Sudan University for Science & Technology

Collage of Computer Science and Information Technology

Prediction of the Generated Amount of Power From a Thermal Power Plant Using Data Mining

التنبؤ بكمية الطاقة الكهربائية المنتجة من محطة توليد حراري بإستخدام التنقيب في البيانات

A thesis submitted in Partial fulfilment of the requirements

for the award of the degree of

Doctor of Philosophy (Computer Science)

Student: Waleed Hamed Ahmed Eisa

Supervisor: Prof. Dr. Naomie Bt Salim

Nov, 2017

DEDICATION

My warm and special thanks to my sheikhs, family: my parents, my wife and kids, my brothers and sister, for supporting me spiritually on writing this thesis and throughout my life.

ACKNOWLEDGEMENT

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr. Naomie Bt Salim for the continuous support of my Ph.D. study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank Prof. Dr. Izeldin Mohammed Osman, who has invented and leads the unique phd program of Computer Sciences in Sudan University of Science and Technology, for his innovation of the program, for his wise comments and encouragement. I have the honor of being his student in my Masters degree.

Also I would like to thank the rest of my thesis committee: Dr. Mohammed Elhafiz, Dr.Talat Wahbi , for their insightful comments and encouragement, but also for the hard question which incentivized me to widen my research from various perspectives. I thank Dr. Yahya Abdullah, Dr. Amir A. Fattah, Prof. Zaroog, for enlightening the first steps of my research. Also I am grateful to the program management team for their dedication and patient.

My sincere thanks also goes to Eng. Zuhair and Eng. Tarig , who provided me great knowledge about power generation, and gave me access to the power plant historical databases. Without they precious support it would not be possible to conduct this research.

ABSTRACT

Electricity has changed everyone's life since the day it was discovered. It is produced all over the world in power plants that uses different types of energy sources like fossil fuel, water falls, nuclear and wind Energy. The amount of electricity produced is the fundamental goal for any power plant. Therefor accurate prediction of this amount is very important for planning and operation activities of the power plant. This study aims to identify the attributes that influence the amount of generated power from a thermal power plant, and to accurately predict that amount. Datasets were prepared form real data that had been collected by SCADA over two years period, from two different units in a thermal power plant of 190 MW capacity. Feature selection was done using wrapper method, and power prediction was done using all available attributes, to give a ranking of the selected attributes, and show the influence of each parameter in the amount of generated power. For power prediction, only controllable parameters like pressure, temperature and steam flow at turbine inlet were used as predictors. Sixteen different algorithms were tested for each dataset, the algorithm that showed higher correlation coefficient and minimum error was selected to build the model. The predicted amount using data mining was found to be more accurate than manufacturers expectations and thermodynamic laws. Models evaluation was done using separate dataset, and cross validation in case of small datasets. Moreover comparison between the predicted and the actual observed amounts was presented in a graph, to visualize the accuracy of the models.

المستخلص

لقد تغير أسلوب حياة الإنسان كثيراً بإكتشاف الطاقة الكهربائية التي أصبحت تستعمل بصورة أساسية في كل مجالات الحياة اليومية. .يتم إنتاج الكهرباء في جميع أنحاء العالم في محطات التوليد الكهربائي عن طريق تحويل الطاقة من شكل أولي الى الطاقة الكهربائية. تستخدم محطات التوليد أنواعاً مختلفةُ من مصادر الطاقة مثل الوقود الأحفوري، وجريان مياه الأنهار, والطاقة النووية وطاقة الرياح . تعتبر كمية الكهرباء المنتجة من محطات التوليد هي الهدف الأساسي لأي محطة توليد, و لذلك فإن معرفة الكمية المنتجة بدقة مهم جدا للتخطيط الإستراتيجي و لمعرفة كفاءة عملية التوليد.

تهدف هذه الدراسة إلى التعرف على الصفات التي تؤثر على كمية الطاقة المنتجة من محطات توليد الطاقة الحرارية، والتنبؤ بدقة بهذه الكمية .تم إعداد البيانات المطلوبة لإجراء هذا البحث من قراءات حقيقية تم جمعها بواسطة نظام التحكم و المراقبة بالمحطة على مدى عامين، من وحدتين مختلفتين من محطة حرارية قدرتها الإنتاجية 190 ميغاواط (محطة الشهيد محمود شريف ببحري). وقد تم اختيار المواصفات المؤثرة على إنتاج الطاقة الكهربائية باستخدام طريقة المجمع Wrapper، وتم التنبؤ بكمية الطاقة ببناء نماذج التنبؤ بطريقتين مختلفتين. الطريقة الأولى باستخدام جميع السمات المتاحة، و من ثم إعطاء ترتيباُ لهذه السمات يحدد مدى أهميتها و أثرها على الكمية المنتجة. أما الطريقة الثانية فإنها تستخدم فقط السمات التي يمكن التحكم بها (وهي الضغط ودرجة الحرارة وسرعة تدفق البخار في مدخل التوربينات). و تم تجربة سبعة عشر خوارزمية مختلفة لكل مجموعة بيانات، وبعد ذلك تم إختيار الخوارزمية التي أظهرت أعلى معامل ارتباط و أقل نسبة خطأ لبناء النموذج النهائي.

لقد أثبت البحث أن الكمية المتوقعة باستخدام تقنيات التنقيب في البيانات تكون أكثر دقة من الطرق التقليدية (معادلات الديناميكا الحرارية و معادلات الكفاءة المعدة بواسطة مصنعي المحطة) .تم إجراء تقييم النماذج باستخدام مجموعة منفصلة من بيانات. وعلاوة على ذلك فقد تم إجراء مقارنة بين الكميات المحسوبة بواسطة معادلات الديناميكا الحرارية, و المحسوبة بواسطة معادلات الكفاءة المعدة بواسطة مصنعي المحطة , و الفعلية , و المتوقعة بواسطة النموذج الجديد. كل ذلك في رسم البياني واحد لتسهل عملية المقارنة.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

KNPP	-	Khartoum North Power Plant
SVM	-	Support Vector machine
MLP	-	Multi Layer Perceptron
LR	-	Linear Regression
PR	-	Pace Regression
DM	-	Data Mining

LIST OF APPENDICES

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Electricity generation is the process of generating electric power from primary energy sources. The basic principle of electricity generation was discovered by Michael Faraday during the early nineteenth century. In this basic method the electricity is generated by the movement of a loop of wire, or disc of copper between the poles of a magnet, then the motion between a magnetic field and a conductor creates an electrical current. This method is the base of power systems (Zhang, 2010).

