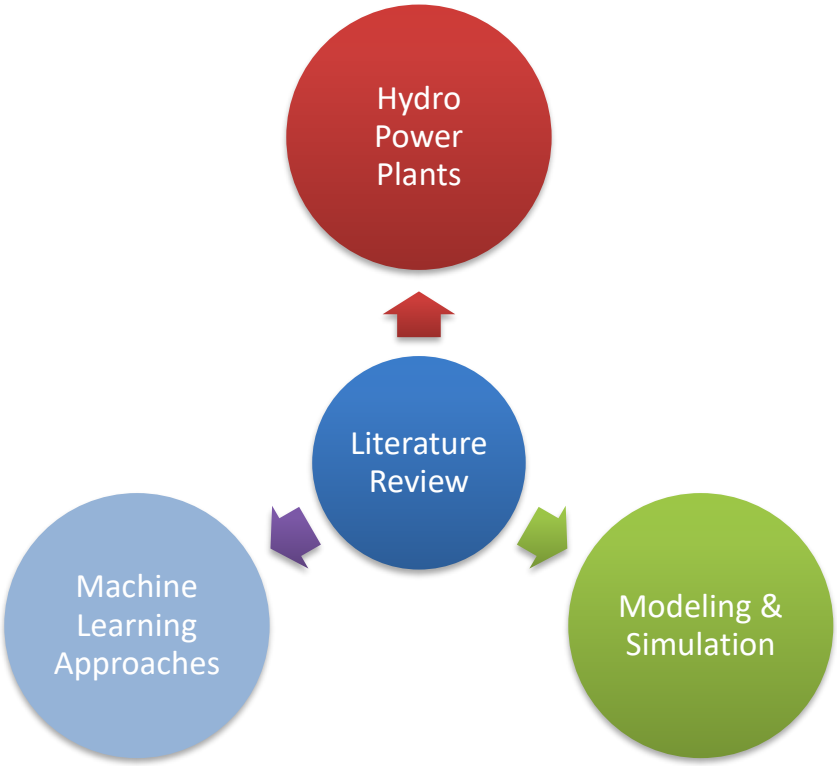

Figure 21: Areas of Research

3.1 Research on Hydro Power Plants Around the World

Table 2; summarizes the literature work in the sequence of research area shown in Figure 20 whereas details of these papers are described later.

Machine Learning Approaches		
Paper	**Title**	**Key Features**
M. Dehghani, Et. al 2019 [21]	Prediction of Hydropower Generation Using ***Grey Wolf Optimization Adaptive Neuro-Fuzzy Inference System***	GWO-ANIFS as a couple showed promising predictions while ANFIS independently failed in multiple scenarios/rules. ANFIS takes long training time to adjust values to parameters of membership functions where GWO is used to optimize the overhead.
Alrayess H. Et. al 2018 [22]	Using ***Machine Learning Techniques*** and ***Deep Learning*** in Forecasting the Hydroelectric Power Generation in Almus Dam, Turkey	Studied and presented the difference and quality of results of ANN, SVM and Deep Learning models.
I. E. Ebukanson Et. al 2017 [23]	Statistical Analysis of Electricity Generation in Nigeria Using ***Multiple Linear Regression Model*** and ***Box-Jenkins Autoregressive Model of Order 1***	Prediction accuracy of both the models was compared on the same set of data. Examined bivariate relationship of climate variables and electric power generation.
A. T. Hammid Et. al 2016 [24]	Prediction of Small Hydropower Plant Power Production in Himreen Lake Dam Using ***Artifical Neural Network***	Discovered non-linear behavior of Small Hydropower Plant energy production. Analyzed variations in power production in mixed situations.

G. Li Et. al 2015 [25]	Applying a **Correlation Analysis Method** to Long-Term Forecasting of Power Production at Small Hydropower Plants	
G. Li Et. al 2014 [26]	Short-term Power Generation Energy Forecasting Model for Small Hydropower Stations Using ***GA-SVM***	Mastered short-term power generation energy of Small Hydropower Plants to understand the generation trend and better utilize resources. Used hybrid technique of GA with SVM to optimize the model parameters
E. Uzlu Et. al 2014 [27]	Estimates of Hydroelectric Generation using ***Neural Networks*** with the ***Artificial Bee Colony*** algorithm for Turkey	ANN with ABC predicted the hydropower generation more accurately than classical ANN.
F. Olsson Et. al 2005 [28]	Modeling the Total Inflow Energy to Hydropower Plants – A Study of Sweden and Norway	Comparison study of MLR and ANN was studied to choose optimal model for Total Inflow.

Table 1: Literature Review in the Area of Machine Learning

In this paper M. Dehghani, Et. al 2019 [21] combined Grey Wolf Optimization (GWO) with the adaptive neuro-fuzzy inference system (ANFIS) based novel model is presented to predict the hydropower generation. Twenty different (independent and dependent) parameters such as water inflow in the dam, average rainfall, hydropower plant last month power production, etc. have been selected as input whereas for output just one parameter, Hydropower plant energy generation, was selected. Moreover, 53 years old Dez basin Dam, Iran data has been used. This integrated GWO-ANFIS model performed in all 20 input while ANFIS fails to perform on 9 out of 20 input parameters. The proposed GWO-ANFIS works better than ANFIS and predicts the hydropower generation with higher accuracy in comparison to the ANFIS, which make it suitable and worth-able for policymaking and industrial use.

In this paper, Alrayess H. Et. al 2018 [22] use three different models, Artificial Neural Network (ANN), Support Vector Machine (SVM) and Deep Learning (DL), to predict the power generation of Almus Dam, Almus Baraji, Turkey. Three different variables such as water inbound, water influx and Lake water levels has been selected as an input while the Almus Hydropower plant electricity production has been opted as an output.

The correlation coefficient for ANN is 0.76, SVM is068 and DL is 0.99 respectively while square correlation of ANN is 0.58, SVM has 0.46 and DL has 0.99 respectively. The RMSE of ANN is results shows that DL works better than ANN and SVM. An in-depth analysis has been done using ANN, SVM and DL models, and it has been found that DL perform better. DL has a RMSE of 0.071 while ANN has a RMSE value of 0.675 and SVM has 1.00 respectively.

This paper discusses detailed statistical analysis of power production in Nigeria using two different models, named Multiple Linear Regression (MLR) and Box Jenkins Autoregressive model of Order 1 (BJAO), has been concluded. I. E. Ebukanson Et. al 2017 [23] selected two independent environmental variables, ambient temperature and rain, were opted. Furthermore, the data of Nigeria electric power production acquired from the Central Bank of Nigeria Statistical Bulletin whereas rain and ambient temperature were obtained from the National Bureau of Statistics (NBS).

The relation between power production and environmental variables has been established and MLR and BJAO has been compared. The results show that rain has an important role with power production; it has correlation of 0.92 with power production. Moreover, ambient temperature has not a pivot role in power production as it only had a relation of 0.13 with power generation. During the comparison it has been noticed that MLR is more rigorous, robust and it outperform BJAO with RMSE of 60.27% and constant of determination of 99.77%.

A. T. Hammid Et. al 2016 [24] proposed an efficient performance predictive framework for a small hydropower plant at the Himreen lake dam by employing Artificial Neural Network (ANN) integrated with Single Layer Perception (SLP) with a feed-forward Backpropagation (BP) algorithm. Moreover, the proposed framework designed by using real data set collected daily from the dam from 2005 to 2015. ANN predicts and improves the performance of the hydropower plant in association with predicted variables and real output is higher than 0.96 with an RMSE 0.0032734 and gradient of epochs 0.000136 respectively.

G. Li Et. al 2015 [25] analyzed spatiotemporal parameters of SHP and a correlation between Large-Medium Hydropower Plant (LHP) and Small Hydropower Plant (SHP) has been proposed. SHP in Yunnan province, China were taken as an example and the correlation between power generation of SHP and LHP (Dachaoshan plant) interval inflow has been designed and verified. By observing the inflow at Dachaoshan plant, the power generation at Puer District, Yunnan, China is predicted. The proposed scheme accuracy has been verified by analyzing the prediction results of 13 districts in Yunnan. The proposed scheme gives us 87.9% prediction accuracy with 18.5% of RMSE and the maximum RMSE is 43.8% and minimum RMSE is 5.6% respectively. Hence, it has been concluded that the prediction accuracy of power generation at SHP is depended upon the LHP.

Authors G. Li Et. al 2014 [26] presented a Genetic Algorithm (GA) based support vector machine (SVM) model to predict the power generation of a small hydropower plant (SHP). SVM is opted because of its robustness, nonlinear and efficient pattern recognition features. An extensive analysis carried out to see the potential and

predictive accuracy of GA-SVM. Hence, the historical data of SHP in Yunlong and Maguan County in Yunnan, China is selected to investigate the potential of GA-SVM. To better understand the potential of GA-SVM, Auto-Regressive Moving Average (ARMA) model is adopted for comparative analysis. GA-SVM outdo ARMA in the calibration period with two measurements. The proposed model outperforms ARMA with 0.24 % in RMSE and 0.41% MAPE value. The results clearly showed that GA-SVM performed better than ARMA model.

E. Uzlu Et. al 2014 [27] studied coupling Artificial Neural Network (ANN) with Artificial Bee Colony (ABC) algorithm for prediction and forecasting of turkey annual hydroelectric energy production. Multiple independent parameters such as Gross Electricity Demand (GED), Residents and Energy Consumption are selected to estimate the accuracy of ANN-ABC model. Hydroelectric production in turkey from 1980 to 2021 was a model based on the parameters. ANN model trained with Backpropagation algorithm opts for comparison with ANN-ABC. RMSE and Relative Error (RE) selected as parameters to evaluate the model accuracy. Average RE for ANN-ABC was 4.599 while average RE for ANN with Backpropagation algorithm was 5.202. Furthermore, ANN-ABC is more robust and has better accuracy than ANN with Backpropagation algorithm.

In this paper, F. Olsson Et. al 2005 [28] carried out an extensive study to find the best model based on Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) for Sweden and Norway hydropower plants. The selected model simulated the total week input in the hydropower plants by using daily data from multiple stations as an input. After close analysis of both models by using the daily hydropower plant stations inflow data provided by Swedish Meteorological and Hydrological Institute (SMHI) and Norwegian Water Resources and Energy Directorate (NVE), MLR has been adopted. Furthermore, the selected model was then integrated into a user-friendly software, named TWh-simulator, which is operational in Sweden. When simulated for validation, MLR remains consistent with RMSE of 0.22 and correlation coefficient of 0.98 respectively.

Different reservoir operation models have been developed and applied for planning studies to formulate and evaluate hydro power generation.

Modeling & Simulation		
Paper	**Title**	**Key Features**
R. Teegavarapu Et. al 2014 [29]	Simulation of Multiple Hydropower Reservoir Operations Using System Dynamics Approach	Monte Carlo simulation is used to assess the variability of power generation due to changing system conditions.
A. Sharifi Et. Al 2012 [30]	A System Dynamics Approach for Hydropower Generation Assessment in Developing Watersheds: A Case Study of Karkeh River Basin, Iran	Studied hydrological impacts of watersheds and their effects on hydroelectricity generation of existing and projected hydropower plants.
T. G. Bosona Et. al 2010 [31]	Modeling Hydropower Plant System to Improve its Reservoir Operations	Analyzed increase in yearly energy production and uniformity of monthly energy production by controlling reservoir release.

Table 2: Literature Review in the Area of Modeling and Simulation

In this paper Ramesh and Slobodan in 2014 [29] simulated a multiple reservoir system using STELLA simulation environment. The real time application of model is done to four reservoirs located at Winnipeg River in the Manitoba province in Canada. These reservoirs built a part of much more complex network of hydropower reservoirs maintained by the local hydropower corporation, Manitoba Hydro. Different built-in mathematical, logical and statistical functions were used along with java built-in functions like RANDOM, IF-THEN-ELSE and DELAY. Sensitivity analysis and graphical inputs were also part of the model. Major components include system configuration, hydraulic coupling and decision process. Future improvements in the model should focus on accurate modeling of flow transport delays and hydraulic coupling.