The power system which is also known as the grid is divided into three components: the generator which produce the power, the transmission system that carries the power from the generators to the load centers, and the distribution which delivers power to the end users. There are many types of generators (also known as power plant) normally these power plants contain one or more generators, which is a rotating machine that converts mechanical power into electrical power.

Mainly, power is generated by different types of turbine. Turbines are commonly driven by wind, water, gas or steam. The turbine then drives an electric generator. The different types of basic energy sources which are used to rotate the turbines include:

1. **Wind**: where the wind moves big fans that rotates the turbines.
2. **Water**: Energy is captured from the movement of water. From falling water (dam), this produces about 16% of the worlds electricity.
3. **Gas**: Natural gas is burned in a gas turbine, by Combined cycle which are driven by both steam and natural gas. At least 20% of the worlds electricity is generated by natural gas.
4. **Steam**: In steam turbines, water is boiled by:

    i. Fossil energy source like coal or oil in a thermal power plant, about 40% of all electricity is generated this way.
    ii. Nuclear fission used heat created in a nuclear reactor creates steam. Less than 15% of electricity is generated this way.

Renewables energy sources are being used in steam like: Biomass, Solar thermal energy , and Geothermal power (Abolhosseini, Heshmati and Altmann, 2014)).

In Sudan Water Energy and Thermal Power Plants (steam) are used. This research focuses on thermal power plants that uses oil as energy source, these types of power plants use ***Rankine Cycle*** to generate electricity. More theoretical investigation about Rankine Cycle could be found in (Kapooria et al, 2008). Rankine Cycle is a closed system consists of four main components, that are interconnected together to build one system (Learn Engineering, 2013). These components are :

1. **Steam Turbine** which uses the superheated steam that is coming from the boiler to rotate the turbine blades.
2. **Condenser**: uses external cooling water to condense the steam which is exhausted from turbine to liquid water.
3. **Feed water Pump**: to pump the liquid to a high pressure and bush it again to boiler.
4. **Boiler** which is externally heated to boil the water to superheated steam.

Nowadays power plants are commissioned with Supervisory Control And Data Acquisition (SCADA) systems, to ease the control and data collection of power plant components. Data is collected instantly by SCADA system via different types of sensors that are connected to different locations of power plant to capture hundreds of

parameters, each of which is given a unique Tag No or Parameter ID. This huge amount of data is recorded in a historical database. Unfortunately, due to disk space limitation, always there is a data purging process to delete data older than a specific period, and keep the space utilization within acceptable threshold. Of course this reduces the amount of data available for analysis.

Power calculations

Data like pressure, temperature and flow rate, which is collected from different locations in the power plant, could be used for different purposes like: monitoring the current status of each component, or calculating the expected amount of generated power to estimate the efficiency of the power plant.

The availability of this huge amount of real time data encourages the adoption of data mining techniques. Data mining is defined as the process of discovering patterns in data (Witten and Frank and Hall, 2011). This research is studying the possibility of using data mining techniques to solve real problems in power plants. However, there are some obstacles that faces researchers and engineers to benefit from data mining in this area. The first one is the interdisciplinary nature of such a research, because it requires deep knowledge in both IT and electromechanical engineering, this obstacle is handled by intensive review of the literature and many meetings with the domain expert. Another obstacle is the lack of standard analysis methods and benchmarks, this leads to the use of different methods and datasets and compare them to come up with the most appropriate one.

Kartoum **N**orth **P**ower **P**lant (KNPP) is one of the biggest thermal power plants in Sudan. KNPP was commissioned in three main phases. Each phase is composed of two identical units. Phase II, which is composed of Unit 3 and Unit 4, was selected as a case study for this research. Steam properties were collected instantly by SCADA system, using sensors that are connected at different locations in the power plant. So all required data is available at the central database. To distinguish between all these parameters, a unique number is assigned for each one. More details about these parameters and dataset will be provided in chapter 3 Research Methodology.

1.2 Problem Background

Data mining could be used to solve many types of problems in power plants, like prediction of generated power, prediction of machine failure, diagnosis of machine failure and many others. There is no standard technique that could be used for a specific problem. In the literature reviewed, a lot of researchers provide many interesting articles. Like the researches of wind turbine performance which were conducted under one research project in University of Iowa in USA at 2012. The project name is "Data-Driven Performance Optimization of Wind Farms research", and leaded by Prof. Andrew Kusiak. (Küçüksille, Selbas and Sencan, 2011) focused on prediction of thermodynamics properties, although it is about refrigerants, but it provides a very clear road map of using data mining in engineering and thermodynamic problems.

Recently the use of data mining applications in power systems is increasing. Many papers were found in the literature, each is focusing on one area of the power system. Some are focusing on the **Distribution System** like (Ramos, 2008) who used decision tree to classify the consumers. (Saibal, 2008) used WN (Wavelet Networking) which is an extension of perceptron networks for the Classification of transients. (Figueiredo, Rodrigues and Gouveia, 2005) used Decision tree for the Classification of Electricity energy consumer. (Dola, 2005) used Decision tree and neural network for Faults classification in distribution system. (Mori, 2002) used Regression tree and neural network for Load forecasting.

Other researcher focused on the **Transmission Line** Problems like: (Hagh, 2007), (Silva, 2006), (Costa, 2006), (Vasilic, 2002) all of them used Neural network to study Faults detection, classification and locations in Transmission Lines. (Dash, 2007) used Support Vector Machine for the classification and identification of series compensated. (Vasilic, 2005) and (Huisheng, 1998) used Fuzzy/ neural network for faults classification.

Some other researchers focused on **Power Generation** part (power plants) like (Kusiak, Zhang and Li, 2011) who used a multi objective optimization model to

optimize wind turbine performance. Others like (Küçüksille, Selbas, and Sencan, 2011) who used data mining **to predict thermodynamic properties**, they used many algorithms to predict enthalpy, entropy and specific volume for specific types of refrigerants.

Other researchers focused on **Work Process Optimization and Performance Monitoring**, Water and Power Plant Fujairah (FWPP) in the United Arabic Emirates is a true success story of data mining usage, where more than 4% of of the total consumption have been achieved (Tyagi and Kumar, 2014). Softstat 'which is statistical analysis consultancy group', showed the superiority of data mining tools to traditional approaches like DOE (design of experiments), CFD (computational fluid dynamics). In their research they started by feature selection then apply DM algorithms to get better performance of Flame temperature. Finally recommendations from the model was deployed.

Investigating and solving power plant problems by traditional ways is very expensive, complicated, and time consuming process. Because of this, data mining is recently used to solve many types of problems in different types of power plants, like predicting: power plant yield, failure detection and diagnoses, emission of Nitrogen oxides, power curves and wind speed, and boilers' efficiency. To predict these targets a lot of attributes could be used, like steam properties (pressure, temperature and flow rate) at turbine inlet and outlet, and many others. SCADA systems, which are now a days available by almost any power plant, instantly capture and store huge amount of data which is stored in historical databases. All these databases are available for data mining researches to solve power plant problems.

Although many problems of power plants were solved by data mining, still a lot of researches are needed. Like predicting the efficiency, detecting and diagnosing failure, especially in thermal power plants. Chapter two provides more details about these problems, data mining techniques, what was solved and what is the gap.

1.3 Problem Statement

The fundamental tasks of operation engineers in the power plant is to control the amount of generated power, to guarantee the stability of power supply. There are two methods to expect the amount of generated power; the first one is by using equations of thermodynamic laws. The second by using the Consumption Graph which is prepared by the power plant manufacturers, and gives the amount of generated power in mega watts, as a function of steam flow rate at turbine inlet. The problem is that, over time, the calculated amount of power using these methods is different from the actual amount. So, the problem is stated as two sub-problems:

1. Can we design a feature selection technique, that can determine the best set of features that influence the current amount of power generated from a thermal power plant.?

2. Can we use only the controllable parameters, to accurately predict the amount of generated power from a thermal power plant?

By dividing the problem into above two sub-problems, we could have better understanding. The first sub-problem is studying the full feature set, to study all features without neglecting any feature, and the second part is focusing only in the controllable parameters.

1.4 Objectives of Study

The goal of this research is to to determine all parameters that influence the amount of generated power and to predict accurately the amount of power generated from thermal power plant using data mining techniques. To achieve this goal the following objectives have been specified:

1. Design a feature selection technique, that can determine the best set of features to predict the amount of generated power from a thermal power plant.
2. Design a prediction technique that can accurately predict the amount of generated power, using only the controllable parameters.

Figure 1.1 shows methodology followed to achieve these goals.

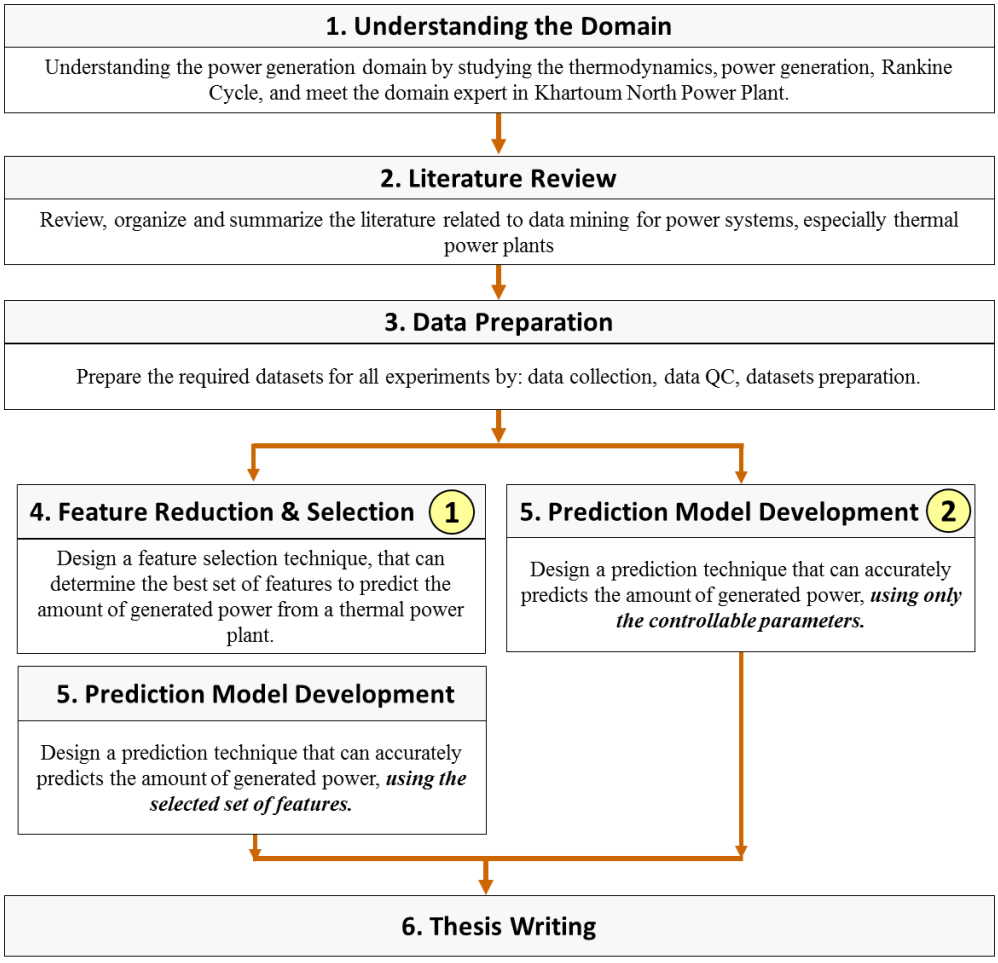


Figure 1.1 Operational Framework

1.5 Scope of Study

The previous section has stated the objectives of this study which focuses on how to produce a good summary using the proposed technique. The following aspects are the scope of research for those objectives.

1. The study focuses only on feature selection and prediction of amount of power generated from thermal power plant, using: first all selected features, then only the controllable parameters.