A. Sharifi et al. 2012 [30] presents a system dynamic model using Vensim software to assess hydropower generation in developing watersheds which are located at Karkheh

River basin in Iran. The goal of this simulation is to analyze the production of newly hydropower units meet the designated energy production requirements after the watershed is fully developed and to assess the hydropower generation and expansion opportunities in future. The paper consists of three sectors i-e irrigation dams in which water storage and release behavior of non-hydropower dams were simulated; hydropower dams which were single-reservoir reliability-based simulation model was used in which dam operation and energy production of hydropower units were simulated; lastly, control hydrometric stations in which river discharges were modified at a location based on upstream irrigation and hydropower dam storage release behavior. Simulation results revealed that an average 88 GWh=year increase in electricity production can be achieved per 100 × 106 m3 of annual environmental flow release out of transferred water from Sirvan to Karkheh River basin.

Melka Wakana Hydropower Plant in Ethiopia (T. G. Bosona, 2010 [31]), modelled and simulated by Swedish University of Agricultural Sciences Sweden. In this paper, different simulations were carried out, using System Dynamics approach in Powersim software, to increase the efficiency of utilization of dam reservoirs by changing the values of initial reservoir storage and acceptable reservoir release for power generation thus maximized yearly energy output with improved uniformity of energy production is obtained. The results of simulation analysis indicated that 5.67% yearly power generation was increased while 38.33% evaporation loss was reduced. Moreover, uniformity of monthly power generation was also improved but the power plant still produced below monthly energy production.

Different researches have been done to improve hydro power generation in Pakistan. Analytical tools can help the decision and policy makers to formulate valuable policies in order to improve the power generation by reducing water losses from reservoir. In this proposed project, an analytical tool will be developed that will model the hydropower plant using Machine Learning approach modeling the power generation visually using bygone data. This will help in analyzing the current power production and will forecast the future production based on hydro resources in Pakistan. In this thesis, we used machine learning techniques to model and forecast the Tarbela Power Plant generation.

Chapter 4

Methodology and Results

This chapter explains the steps of Machine Learning Approach to develop proposed model of Taberla Power Plant.

4.1 Data Analysis

4.1.1 Case Study: Tarbela Power Plant

One of main sources of hydro power in Pakistan is Tarbela Dam which is located on the Indus River about 60 km northwest of Islamabad. It is one of the largest earths filled dam in the world used for power generation and irrigation purposes. Top of FormTarbela dam was built in 1970's to control water flow of the upper Indus for irrigation of the fields downstream and for generating hydro power. Indeed, even since today it is the main water reservoir on the Indus. Besides, its irrigation system discharges more than 6.4 MAF it creates up to 3478 MW [32] of electricity and contributes to 32% of Pakistan's energy needs. It has two spillways i.e., the service spillways and the auxiliary spillways and six tunnels. service spillways have 7 Gates; auxiliary spillway has 9 Gates. There are three tunnels in the right projection are utilized for water system and electricity generation. Rest are used for irrigation purposes only.

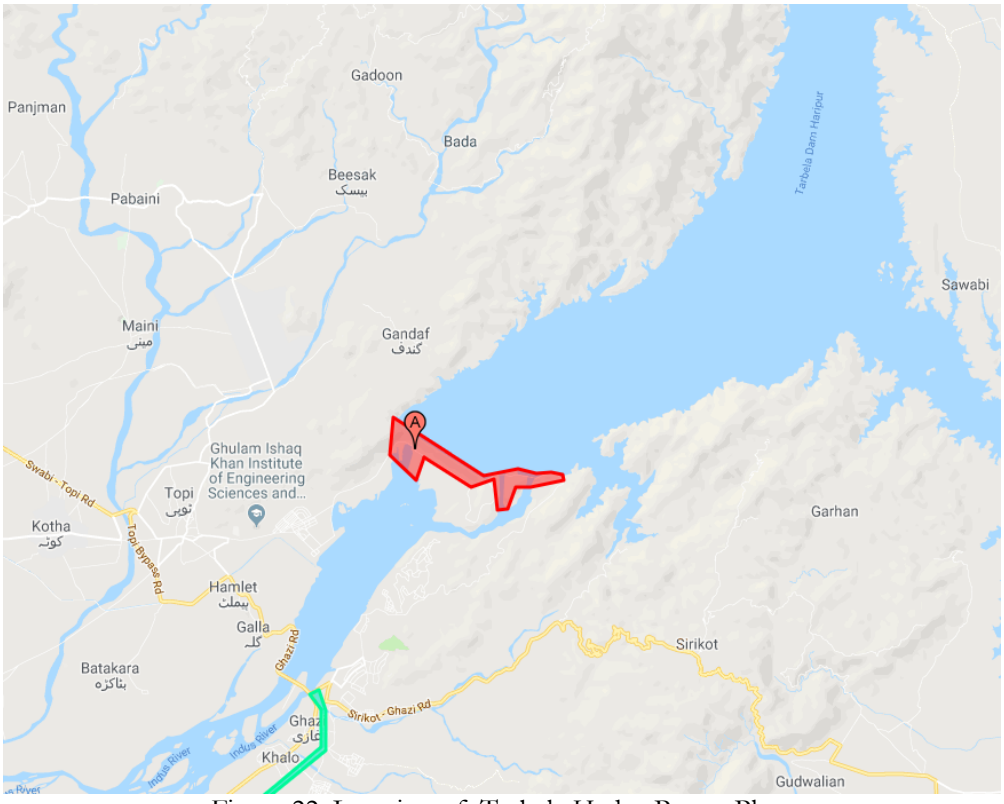

Figure 22: Location of Tarbela Hydro Power Plant

The comparison of energy generation of Tarbela dam with other projects is shown in the figure below.

Existing Hydropower projects in operation in Pakistan

S. No	Project Name	Location	Province	Capacity (MW)
A. WAPDA				
1	Tarbela	Indus River	Khyber Pakhtunkhwa	3478
2	Warsak	Kabul River, Peshawar	Khyber Pakhtunkhwa	240
3	Jaban (Malakand-I)	Swat River, Malakand	Khyber Pakhtunkhwa	20
4	Dargai (Malakand-II)	Swat River, Malakand	Khyber Pakhtunkhwa	20
5	Kurram Garhi	Kurram Garhi (Canal)	Khyber Pakhtunkhwa	4
6	Mangla	Jehlum River, Mirpur	AJ&K	1000
7	Ghazi Barotha	Indus River, Attack	Punjab	1450
8	Chashma	Indus River, Chashma	Punjab	184
9	Rasul	Chenab River, Rasul	Punjab	22
10	Shadiwal	Gujrat	Punjab	14
11	Nandipur	Upper Jehlum Canal, Gujranwala	Punjab	14
12	Chichoki Hydel	Upper Jehlum Canal, Sheikhupura	Punjab	13
13	PAEC Chashma Hydel	Chashma, Mianwali	Punjab	1.2
14	Renala	Lowerbari Doab Canal, Okara	Punjab	1
15	Satpara	Satpara River, Sakardu	Gilgit-Baltistan	16
16	Kar Gah Phase VI	Gilgit	Gilgit-Baltistan	4
			Sub Total	6481

Figure 23: Capacity of Hydropower Projects in Pakistan

Table 4 provides output power at each turbine.

Units (Turbines)	Output Power (MW)
1	175
2	175
3	175
4	175
5	175
6	175
7	175
8	175
9	175
10	175
11	432
12	432

13	**432**
14	**432**

Table 3: Output Power of Installed Turbines

4.1.2 Data Set

We first visited Tarbela Power Plant in August 2016 and had a detailed discussion about their process and data variables. They provided us a dataset of 23 years at a daily resolution with the following parameters:

	HYDROLOGICAL DATA FOR THE YEAR 1993-2016						
	TARBELA POWER STATION						
DATE	HRL	IRRIGATION INDENT (CFS)	INFLOW	Discharge		TOTAL	Generation (Million KWH)
			CFS	(1~10)	(11~ 14)	OUTFLOW	

Figure 24: Structure of Energy Generation Data Set

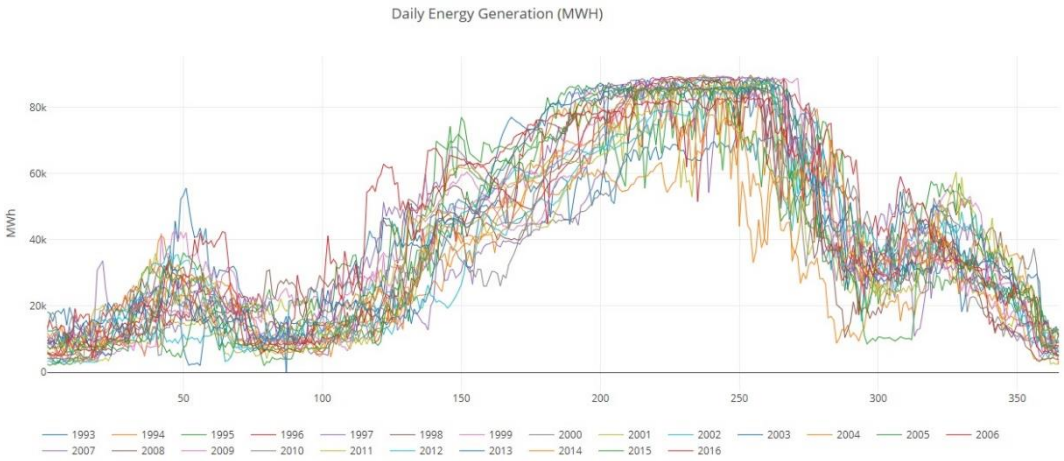

Figure 25: Daily Energy Generation of Tarbela Power Plant 1993-2016

Data related to inflow at rivers, evaporation and rainfall data at different dams are gathered from different resources like NESPAK and Hydra Consulting services. The following graph shows the average evaporation at Tarbela Dam

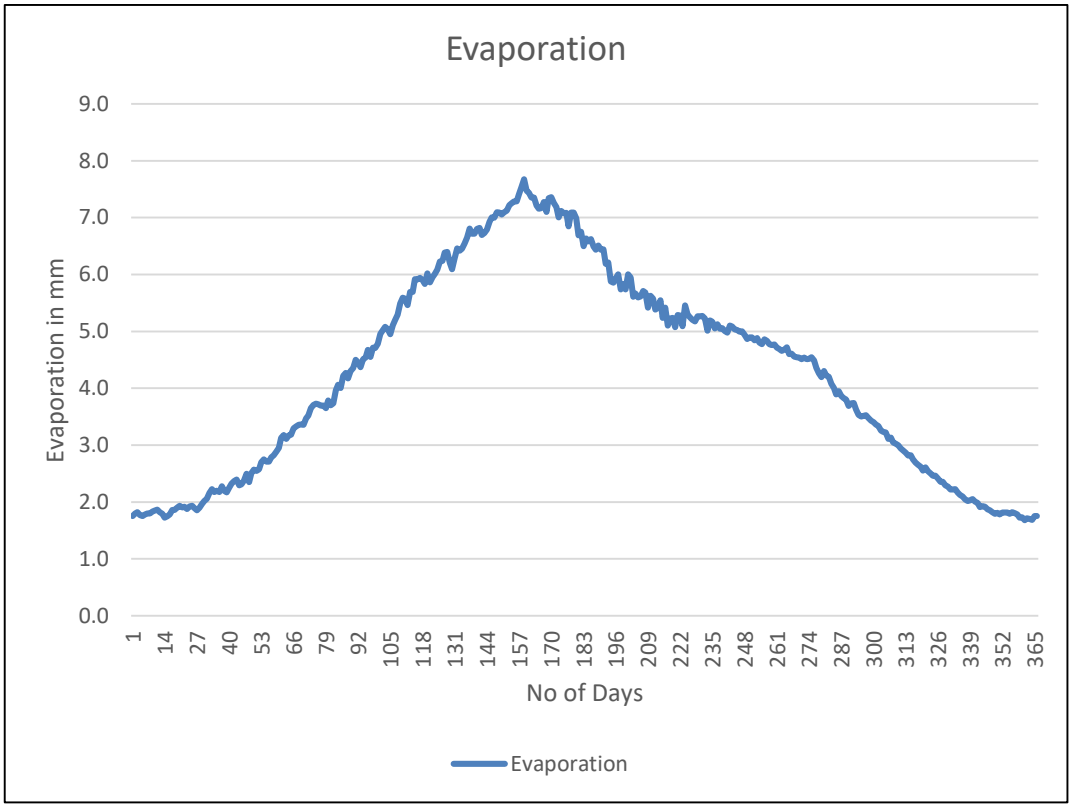

Figure 26: Evaporation Rate in Tarbela

The daily inflow of River Indus at Tarbela is plotted as below.