2. The research uses as a case study, data collected from Khartoum North Power Plant, specifically Unit 3 and Unit 4. Five sets of data were used, three for the first sub problem and two sets for the second sub problem

3. The benchmark used to compare results is the actual amount of generated power. While model evaluation is done by cross validation for small datasets, and separate test sets with bigger sets.

1.6 Significance of Study

The study investigates the prediction of generated power amount from thermal power plant, and the features influence that amount. The significance of this research is to propose methods for power prediction from existing data, using data mining techniques. This research identifies the features that affects the amount of generated power and how much they influence this amount, by ranking the attribute selection results and prediction models that uses these selected features. Using controllable parameters, the research provides accurate prediction models through different prediction algorithms like Linear Regression, Pace Regression and SMOreg. Algorithms gives different results in different datasets. The research reveals the superiority of data mining techniques over thermodynamic laws and manufacturers expectations, in predicting the amount of generated power. Because it depends on real data, while other methods depends on theoretical equations and optimum environment. The performance of the proposed methods is evaluated by cross validation and separate test sets, and compared with actual readings of generated power.

1.7 Expected Contribution

The contributions from this research can be described as follows:

1. Identification of features that influence amount of generated power in thermal power plants.
2. Better understanding of the effects of selected parameters in thermal power plant using feature selection and prediction algorithms.
3. More accurate methods to predict the amount of generated power based on real data collected from the plant through the use of data mining techniques.

1.8 Thesis Organization

This thesis is organized in six chapters to cover the objectives of the study and details of methodologies followed to achieve the goals. These eight chapters are organized as follows:

Chapter 1, *Introduction*: by now the reader already covered this chapter, which presented a general discussion about this research by giving a brief background about the topic. Then the problems was stated, and the research objectives were shown, which is followed by scope and significance of the study. Finally the research contribution of this thesis is summarized.

Chapter 2, *Literature Review*: this chapter presents a comprehensive review of all areas related to this research. The chapter starts by briefing data mining and data mining process (the CRISP-DM mode), then important prediction algorithms are shown. After that various applications, techniques and researches of data mining in power systems and in power plants are presented. The chapter summarizes the literature in simple and clear tables to reveal the variety of researches in this area.

Chapter 3, Research Methodology: this chapter describes the methodology used and the steps followed to achieve the objectives of this research. The chapter presents the data preparation processes, starting from the data collection, until the datasets are ready for machine learning tools. The chapter also describes the datasets by showing the attributes and number of instances for each dataset.

Chapter 4, *Features Selection and Prediction Models using all features*: this chapter gives the research details by showing the methods followed for feature selection, and algorithms used to build the models using the selected set of features. Also the chapter presents the initial comparison between algorithms, the details of each prediction model. Also it shows the evaluation of each model and a comparison of each model result with the actual values using comparison graphs. the chapter also presents a detailed discussion and interpretation of the obtained results.

Chapter 5, *Prediction Models using the Controllable Parameters*: this chapter gives the research details by showing the algorithms used to build the models, the initial comparison between algorithms, the details of each prediction model. The chapter provides accurate models to predict the power using only the controllable parameters. Also it shows the evaluation of each model and a comparison of each model result with the actual values using comparison charts. the chapter also presents a detailed discussion and interpretation of the obtained results.

Chapter 6, *Conclusion and future work*: this chapter presents the research conclusion by highlighting the research contributions and the findings of its work. The chapter also presents suggestions and recommendations for future study.

CHAPTER TWO

LITERATURE REVIEW

Data mining techniques and their applications have developed rapidly during the last two decades. This chapter provides a review of the application of data mining techniques in power systems, specially in power plants, through a survey of literature from 2000 to 2015. Keyword indices and article abstracts and conclusions were used to classify more than ninety articles concerning application of data mining in power plants, from many academic journals and research centers. Because this research is concerned with application of data mining in power plants; this review started by providing a brief introduction about data mining and power systems to give clear vision about these two different disciplines. This review presents comprehensive surveys of the collected articles and classifies them according to three categories: the used techniques, the problem and the application area.

From this review it is proven that data mining could be used to solve many types of problems in power plants, like prediction of generated power, failure prediction, failure diagnosis, failure detection and many others. Also there is no standard technique that could be used for a specific problem. Application of data mining in power plants is a rich research area and still needs more researches.

2.1. Introduction

Most of the electric power systems now a days are equipped with SCADA (Supervisory Control And Data Acquisition) systems, that eases the collection of real

time data. This huge amount of data which is collected instantly encourages the application of data mining techniques in power systems. However, this area is new and still faces difficulties to benefit from data mining (Morais et al, 2009). The first difficulty is that: mining power systems data is an interdisciplinary task, that requires electromechanical engineers and data scientists to work as a team in order to achieve their goals. To handle this difficulty, many meetings and discussions were conducted with domain experts. The Second one is the limitation in data storage capacities, which leads to implementation of automatic purging policies, consequently data is always available for short periods, less than what is required by a data mining tool. This was handled by collecting sample data every month for the last two years, all these samples were combined together to form single dataset for each unit. A third difficulty is the lack of standardized benchmarks, this is very clear from all researches presented here after, where researchers are using proprietary datasets, which makes it difficult to compare algorithms and reproduce results.

Because it is an interdisciplinary research, this research started by giving an introduction about power systems and Rankine cycle for data scientists and computer engineers, on the other hand an introduction to data mining and CRISP-DM model is provided for electromechanical engineers. After that, principles of Feature Reduction and Selection techniques is provided, followed by a review of its application in power systems. Then prediction methods and famous algorithms are presented, followed by a comprehensive review about their application in power systems. After that a discussion is done to lead to research objectives. Finally, this chapter is concluded with a summary.

2.2. Power Systems

A typical power system which is also known as the grid is shown in figure 2.1, it is divided into three components: the generator which produce the power, the transmission system that carries the power from the generators to the load centers, and the distribution which delivers power to the end users.

There are many types of generators (also known as power plant) normally these power plants contain one or more generators which is a rotating machine that converts mechanical power into electrical power. Then the motion between a magnetic field and a conductor creates an electrical current. Most power plants in the world burn fossil fuels such as coal, oil, and natural gas to generate electricity. Others use nuclear power, but there is an increasing use of cleaner renewable sources such as solar, wind, wave and hydroelectric (Morais et al, 2009).