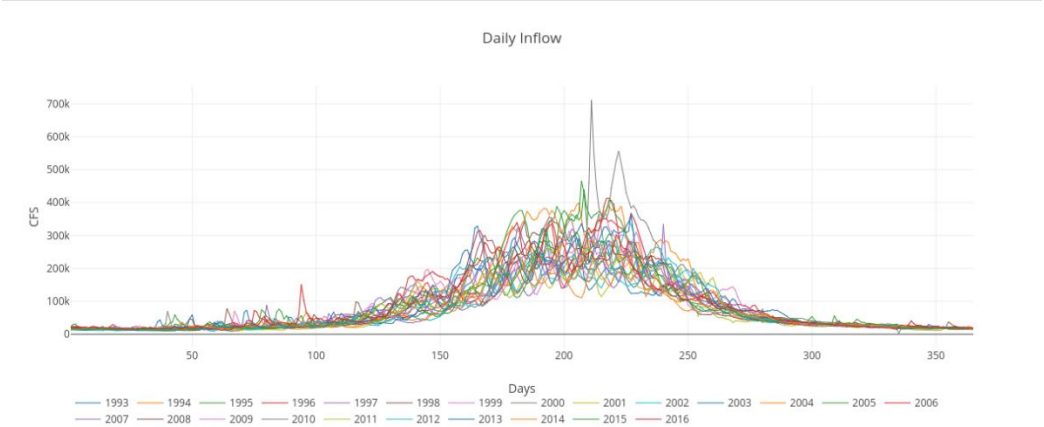

Figure 27: Daily Inflow of River Indus at Tarbela

The following plot shows the average precipitation at Tarbela Dam

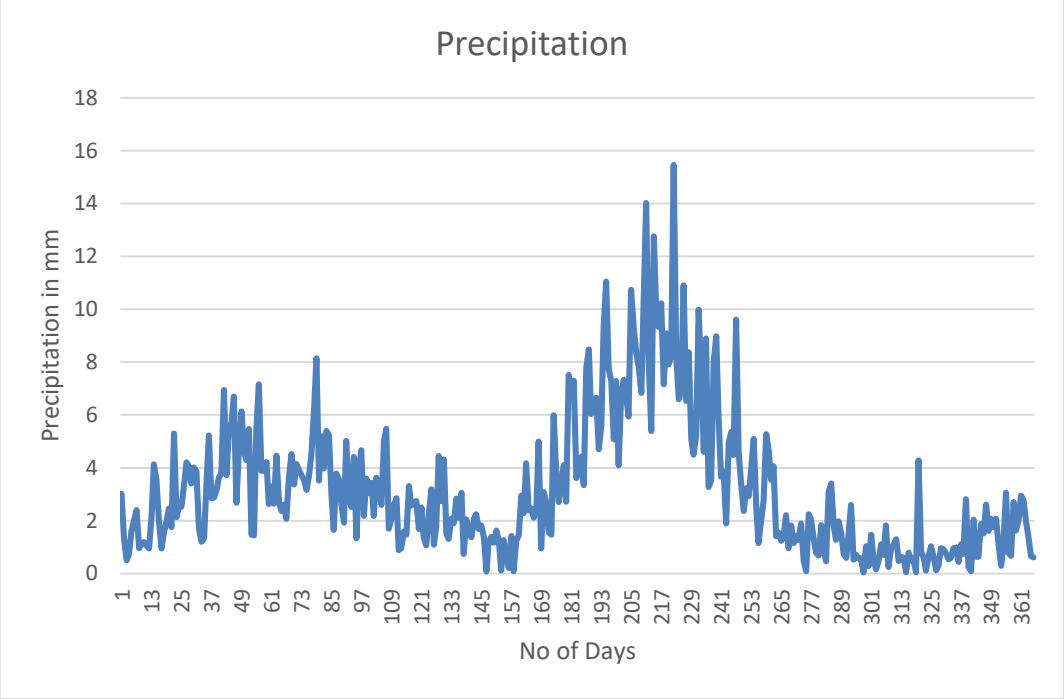

Figure 28: Daily Precipitation of River Indus at Tarbela

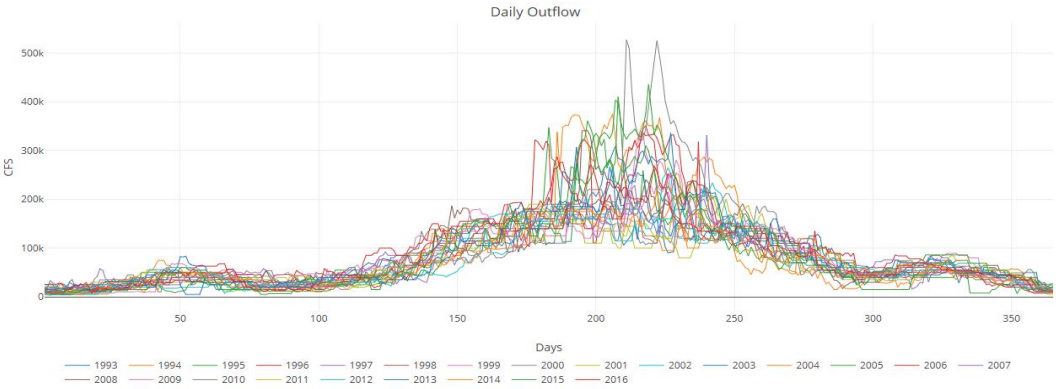

Figure 29: Daily inflow of River Indus at Tarbela

This collected data was for the use of development of a hierarchical and multi-scale system dynamics model based on composition of sub-models in a hierarchical order. The purpose was to develop a hybrid model where System Dynamics and Agent based approach were integrated. At highest abstraction level (Macro-level), system dynamics dealt with aggregates. Since it works on continuous processes, they used System

dynamics in Tarbela hydro power plant to get simulated results. While Agent based modeling is known for micro level abstraction, a part of the agent behavior is defined as a state chart represented in the environment model.

We re-visited Tarbela Power Plant in March 2019 and further discussed each data variable and its significance in power generation to apply Machine Learning approaches in predicting energy generation. We were provided with the latest dataset of the last 5 years recorded at a daily resolution in the same format with an addition of few parameters:

	HYDROLOGICAL DATA FOR THE YEAR 2013 - 2017												
	TARBELA POWER STATION												
	H.R.L	T.R.L	G.HEAD	Irrigation	INFLOW	OUTFLOW THROUGH (CFS)							TOTAL
DATE	SPD	SPD	FT	Indent (CFS)	CFS	Units(1-10)	Units(11-14)	P/H (1-14)	Auxiliary Spillway	Service Spillway	T-4	T-5	OUTFLOW (CFS)

Figure 30: Structure of Energy Generation Data Set

4.1.3 Pre-processing

Data is raw form of text. It needs to be processed, structured in order to extract connections and transform into a meaningful information. Discussion with authorities of Tarbela Power Plant assisted in understanding the meaning of each recorded variable. It is explained as below,

Head Reservoir Level (H.R.L): It is measure of water level in reservoir recorded daily.

Tail Reservoir Level (T.R.L): It is the measure of water level located exactly downstream from a dam.

Gross Head (G. Head) is a function of input and output flow of water through a hydraulic structure. It is calculated as difference of H.R.L and T.R.L and is a vertical height measured in feet. The higher is the values of head, the more is the pressure gain resulting in more power.

Irrigation: It is the amount of water released through the system every day for the purpose of irrigation. Indus River System Authority (IRSA), an authority in Pakistan responsible for supervising activities of Indus River directs Tarbela Power Plant the amount of water to release daily. This is the water quantity used by Powerhouse to generate electricity.

Inflow: It is the amount of water entered the reservoir from the discharge of an upstream source i.e Indus River. It is calculated using different statistical methods, one of the common being mechanical current meter method. It is measured in CFS (Cubic Feet per Second).

Outflow: It is the discharge of water through a hydro system and/or spillway. The amount of water released to downstream flows is calculated daily through each output and total is computed.

Parameters	Units
Date	DD/MM/YYYY
H.R.L	Survey of Pakistan Datum
T.R.L	Survey of Pakistan Datum
G. Head	Feet
Irrigation	Cubic feet/second
Inflow	Cubic feet/second
Outflow (Tubines/Spillway)	Cubic feet/second

Table 4: Parameters

Once we have knowledge of all the parameters of the dataset, we filtered out the columns needed for analysis of our problem study. Initially theses were as follows,

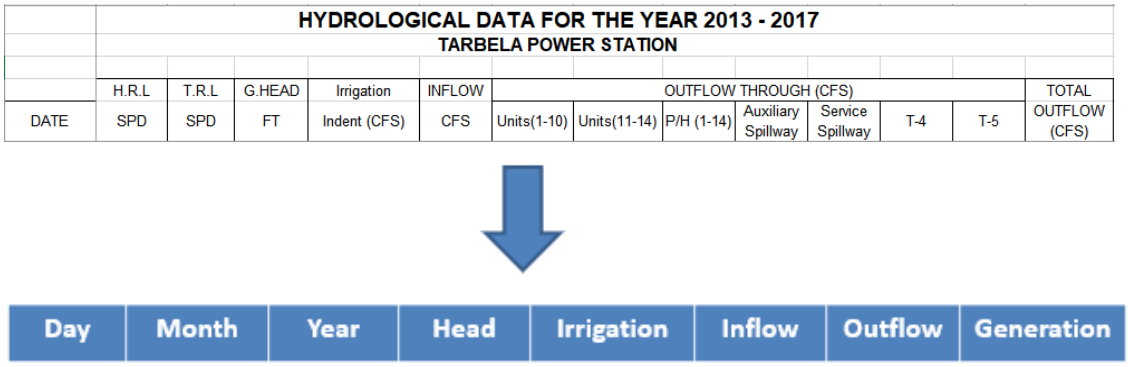

Figure 31: Features of Tarbela Power Plant

Later, expert from industry advised to use the following key columns relevant for the study of forecasting energy generation. It consisted of Date, Head level, discharge of water from Turbines 1-10 denoted by D1, the discharge of water from Turbines 11-14 represented by D2 and Generation which is the energy generated in KWh.

Once applicable features were finalized, individual columns needed to be pre-processed in the eligible format.

Date was split into Day, Month and Year attributes. We selected five years of historical data to study the trends in energy production out of which 4 years were selected for training each model and 1 year of data was used to test its accuracy. These were not limited for which reason Year was removed from the final set of input features.

Machine learning experts showed concern of using Day and Month as categorical values. The connection of two numerical values, say Day 1 with Day 5 or Month 2 with Month 7 does not relate a link. It was recommended to use One-Hot Encoding or also knows as 1-of-K method to convert it into binary values. This would require adding dummy variables in order to categorize each tuple.

Day having a maximum of 31 values was split into 31 columns, starting Day1 to Day31 while Month was replaced with 12 columns, Month 1 to Month 12. For each row, binary values were introduced for columns Day1 to Month12. For example, data recorded on January 01 will have 1 in Day1 and Month1 while all the remaining dummy variables will have 0 as its value. This ensured uniformity and supported in scalability for algorithms that requires data to be in a specific range.

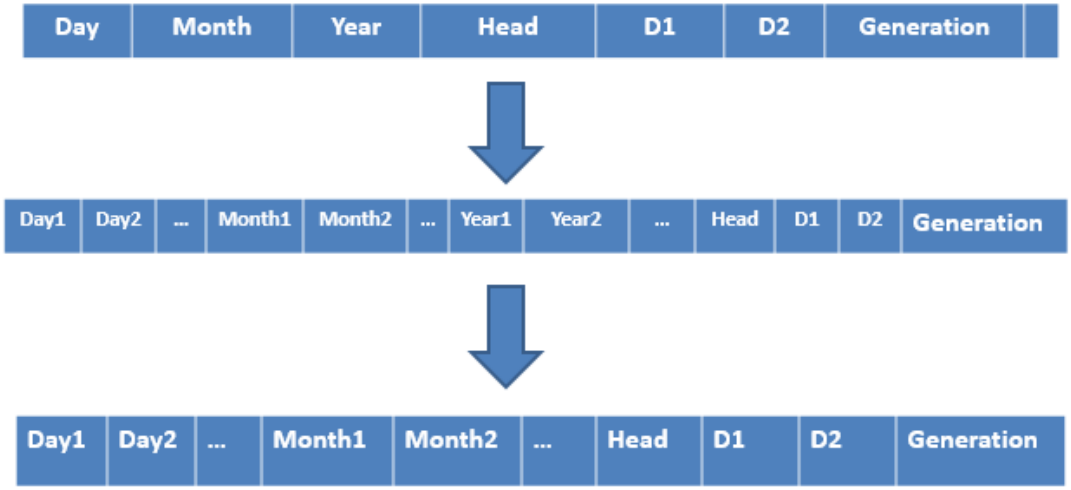

Figure 32: Features Extracted for Energy Forecasting Model

4.1.4 Training & Test Sets

The data gathered in previous visits showed anomalies and discrepancies. The officials of Tarbela Power Plant were kind enough to provide us with latest data of energy produced by the plant. This included recorded data from year 2013 – 2017.

Usually dataset is used as a whole and is divided into training and test sets during run time. These portions are randomly selected based on the ratio defined in the tool. However, we chose to divide the data into two independent sets. The training set has recordings of years 2013 – 2016 and testing of each model was performed data of Year 2017. This technique for dataset division was selected to avoid overfitting issues and increase model's robustness while achieving goal of high performance.

4.2 Model Development

4.2.1 Feature Extraction

Feature Selection step is the basis for Machine Learning. It is a data driven technique in which non-related variables can lead to meaningless results. Aim of performing machine learning is to achieve substantial accuracy in comparison to real-world data and gain insights to better idealize situations in future. Algorithms made for the purpose can better train to understand and study the correlation only if input features have association in predicting the dependent variable which is the essence of machine learning.

The key is to efficiently choose features that shows great insights with the predicted variable. When solving a real-world problem there could be use cases where data may have multiple inputs having no or little connection with the output which in turn can lead to incorrect and biased results. The concept of this step is to study the relation and best select features showing strong relationship.

A few advantages of using this technique are,

- Training the model is quicker and easy
- Less computation complexity involved
- It perfectionates and enhances productivity
- Reduced *under-fitting* and *over-fitting,* concepts used when either model predicts worse or extensively better against actual data which is an ideal result.

There are numerous approaches and techniques available for this step. A few are illustrated below.

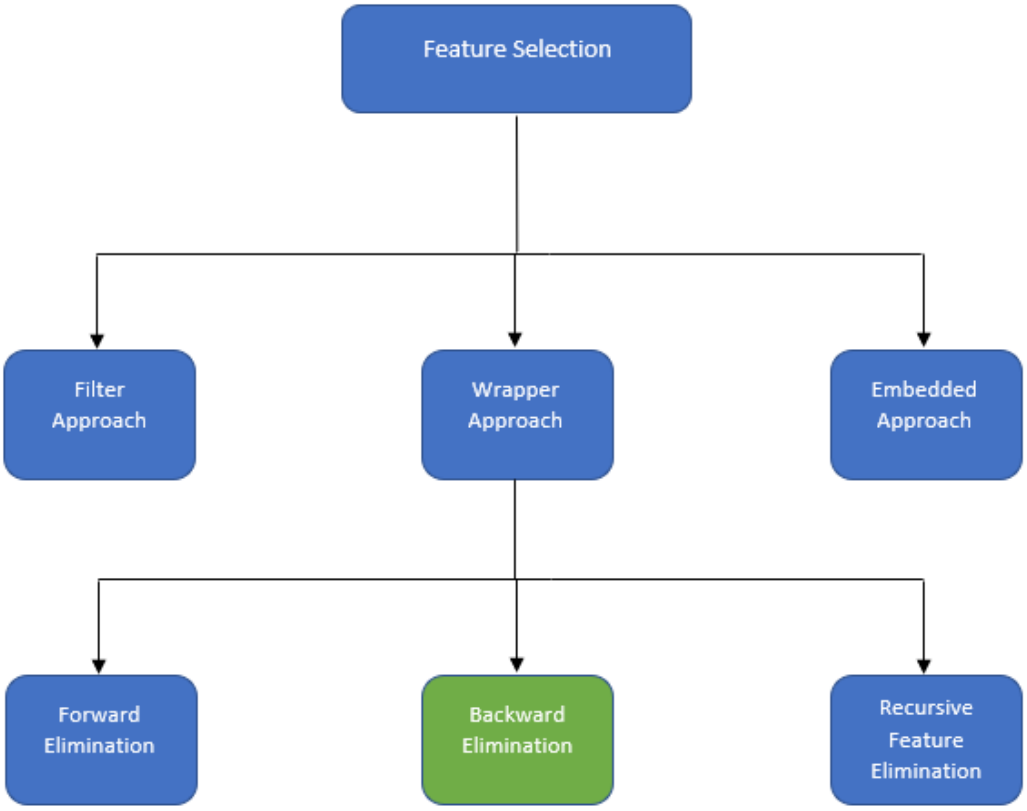

Figure 33: Types of Feature Selection Approaches

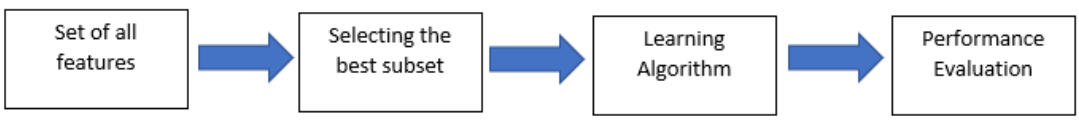

Figure 34: Steps of Filter Method

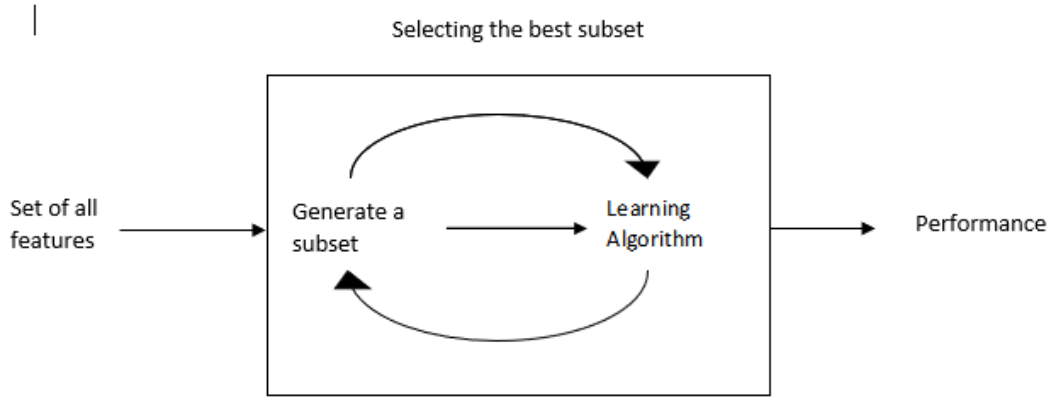

Figure 35: Steps of Wrapper Method

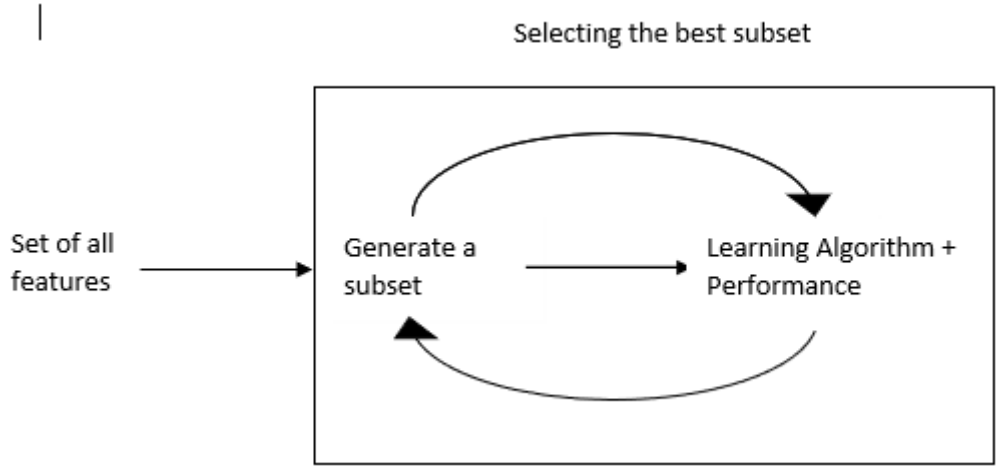

Figure 36: Steps of Embedded Method

We selected Backward Elimination process, a wrapper-based approach of feature selection for the analysis of this study. Its better suits our case study where we had limited number of features. With its advantage of delivering high performance, also comes a few disadvantages of this technique,

- It has a risk of overfitting where observations are inadequate
- Computation complexity is high when the count of variables is large.

Backward elimination steps followed are explained below,

Step 1: Select Significance Level (SL) e.g., 0.05.

Step 2: Fit the model with all possible independent variables.

Step 3: Consider the variable with highest P-value. If P > SL, go to next step
Else Stop the process and **Finish**.

Step 4: Remove the variable selected to be have P > SL.

Step 5: Fit model without the variable removed in previous Step and go to **Step 3**.

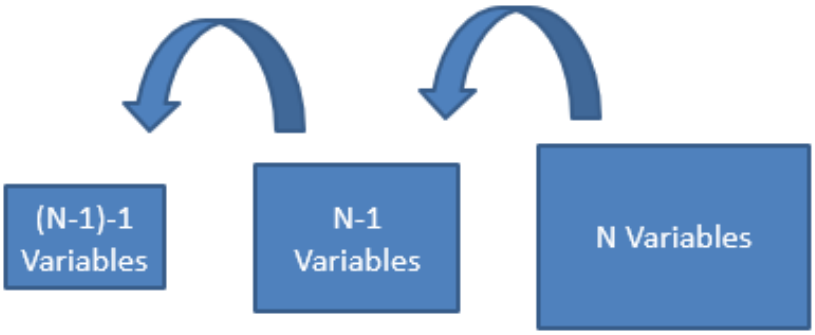

Figure 37: Overview of Backward Elimination Method

The results below show following variables chosen as subset of features based on the selected significance level. It is these features which proves to be strongly correlated and contributes to effective prediction of output variable. We used Simple Multiple Linear Regression with Backward Elimination to test which features are best selected to optimize accuracy of the algorithm in predicting output.

It is also important to understand statistical hypothesis and the concept of P-value in order to perform hypothesis testing. A **statistical hypothesis** is an assumption about a population. **Hypothesis testing** refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. Hypothesis testing using statistical techniques is carried out to examine the correctness of a hypothesis. The P-value is calculated for each predictor aids in deciding if the hypothesis of its association with output variable holds true or not in which case it is either kept in dataset or removed to avoid noise.

Variables	P-Value	Significance
Month1	2.80E-07	***
Month2	0.00271	**
Month5	0.00944	**
Month6	0.00012	***
Month7	0.01672	*
Month8	9.24E-11	***
Month9	0.00713	**
Month10	1.31E-08	***
Month11	2.34E-06	***
Year2013	6.69E-06	***
Year2015	5.77E-10	***
Head	< 2e-16	***
D1	< 2e-16	***
D2	< 2e-16	***

```
Residual standard error: 3222 on 1446 degrees of freedom
Multiple R-squared:  0.9864,    Adjusted R-squared:  0.9863
F-statistic:  7488 on 14 and 1446 DF,  p-value: < 2.2e-16
```

Figure 38: Backward Elimination Results

4.2.2 Model Development Process

Development of each model is divided among 5 major steps followed in each of the algorithms used for analysis in this study.