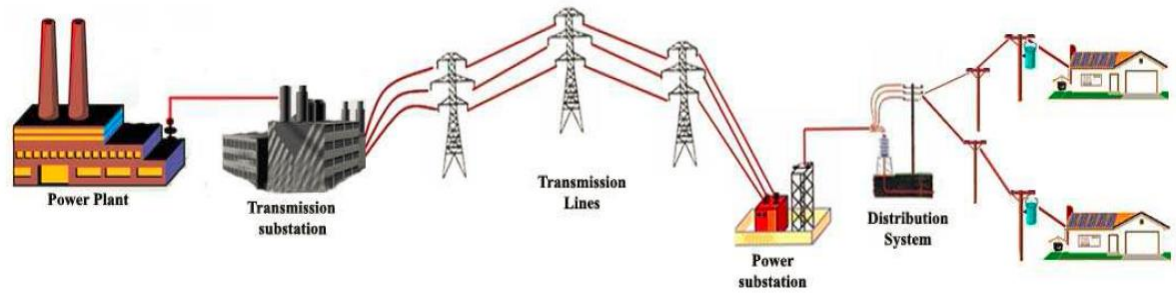


Figure 2.1 Components of power system (Morais et al, 2009).

2.2.1 Carnot Cycle, Rankine Cycle and Thermal Power Plants

The Carnot cycle is a theoretical thermodynamic cycle proposed by Sadi Carnot in 1824. It provides an upper limit on the efficiency that any classical thermodynamic engine can achieve during the conversion of heat into work. Carnot cycle is a theoretical construct that cannot be built in practice (Martínez et al, 2016). The practical implementation of Carnot Cycle is Rankine Cycle, which is applied by thermal power plants to generate power by converting heat into work. This research focus on thermal power plants that uses oil as energy source. More theoretical investigation about Rankine Cycle could be found in (Kapooria et al, 2008). Rankine Cycle is a closed system consists of four main components, that are interconnected together to build one system as shown in figure 2.2. These components are:

1. **Steam Turbine which** uses the superheated steam that is coming from the boiler to rotate the turbine blades.
2. **Condenser**: uses external cooling water to condense the steam which is exhausted from turbine to liquid water.
3. **Feed water Pump**: to pump the liquid to a high pressure and bush it again to boiler.
4. **Boiler** which is externally heated to boil the water to superheated steam.

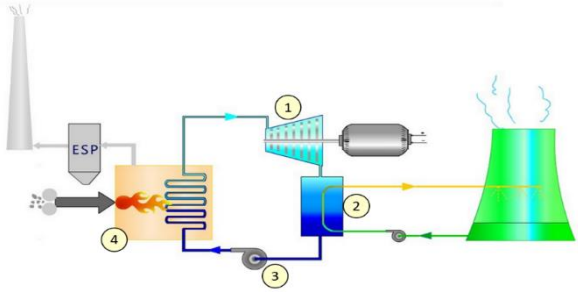


Figure 2.2 Thermal Power Plant using Rankine Cycle (Learn Engineering, 2013)

2.2.2 Power calculation using thermodynamic laws

2.2.2.1 Basic equation of power generation

The amount of generated power from a thermal power plant could be calculated by applying thermodynamic laws, using steam properties at specific points of power plant (turbine inlet turbine outlet). All required steam properties at various locations in power plant are captured by SCADA. Simply Equation (2.1) could be used to calculate the amount of generated power.

$$\text{Generated Power in MW} = \dot{m}_s X \ (h_{in} - h_{out}) \qquad (2.1)$$

Where:

$\dot{m}_s$: is the flow rate of steam at turbine inlet, its value is found in parameter (6) in available dataset.

h_{in}: is the enthalpy at turbine inlet, which could be calculated by applying parameters 8,9 (Main steam header pressure, and T/A inlet steam temperature) values in steam tables and Mollier diagram or enthalpy calculator. Both of these parameter are controllable.

h_{out} : enthalpy at turbine inlet, could be calculated using parameters (29,30) using steam tables and Mollier diagram or enthalpy calculator. ***Parameter 29*** is the steam pressure at turbine outlet and condenser inlet, while ***parameter 30*** represents the temperature of steam at the same point.

In order to be able to control the amount of power generated; we should have full control of the three parameters **$\dot{m}s$, h_{in} and h_{out} .** From the above equation we notice that both **$\dot{m}s$** and **h_{in}** are controllable, while **h_{out}** is not fully controllable.

2.2.2.2 Enthalpy Calculation

Enthalpy is a measurement of energy in a thermodynamic system. It includes the internal energy, which is the energy required to create a system, and the amount of energy required to make room for it by displacing its environment and establishing its volume and pressure (Zemansky and Mark,1968). Enthalpy is defined as a state function that depends only on the prevailing equilibrium state identified by the variables internal energy, pressure, and volume. It is an extensive quantity. The unit of measurement for enthalpy in the International System of Units (SI) is the joule, but other historical, conventional units are still in use, such as the British thermal unit and the *calorie*.

The enthalpy is the preferred expression of system energy changes in many chemical, biological, and physical measurements at constant pressure, because it simplifies the description of energy transfer. At constant pressure, the enthalpy change equals the energy transferred from the environment through heating or work other than expansion work. The total enthalpy, H, of a system cannot be measured directly. The same situation exists in classical mechanics: only a change or difference in energy carries physical meaning. Enthalpy itself is a thermodynamic potential, so in order to measure the enthalpy of a system, we must refer to a defined reference point; therefore

what we measure is the change in enthalpy, ΔH. The ΔH is a positive change in endothermic reactions, and negative in heat-releasing exothermic processes.

Enthalpy is the most important factor to calculate the generated power. Because no sensor can read enthalpy directly, its value should be calculated using pressure and temperature. Enthalpy could be calculated applying steam pressure and temperature values to steam tables (Appendix A), also it could be calculated using Mollier chart (Appendix B). Alternatively some applications that could be found in the internet like **WASP (W**ater **A**nd **S**team **P**roperties), could be used to calculate the enthalpy. Figure 2.3 shows a snap shot of WASP on which you can enter the temperature and pressure, then immediately WASP will give you the steam properties like enthalpy. After getting enthalpy values, the generated power could be calculated easily using equation (2.1).