4.2.2.1 Hyperparameter Tuning:

Hyperparameters are the variables whose values are used by the algorithm for learning. Each algorithm has its own set of parameters which needs to be optimized in order to yield best results. R provides Caret package that assists in providing functionality to find optimal set of parameters and its values for learning process.

4.2.2.2 Fitting Model on Training Dataset:

Each model is trained using pre-defined function in R. Functions expects data, formula (input and output variables) and multiple hyperparameters to learn the scheme. This once completes learns a pattern and model parameters which are used to predict when an unseen set of data is provided to anticipate behavior.

Model Parameter vs Hyperparameter:

Model Parameters	Hyperparameters
Model parameters are learned attributes that define individual models.	Hyperparameters indicate higher-level structural settings for algorithms.
e.g. Regression coefficients Decision Tree split locations	e.g. the number of trees to include in a Random Forest
Learns directly from the (training) data	Decides before fitting the model because they cannot learn from (training) data

Table 5: Comparision of Model Parameters vs Hyperparameters

4.2.2.3 Testing Model Accuracy:

In each of the model, statistical methods were adopted to verify the model. Model's results were evaluated based on the historical data. The predicted results of the model were compared with the real data collected from the partnering organization. The model was validated when the simulated curve came close to the actual data, which enable us to believe that the prediction done is correct and according to the given requirements

This is measured by various statistical methods. We used Root Mean Square Error (RMSE) and Mean Average Percentage Error (MAPE). RMSE is measured in the same unit as the output while MAPE shows the percentage of error in result.

4.2.2.4 Results Visualization:

R facilitates the use of functions which enable to provide pictorial representation of data and results in various forms. GGPLOT package was used to map actual data and predicted values in a 2-Dimensional graph representing in scatter plot and a line chart respectively. This helps to better analyze results with respect to numerical results.

4.2.2.5 Cross Validation:

It is technique used to evaluate the model's performance on a sample set of data. K-folds cross validation method was selected and below procedure was used.

1. Shuffle the dataset randomly
2. Split the dataset into k groups
3. For each group:
   3.1. Take the group as a hold out or test data set,
   3.2. Take the remaining groups as a training set,
   3.3. Fit a model on the training set and evaluate it on the test set,
   3.4. Retain the scores and repeat the steps for next group
4. Summarize model using the sample of model evaluation scores

The value of K was set to 10. Results of each group was evaluated based on the scores of RMSE and MAPE and average was applied. This was compared with the scores of training and test data for validation.

4.2.3 Multiple Linear Regression

4.2.3.1 Model Training

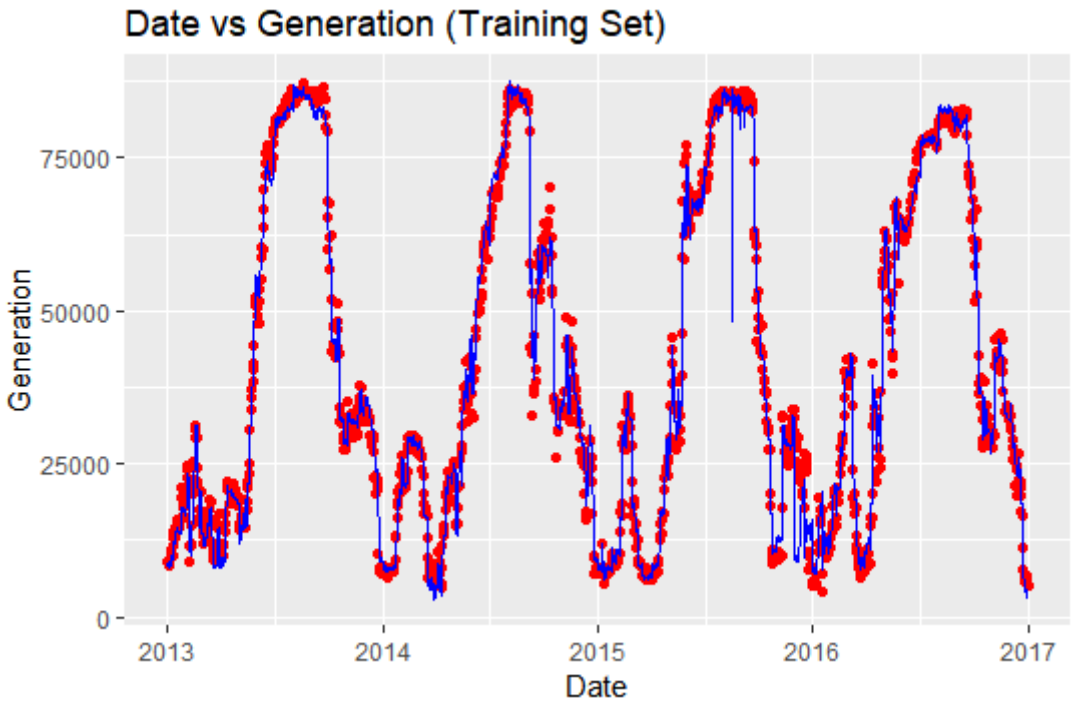

Figure 39: Actual vs Predicted Result of MLR on Train Data

4.2.3.2 Model Testing

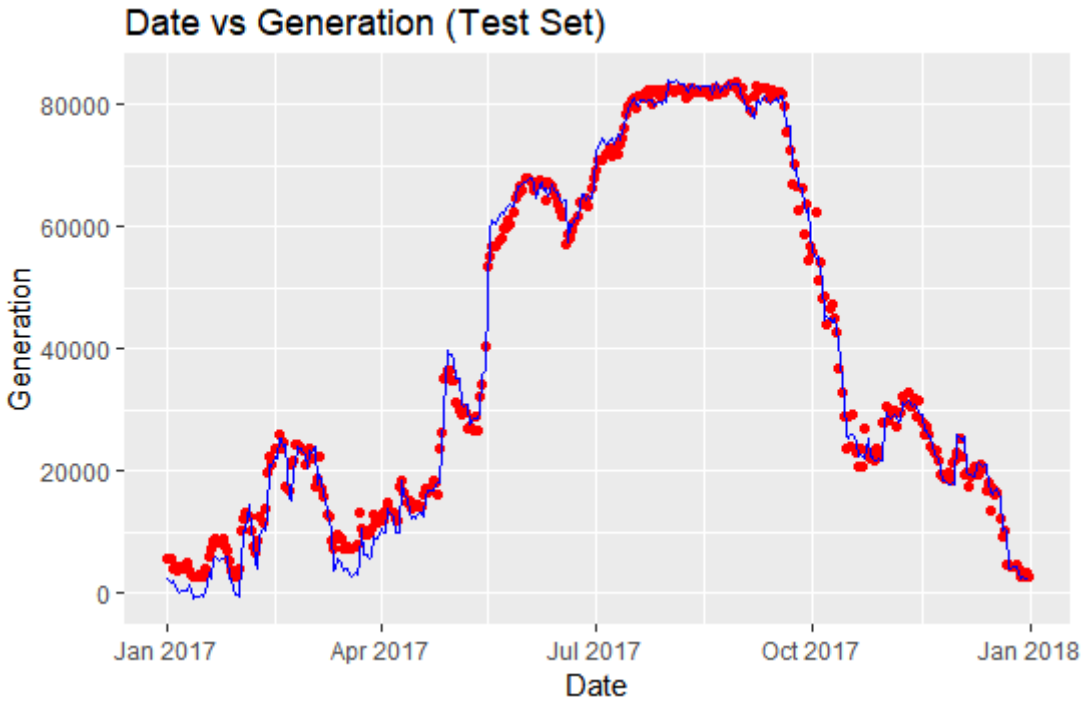

Figure 40: Actual vs Predicted Result of MLR on Test Data

4.2.3.3 Results

Models	RMSE (MWh)		MAPE (%age)	
	Training Set	Test Set	Training Set	Test Set
Multiple Linear Regression	3236.748	2540.105	0.08955796	0.1434325

Table 6: RMSE and MAPE of MLR on Training and Test Set

4.2.3.4 Cross Validation

Models	K-Fold Cross Validation (Training Set)		Error %		Accuracy %	
	RMSE (MWh)	MAPE (%age)	Training Set	Test Set	Training Set	Test Set
Multiple Linear Regression	3324.333	0.09250103	8.955796	14.3432	91.044204	85.65675

Table 7: Cross Validation Results and Accuracy of MLR Model on Training and Test Set

4.2.4 Support Vector Regression

4.2.4.1 Model Training

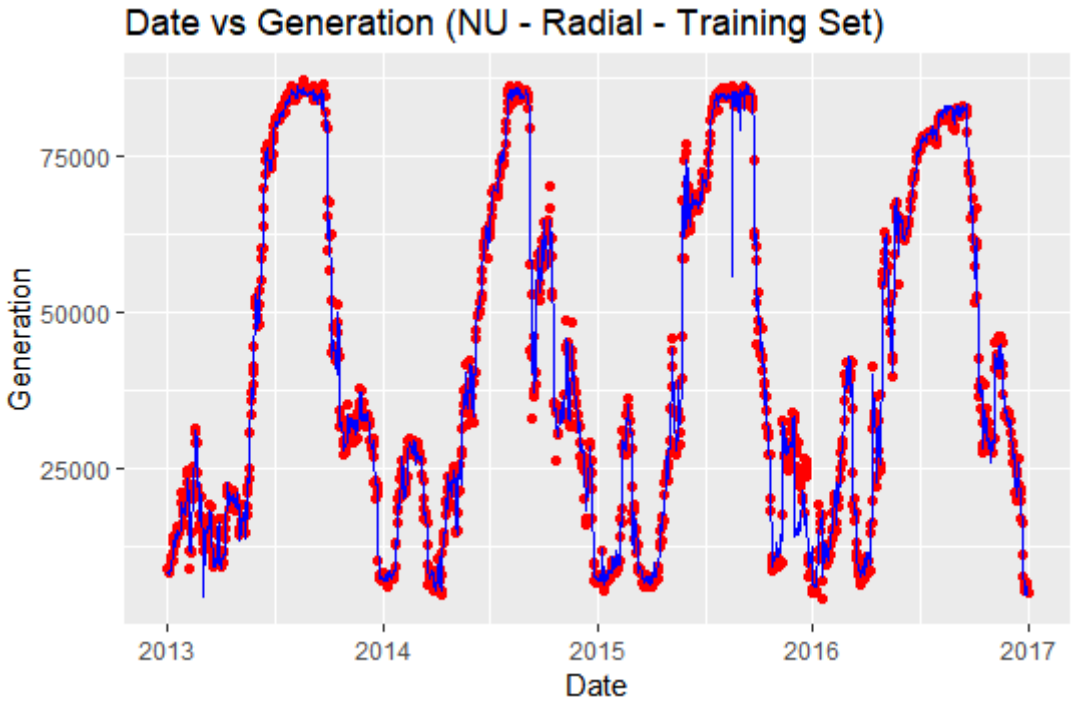

Figure 41: Actual vs Predicted Result of SVR on Train Data

4.2.4.2 Model Testing

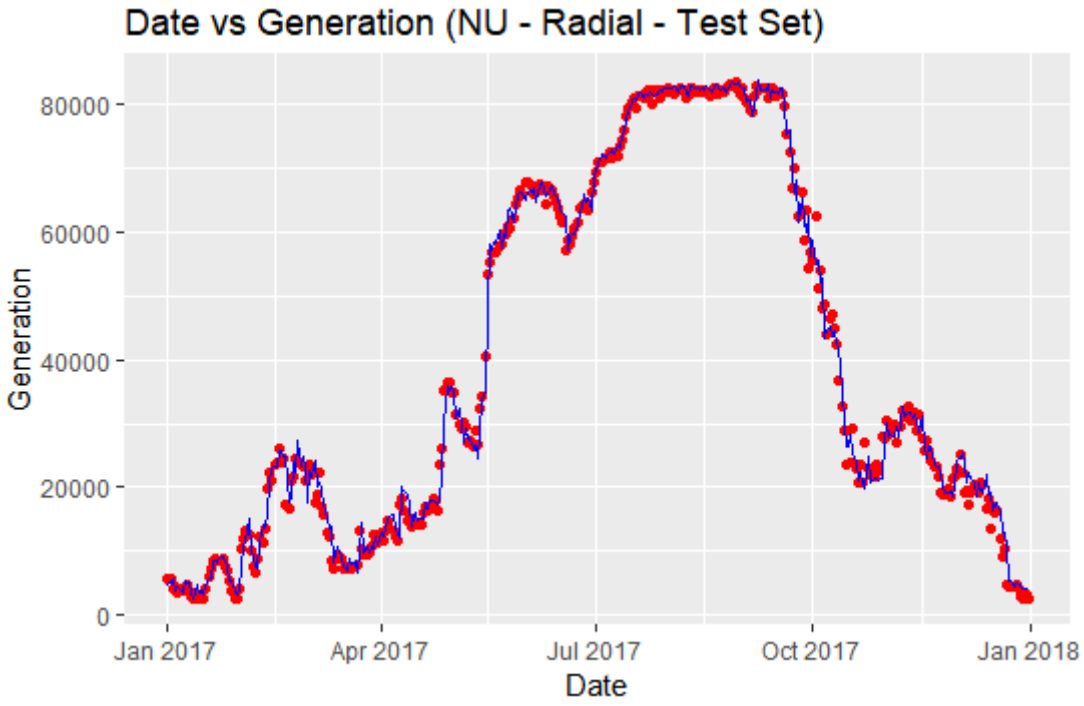

Figure 42: Actual vs Predicted Result of SVR on Test Data

4.2.4.3 Results

Models	RMSE (MWh)		MAPE (%age)	
	Training Set	Test Set	Training Set	Test Set
Linear	3311.037	2688.8	0.0877475	0.180593
Radial	2423.604	2060.3	0.0584842	0.081768
Polynomial	3312.621	2681	0.0877033	0.179725