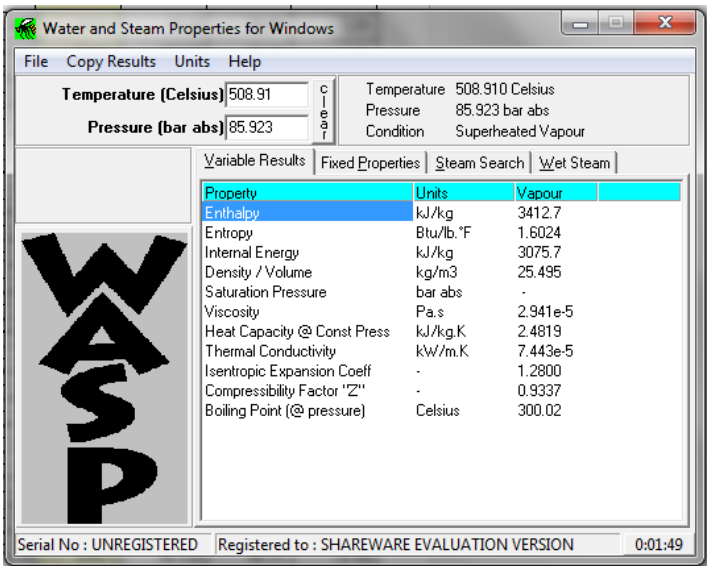


Figure 2.3 Water And Steam Properties

2.2.3 Expected Amount of Generated Power according to Manufacturer

Upon power plant installation, manufacturers provide the Steam Consumption Graph. It is a graph that shows how much power will be generated if steam with specific properties provided to the turbine. Figure 2.4 below shows the steam consumption

graph for Khartoum North Thermal Power Plant, for both Unit 3 and Unit 4. The X-axis shows the amount of power that will be generated in MW, while the Y-axis shows the corresponding steam flow in kg/s, but the steam is supposed to be 510 ˚C temperature and under 87 bar pressure. From the figure 2.4 Steam Consumption Graph, we can simply derived equation (2.2) below.

Output AT Terminal [MW] = Live Steam Flow [kg/s] – 2 **(2.2)**

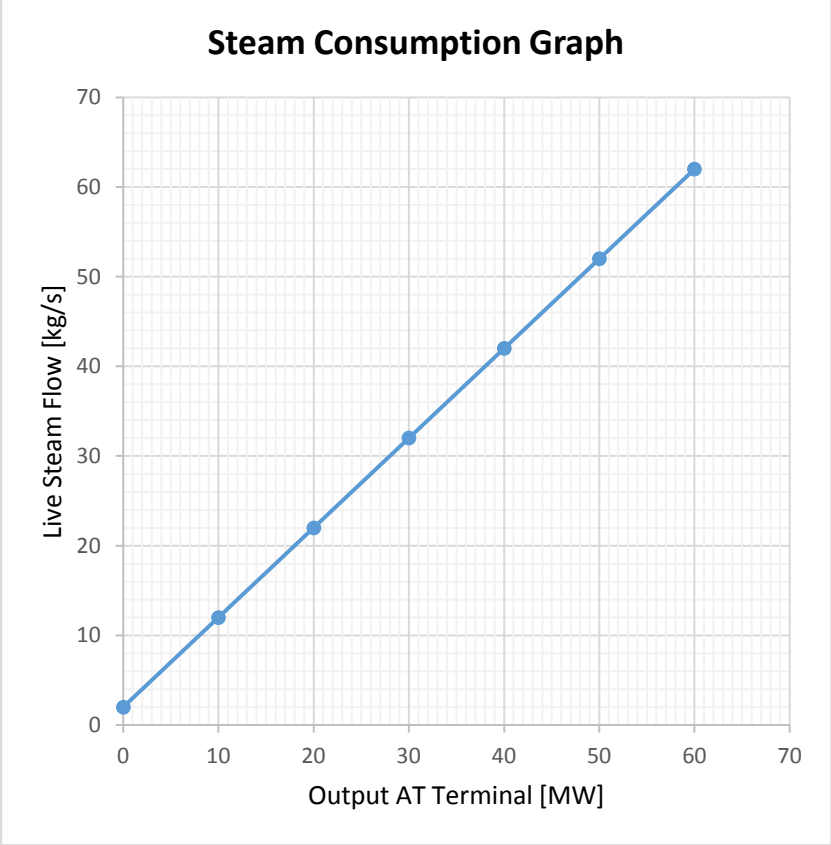


Figure 2.4 Steam Consumption Graph for KNPP for Unit 3&4

When we apply this simple equation to *MainSteamFlow_kg_s* we got the expected power According to *Steam Consumption Graph*. But unfortunately, the power plant can never achieve this value, specially when it becomes old. Efficiency engineers stated that, the power plant can not reach the full capacity even after maintenance. The deviation of *actual* values from the manufacturers' expectation is due to many reasons like degradation of performance of some parts, or any other physical reasons. Figure

2.5 and 2.6 (for Unit 3 and Unit 4 respectively) shows that both *calculated amount* (using thermodynamic laws), and *manufacturer's expected amount* (as per Steam Consumption Graph) are different from the *actual* amount of generated power. This research will show the superiority of data mining techniques to: investigate the reasons behind this degradation, and to predict the power yield in thermal power plants using instances of data collected from two units of a thermal power plant.