Table 8: RMSE and MAPE of NU-SVR on Training and Test Set

4.2.4.4 Cross Validation

Models	K-Fold Cross Validation (Training Set)		Error %		Accuracy %	
	RMSE (MWh)	MAPE (%age)	Training Set	Test Set	Training Set	Test Set
Linear	3309.798	0.09079218	8.774751	18.0593	91.225249	81.94067
Radial	3290.213	0.0881765	5.848422	8.17680	94.151578	91.82319
Polynomial	3343.129	0.08995808	8.770334	17.9725	91.229666	82.02748

Table 9: Cross Validation Results and Accuracy of NU-SVR Model on Training and Test Set

4.2.5 Decision Trees

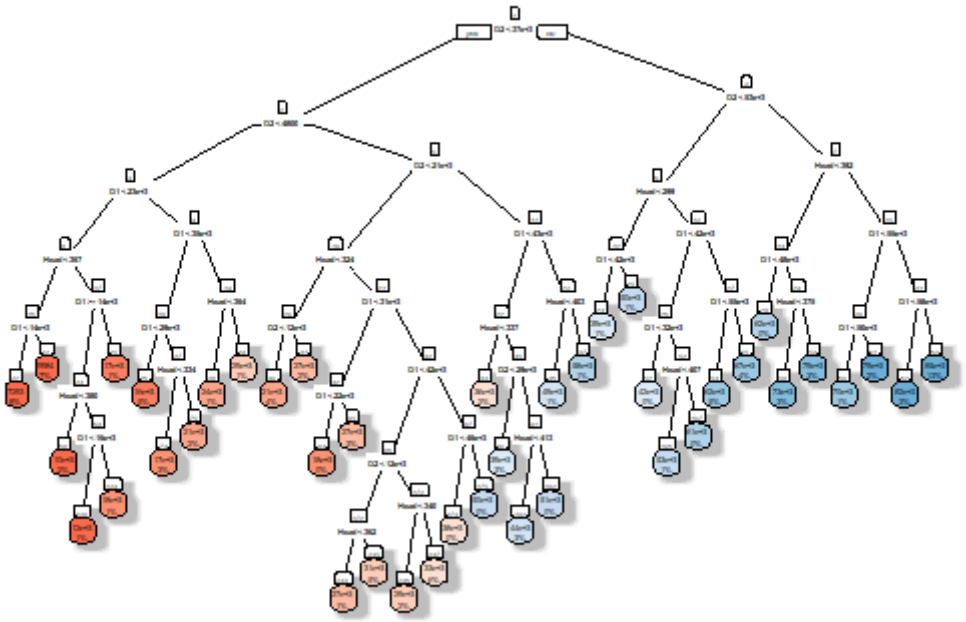

Figure 43: Graphical Representation of Decision Tree

4.2.5.1 Model Training

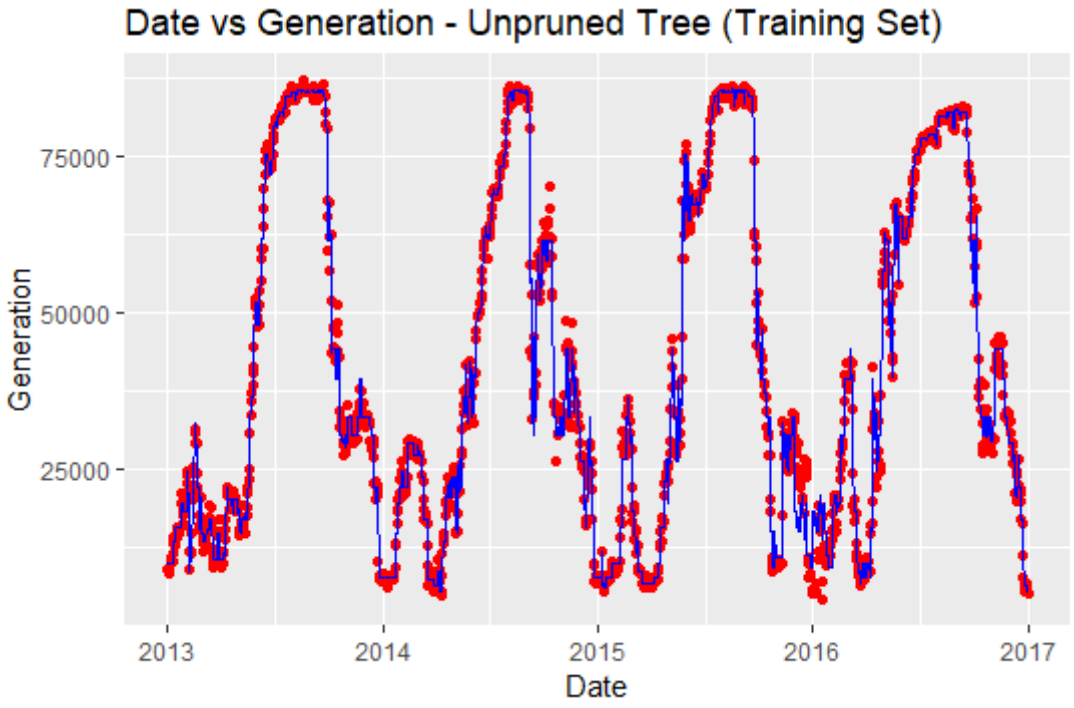

Figure 44: Actual vs Predicted Result of Decision Tree on Train Data

4.2.5.2 Model Testing

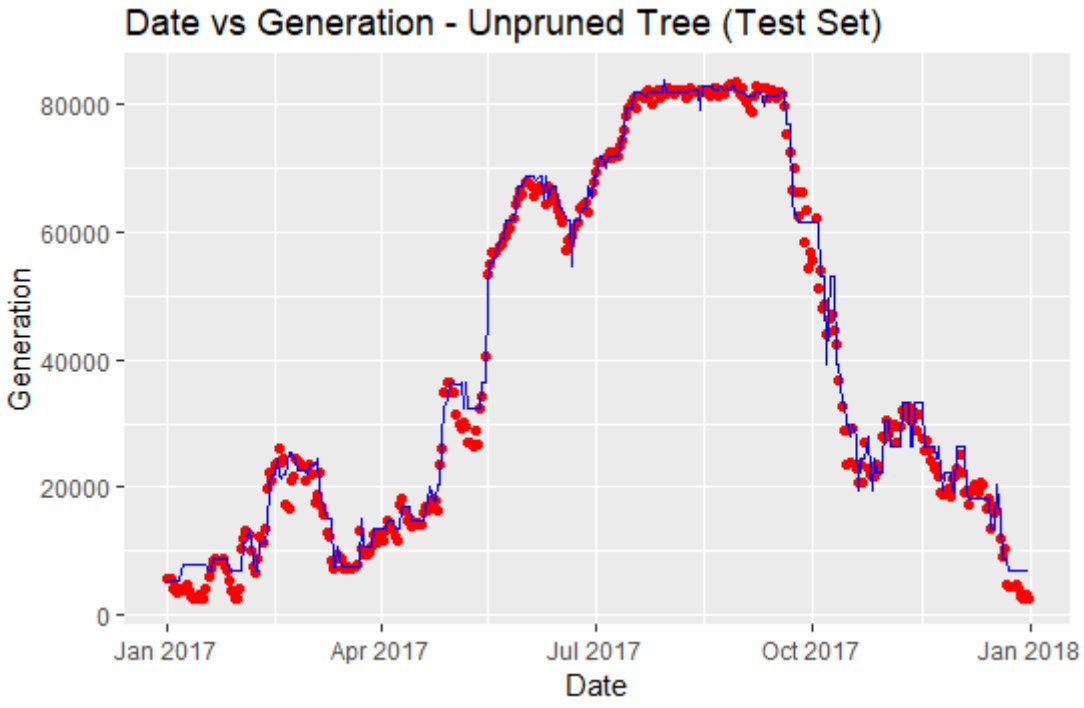

Figure 45: Actual vs Predicted Result of Linear Regression on Test Data

4.2.5.3 Results

Models	RMSE (MWh)		MAPE (%age)	
	Training Set	Test Set	Training Set	Test Set
Decision Tree	2348.527	2751.9	0.0669384	0.156875

Table 10: RMSE and MAPE of Decision Tree on Training and Test Set

4.2.5.4 Cross Validation

Models	K-Fold Cross Validation (Training Set)		Error %		Accuracy %	
	RMSE (MWh)	MAPE (%age)	Training Set	Test Set	Training Set	Test Set
Decision Tree	4082.821	0.1105791	6.693837	15.6875	93.306163	84.31246

Table 11: Cross Validation Results and Accuracy of Decision Tree Model on Training and Test Set

4.2.6 Random Forest

4.2.6.1 Model Training

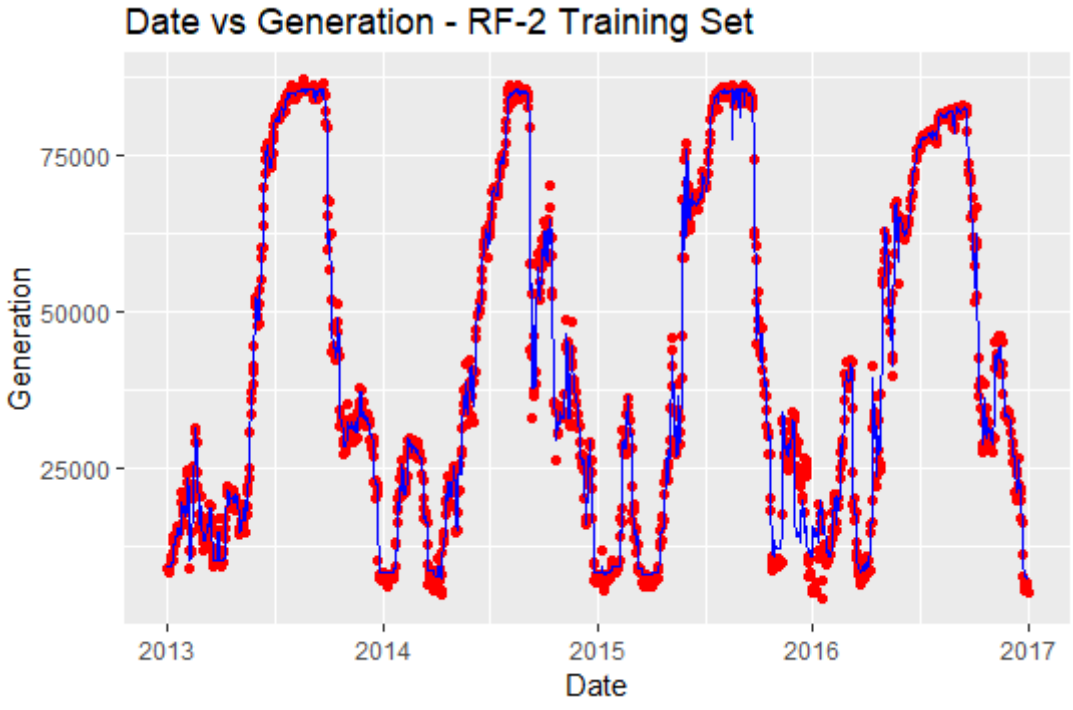

Figure 46: Actual vs Predicted Result of Random Forest on Train Data

4.2.6.2 Model Testing

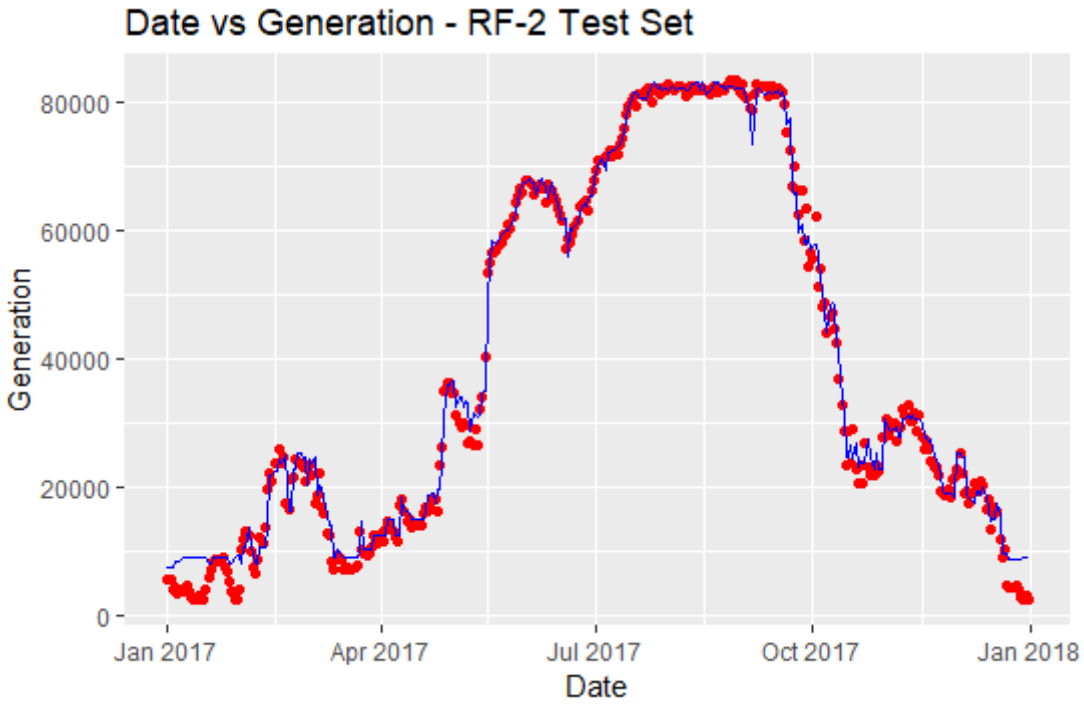

Figure 47: Actual vs Predicted Result of Random Forest on Train Data

4.2.6.3 Results

Models	RMSE (MWh)		MAPE (%age)	
	Training Set	Test Set	Training Set	Test Set
Random Forest	2092.985	2514.9	0.069168	0.19228

Table 12: RMSE and MAPE of Random Forest on Training and Test Set

4.2.6.4 Cross Validation

Models	K-Fold Cross Validation (Training Set)		Error %		Accuracy %	
	RMSE (MWh)	MAPE (%age)	Training Set	Test Set	Training Set	Test Set
Random Forest	3062.631	0.09373203	6.916804	19.2280	93.083196	80.77198

Table 13: Cross Validation Results and Accuracy of Random Forest Model on Training and Test Set

4.2.7 Artificial Neural Networks

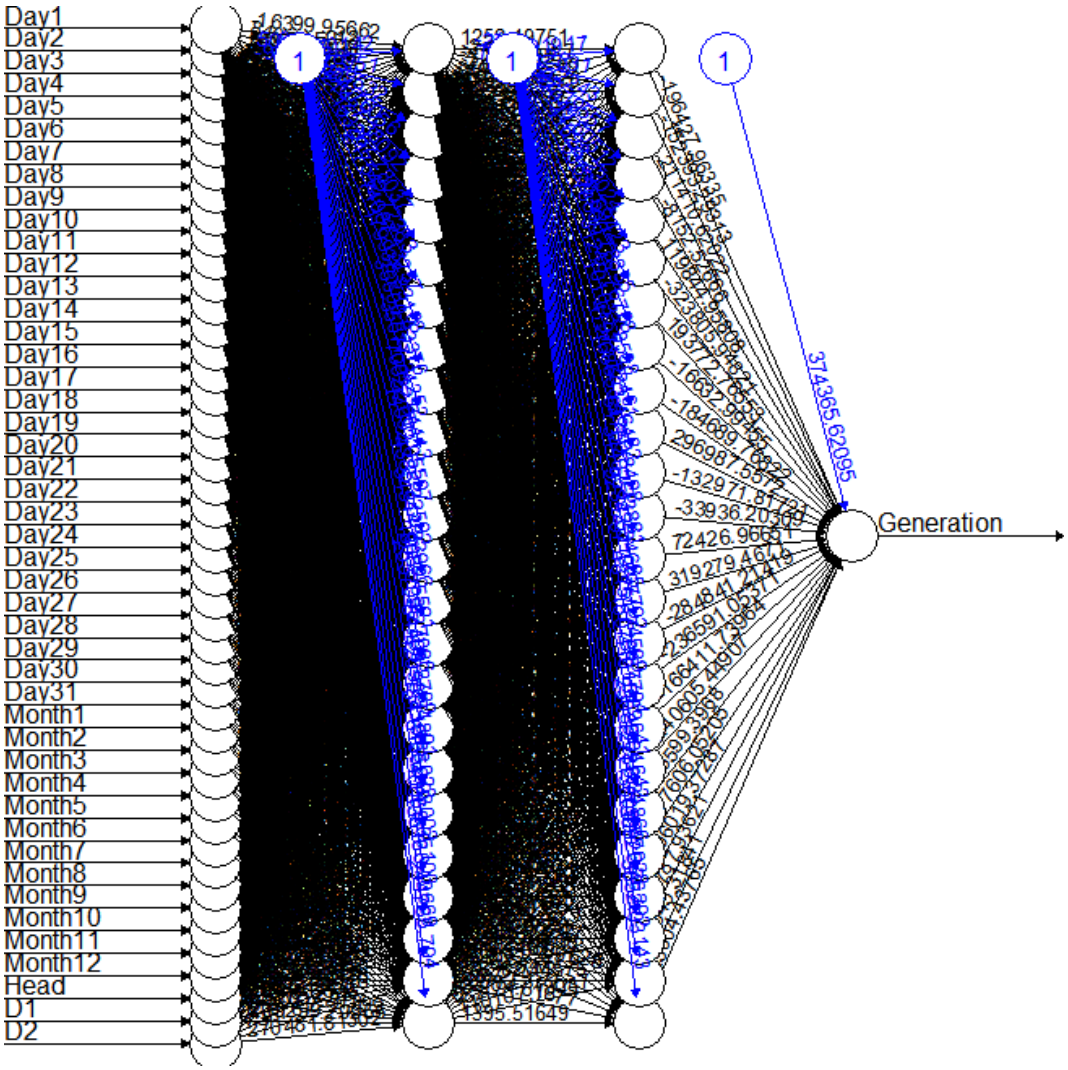

Figure 48: Graphical Representation of a 2 Layer ANN

4.2.7.1 Model Training

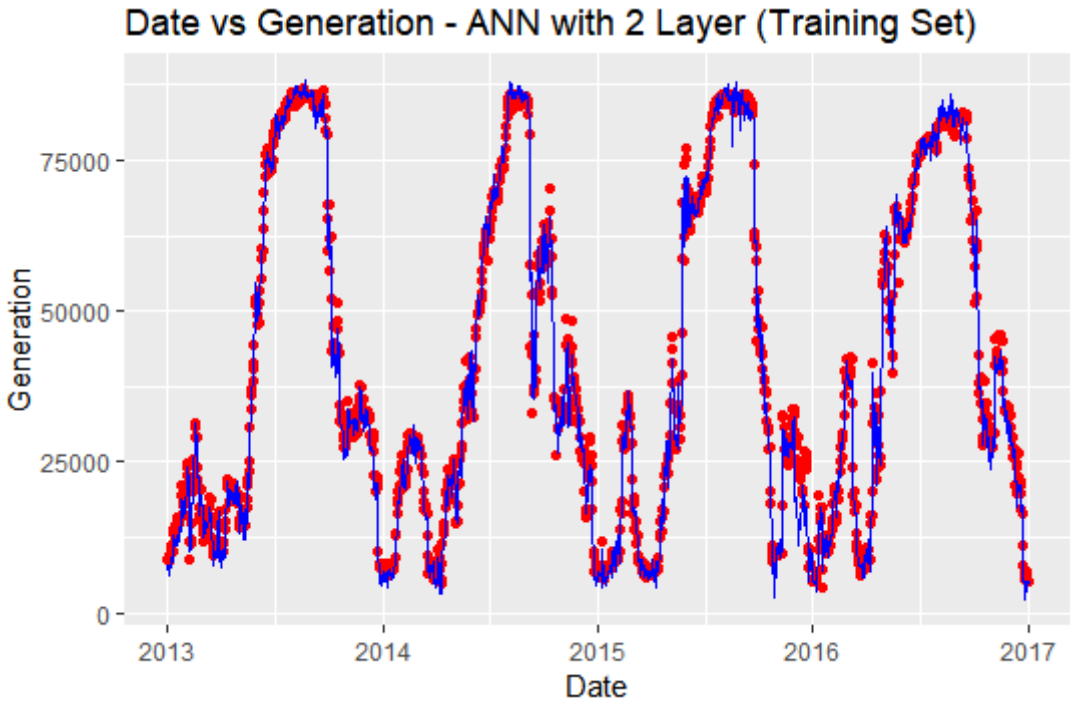

Figure 49: Actual vs Predicted Result of ANN on Train Data

4.2.7.2 Model Testing

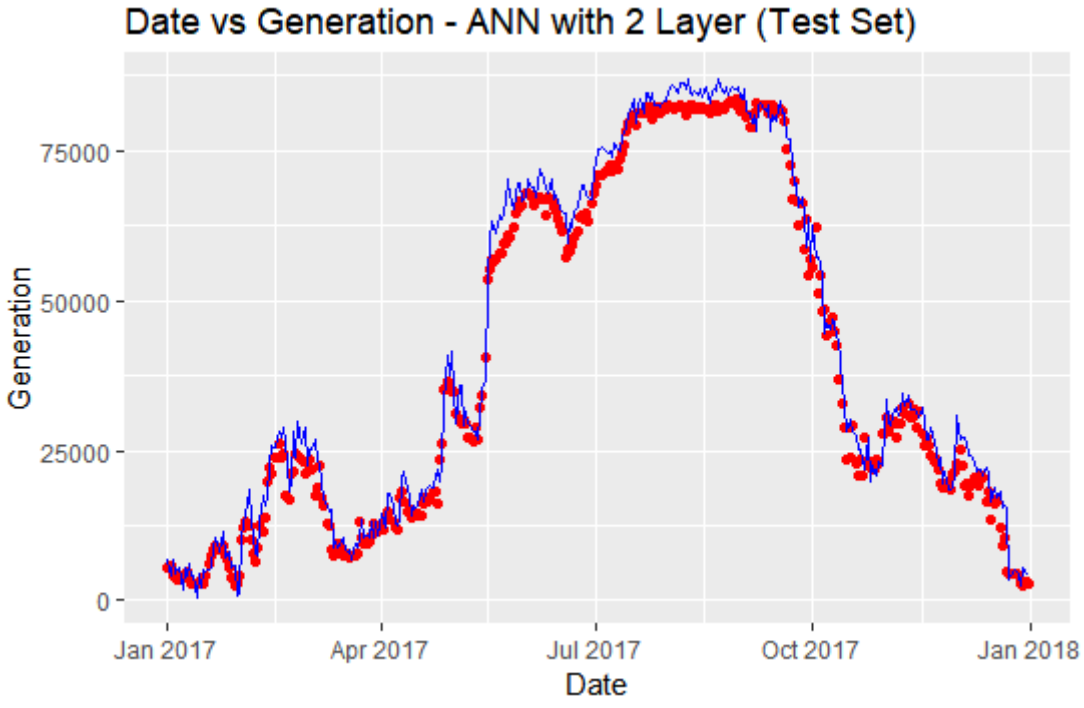

Figure 50: : Actual vs Predicted Result of ANN on Test Data

4.2.7.3 Results

Models	RMSE (MWh)		MAPE (%age)	
	Training Set	Test Set	Training Set	Test Set
Rectifier (24)	2657.002	3536.3	0.0763236	0.172215
Rectifier (24,24)	2676.286	3199.9	0.0835787	0.160203