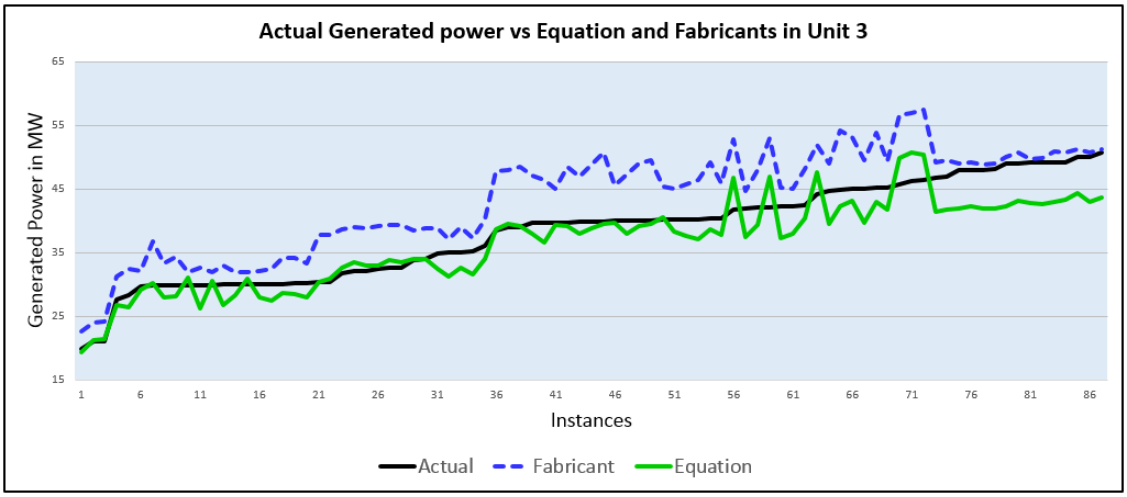


Figure 2.5 Actual power vs Equation and Manufacturer expected values of Unit 3

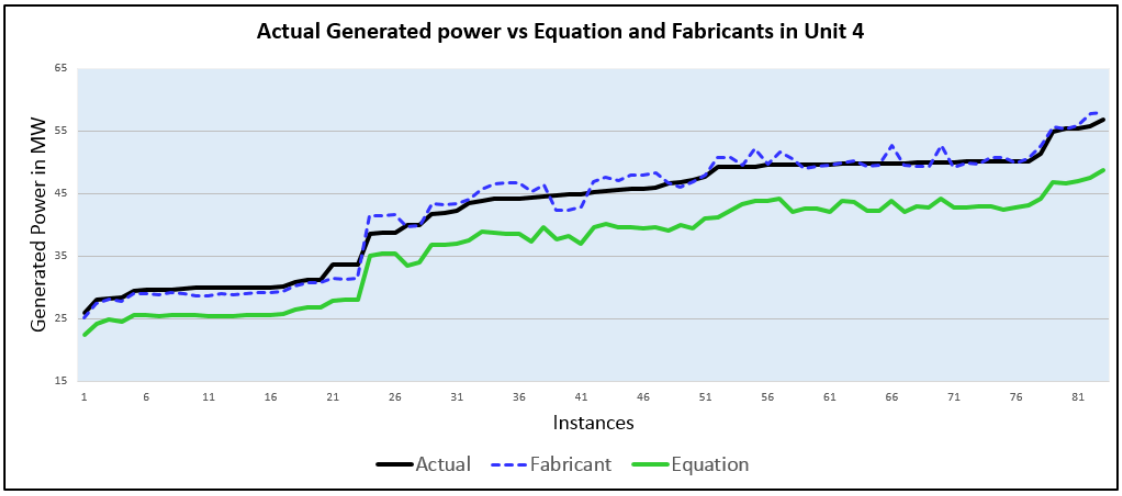


Figure 2.6 Actual power vs Equation and Manufacturer expected values of Unit 4

2.3. Data mining & Data mining Processes

This section is divided into three parts, the first one gives an introduction of data mining, the second is about CRISP-DM which is a complete blueprint for conducting a data mining project. Then data exploration and analysis is presented because it provides a clear method of understanding the datasets, as stated by the CRISP-DM model. Finally a review about building prediction models is provided. The last two parts are illustrated according to data map as shown in figure 2.7, which is organized by Prof Sayed Sayed (Saed, 2017).

2.3.1 Introduction about Data Mining

Data mining is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage. The data is invariably presented in substantial quantities (Witten and Frank and Hall, 2011). Data Mining is explaining the past and predicting the future by means of data analysis. Data mining is a multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database technology. Many businesses have stored large amounts of data over years of operation, and data mining is able to extract very valuable knowledge from this data. The businesses are then able to leverage the extracted knowledge into more clients, more sales, and greater profits. This is also true in the engineering and medical fields (Saed, 2017). This part starts by giving brief introduction about datasets and data types. Then data exploration methods are presented, to show the concepts of exploring the past about data, as this will assist in predicting the future. After that a taxonomy of data mining modeling is provided.