Table 14: RMSE and MAPE of ANN on Training and Test Set

4.2.7.4 Cross Validation

Models	K-Fold Cross Validation (Training Set)		Error %		Accuracy %	
	RMSE (MWh)	MAPE (%age)	Training Set	Test Set	Training Set	Test Set
Rectifier (24)	4624.471	0.1175818	7.632363	17.2214	92.367637	82.77851
Rectifier (24,24)	3290.213	0.0881765	5.848422	8.17680	94.151578	91.82319

Table 15: Cross Validation Results and Accuracy of ANN Model on Training and Test Set

4.2.8 Comparison of Results

Models	RMSE (MWh)		MAPE (%age)	
	Training Set	Test Set	Training Set	Test Set
Support Vector Regression	2423.604	2060.261	0.05848422	0.08176802
Artificial Neural Network	2676.286	3199.863	0.08357867	0.1602034
Decision Tree	2348.527	2751.948	0.06693837	0.1568754
Random Forest	2092.985	2514.932	0.06916804	0.1922802
Multiple Linear Regression	3236.748	2540.105	0.08955796	0.1434325

Table 16: RMSE and MAPE of all Models on Training and Test Set

Models	K-Fold Cross Validation (Training Set)		Error %		Accuracy %	
	RMSE (MWh)	MAPE (%age)	Training Set	Test Set	Training Set	Test Set
Support Vector Regression	3290.213	0.0881765	5.848422	8.176802	94.151578	91.82319
Artificial Neural Network	4145.289	0.1443172	8.357867	16.02034	91.642133	83.97966
Decision Tree	4082.821	0.1105791	6.693837	15.68754	93.306163	84.31246
Random Forest	3062.631	0.09373203	6.916804	19.22802	93.083196	80.77198
Multiple Linear Regression	3324.333	0.09250103	8.955796	14.34325	91.044204	85.65675

Table 17: Cross Validation Results and Accuracy of all Models on Training and Test Set

Chapter 5

Conclusion and Future Work

This chapter provides the discussion, conclusion and future work of this study.

5.1 Conclusion

The statistical computation of will provide a concept and visualization of the future demand and supply of hydro electric energy in Pakistan with focus on Tarbela Power Plant. This indeed can be extended to other hydel projects combining all to form a tool for better management of energy production.

Our study for electricity supply will be useful in forecasting future energy demand and will play a key role in energy planning and distribution. The results from the model will provide insights for analysts to answer and better equipped to address different challenges faced during of the year. It will further lead decision makers to adopt optimal choices for future energy planning in the country.

A brief aim of this analysis is as follows,

- The aim is to provide a means to forecast generation capacity of Tarbela Power Plant using antecedent values.
- Allow stakeholders take timely decisions to alter generation parameters.
- Help improve and maintain uniformity of electricity generation.

5.2 Future Work

This study is the key need of the hour for our country. It can help lead us solve the bigger challenge of energy planning. This study proposes a generalized solution to the problem of demand and supply gap.

This study can be extended to other hydro installations and we propose to run our model for different Hydro Power Plants to predict their energy generation patterns. This will allow stakeholders to get a better insight of their overall installed production capacity. Moreover, it will assist distribution authorities to manage electric shortfalls and will help cope with current energy deficit.

We also aim to extend this study and analyze micro level energy production capacity of each generator using same techniques. This will require additional data from the source authority and will result in improving energy generation process and provide an effective utilization of Power Plant resources.

References

[1] B. Thinnes, "Energy for economic growth," *Hydrocarb. Process.*, vol. 91, no. 4, 2012.

[2] X. Luo, J. Wang, M. Dooner, and J. Clarke, "Overview of current development in electrical energy storage technologies and the application potential in power system operation," *Appl. Energy*, vol. 137, pp. 511–536, 2015.

[3] N. Simcock and C. Mullen, "Energy demand for everyday mobility and domestic life: Exploring the justice implications," *Energy Res. Soc. Sci.*, vol. 18, pp. 1–6, 2016.

[4] I. N. Kessides, "Chaos in power: Pakistan's electricity crisis," *Energy Policy*, vol. 55, pp. 271–285, 2013.

[5] G. Das Valasai, M. A. Uqaili, H. U. R. Memon, S. R. Samoo, N. H. Mirjat, and K. Harijan, "Overcoming electricity crisis in Pakistan: A review of sustainable electricity options," *Renew. Sustain. Energy Rev.*, vol. 72, no. January, pp. 734–745, 2017.

[6] Safiya Aftab, "Pakistan's energy crisis: causes, consequences and possible remedies," no. January, pp. 1–4, 2014.

[7] U. Zafar, T. Ur Rashid, A. A. Khosa, M. S. Khalil, and M. Rahid, "An overview of implemented renewable energy policy of Pakistan," *Renew. Sustain. Energy Rev.*, vol. 82, no. June 2017, pp. 654–665, 2018.

[8] M. Jalas and J. K. Juntunen, "Energy intensive lifestyles: Time use, the activity patterns of consumers, and related energy demands in Finland," *Ecol. Econ.*, vol. 113, pp. 51–59, 2015.

[9] M. Amer and T. U. Daim, "Selection of renewable energy technologies for a developing county: A case of Pakistan," *Energy Sustain. Dev.*, vol. 15, no. 4, pp. 420–435, 2011.

[10] D. Gielen, F. Boshell, D. Saygin, M. D. Bazilian, N. Wagner, and R. Gorini, "The role of renewable energy in the global energy transformation," *Energy Strateg. Rev.*, vol. 24, no. January, pp. 38–50, 2019.

[11] British Petroleum, "BP Statistical Review of World Energy What's inside ?," *Nucl. Energy*, vol. 4, no. June, pp. 1–50, 2010.

[12] G. D. Towards, *100 % Renewable Energy Renewables Global Futures Report*. 2017.

[13] F. U. Qureshi and B. Akıntug, "Hydropower Potential in Pakistan," *IEEE 11th Int. Congr. Adv. Civ. Eng.*, no. November, pp. 1–6, 2014.

[14] M. Praveena and V. Jaiganesh, "A Literature Review on Supervised Machine Learning Algorithms and Boosting Process," *Int. J. Comput. Appl.*, vol. 169, no. 8, pp. 32–35, 2017.

[15] B. S., "Brandt S. (1999) Linear and Polynomial Regression. In: Data Analysis. Springer, New York, NY," *Springer, New York, NY*, no. 1, 1999.

[16] J. L. Crowley and E. M. M, "Intelligent Systems : Reasoning and Recognition Introduction to Bayesian Recognition," vol. 2009, pp. 1–8, 2012.

[17] S. P. Mahila, A. Pradesh, and S. P. Mahila, "E NSEMBLE D ECISION T REE

C LASSIFIER F OR," vol. 2, no. 1, pp. 17–24, 2012.

[18] S. Salzberg and A. Segre, "Book Review:" C4. 5: Programs for Machine Learning" by J. Ross Quinlan," *Mach. Learn.*, vol. 240, pp. 235–240, 1994.

[19] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, 2006.

[20] V. Y. Kulkarni and P. K. Sinha, "Pruning of random forest classifiers: A survey and future directions," *Proc. - 2012 Int. Conf. Data Sci. Eng. ICDSE 2012*, pp. 64–68, 2012.

[21] M. Dehghani *et al.*, "Prediction of hydropower generation using grey Wolf optimization adaptive neuro-fuzzy inference system," *Energies*, vol. 12, no. 2, pp. 1–20, 2019.

[22] H. Alrayess, S. S. Gharbia, and N. Beden, "Using Machine Learning Techniques and Deep Learning in Forecasting The Hydroelectric Power Generation in Almus Dam, Turkey," no. October, 2018.

[23] I. Enoidem Ebukanson, "Statistical Analysis of Electricity Generation in Nigeria Using Multiple Linear Regression Model and Box-Jenkins' Autoregressive Model of Order 1," *Int. J. Energy Power Eng.*, vol. 6, no. 3, p. 28, 2017.

[24] A. T. Hammid, M. H. Bin Sulaiman, and A. N. Abdalla, "Prediction of small hydropower plant power production in Himreen Lake dam (HLD) using artificial neural network," *Alexandria Eng. J.*, vol. 57, no. 1, pp. 211–221, 2018.

[25] G. Li, C. X. Liu, S. L. Liao, and C. T. Cheng, "Applying a correlation analysis method to long-Term forecasting of power production at small hydropower plants," *Water (Switzerland)*, vol. 7, no. 9, pp. 4806–4820, 2015.

[26] G. Li, Y. Sun, Y. He, X. Li, and Q. Tu, "Short-term power generation energy forecasting model for small hydropower stations using GA-SVM," *Math. Probl. Eng.*, vol. 2014, 2014.

[27] E. Uzlu, A. Akpinar, H. T. Özturk, S. Nacar, and M. Kankal, "Estimates of hydroelectric generation using neural networks with the artificial bee colony algorithm for Turkey," *Energy*, vol. 69, pp. 638–647, 2014.

[28] F. Olsson and M. Pearson, "Modeling the total inflow energy to hydropower plants -a study of Sweden and Norway."

[29] R. S. V. Teegavarapu and S. P. Simonovic, "Simulation of Multiple Hydropower Reservoir Operations Using System Dynamics Approach," *Water Resour. Manag.*, vol. 28, no. 7, pp. 1937–1958, 2014.

[30] A. Sharifi, L. Kalin, and M. Tajrishy, "System Dynamics Approach for Hydropower Generation Assessment in Developing Watersheds: Case Study of Karkheh River Basin, Iran," *J. Hydrol. Eng.*, vol. 18, no. 8, pp. 1007–1017, 2012.

[31] "Modeling hydropower plant system to improve its reservoir operation," *Int. J. Water Resour. Environ. Eng.*, vol. 4, no. 2, pp. 87–94, 2010.

[32] M. Abid and M. U. R. Siddiqi, "Multiphase Flow Simulations through Tarbela Dam Spillways and Tunnels," *J. Water Resour. Prot.*, vol. 02, no. 06, pp. 532–539, 2010.

[OL-1] L. Statistics, "How to perform a Multiple Regression Analysis in SPSS Statistics | Laerd Statistics", Statistics.laerd.com, 2018. [Online]. Available:

https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php.

[OL-2] S. Ray, "Understanding Support Vector Machines algorithm (along with code)", Analytics Vidhya, 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

[OL-3] S. Suriya, "Support Vector Machines – Kernel Explained", Machine Learning & Artificial Intelligence, 2018. [Online]. Available: https://codingmachinelearning.wordpress.com/2016/07/25/support-vector-machines-kernel-explained/

[OL-4] L. Chen, "Support Vector Machine — Simply Explained", Medium, 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496

[OL-5] R. S. Bird, "Decision Trees — A simple way to visualize a decision", Medium, 2019. [Online]. Available: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

[OL-6] G. Drakos, "Random Forest Regression model explained in depth", GDCoder, 2018. [Online]. Available: https://gdcoder.com/random-forest-regressor-explained-in-depth/