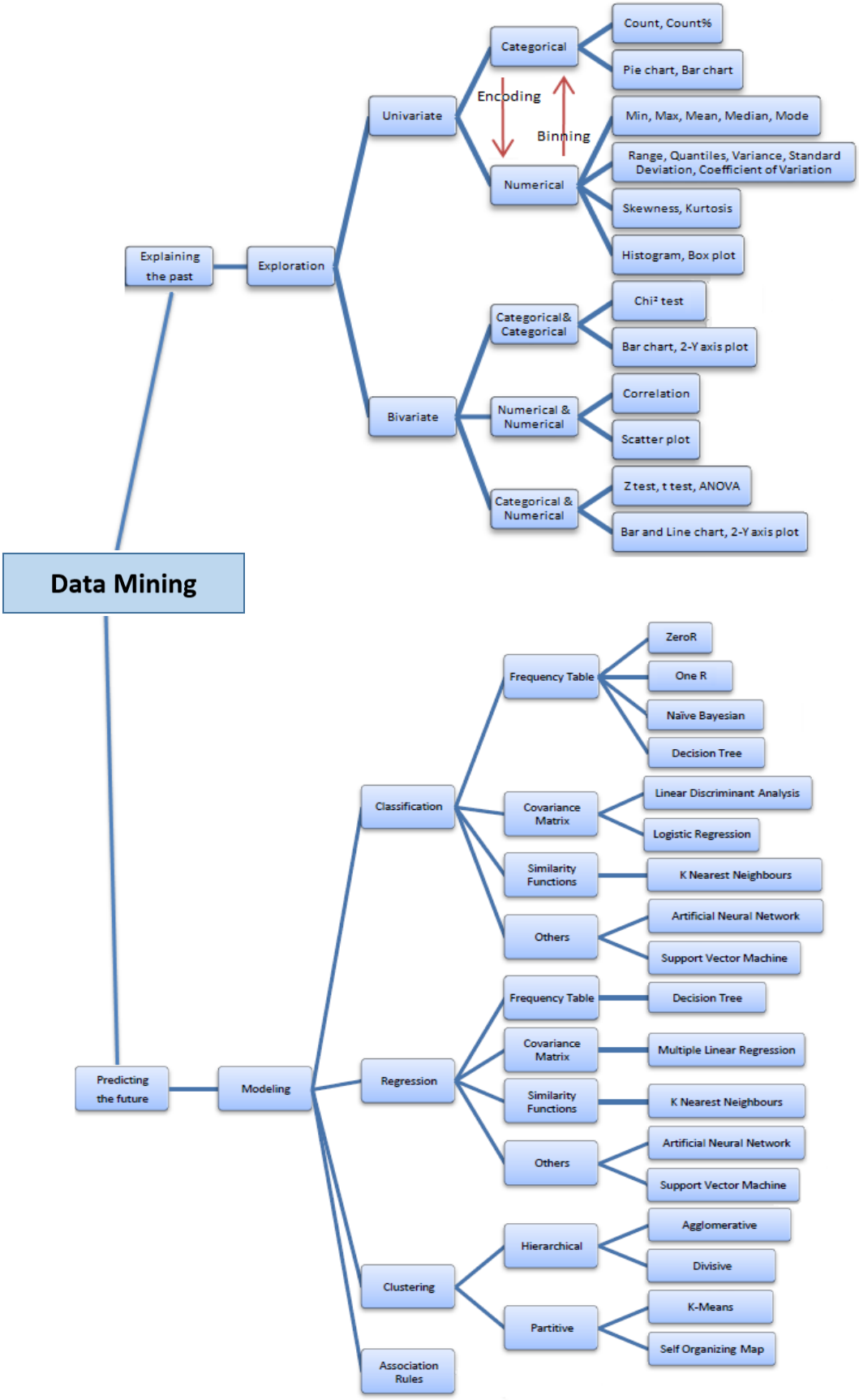


Figure 2.7 Data mining Map (Saed, 2017)

2.3.1.1 Datasets

Dataset: Dataset is a collection of data, usually presented in a tabular form. Each column represents a particular variable (also called feature or attribute), and each row corresponds to a given member of the data (also called records, objects, cases, instances, examples, vectors). There are some alternatives for columns, rows and values. In predictive modeling, predictors or attributes are the input variables and target or class attribute is the output variable whose value is determined by the values of the predictors and function of the predictive model.

2.3.1.2 Data Types for Data Mining

Features have different types that distinguishes the amount of information they encode. Below is a brief review about these data types starting from the "simplest" which carries the least information to those which provide the most information.

1. **Nominal variables**: These are simply labels identifying unique entities. Personal names are nominal labels identifying unique individuals. So too are order numbers, serial numbers, tracking code and many other similar labels.
2. **Categorical variables**: These are group labels identifying groups of entities that shares some characteristics implied by the category. For example group of human, group of palm trees, group of birds, and so on. Some authors look at Nominal and Categorical variables as mutual exclusive, but not ordered.
3. **Ordinal variables**: These are categories that can be listed in some order. Example of this is: small, medium, large or hot, warm, cool. This order is the special about this type, because neither nominal nor categorical variables can be ordered; they are simple unordered labels.
4. **Interval variables**: These are ordinal variables in which it is possible to determine a distance between the ordered categories. Additive distances between equidistant points are meaningful, but ratios aren't. For example temperature, expressed in degrees, it makes sense to talk about the difference between two temperatures, and compare that with the difference between another two temperatures. But we can't say that degree 40 is twice hot than degree 20.

5. **Ratio variables**: these are interval variables in which ratios are valid, and could have a true zero point. An example is the bank account. The zero point is an empty account, and the ratio between 100$ and 200$ is 1 to 2. Any mathematical operations are allowed. It certainly does make sense to talk about three times the distance and even to multiply one distance by another to get an area. All datasets in this research are ratio variable.

It is important in data mining to understand the nature of your data, because that will determine the data mining methods that could be used. Mainly, data either numerical or categorical. Ratio and Interval variables above are both numerical data. While ordinal and nominal are categorical.

2.3.2 The CRISP-DM Reference Model

Cross-industry standard process for data mining (CRISP-DM) is a comprehensive process model and data mining methodology that provides anyone with a complete blueprint for conducting a data mining project (Shearer, 2000). As shown in figure 2.8, the CRISP-DM breaks down the life cycle of a data mining project into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The arrows in the figure indicate the most important and frequent dependencies between the phases, while the outer circle shows the cyclic nature and continual improvement of data mining itself , i.e. lessons learned during the data mining process and from the deployed solution can trigger new, business questions.

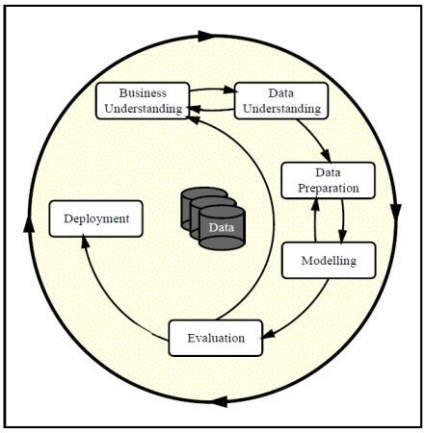


Figure 2.8 Phases of the CRISP-DM (Shearer, 2000)

1. Business Understanding : This is the most important phase, it is the initial business understanding phase which focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. In order to understand which data should later be analyzed, and how. This phase involves several key steps, including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan.

2. Data Understanding : This phase starts with an initial data collection, then the analyst start to increase familiarity with the data, to identify data quality problems, discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. This phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality.

3. Data Preparation : This phase covers all activities to construct the final data set that will be fed into the modeling tool(s) from the initial raw data. This phase consists of five steps : the selection, cleansing, construction, integration, and data formatting.

4. Modeling : In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Several techniques exist for the same data mining problem type. Some techniques have specific data requirements, therefore, stepping back to data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models.

5. Evaluation : In this phase the model is evaluated and its construction is reviewed to be certain it properly achieves the business objectives, and consider all important business issues. At the end of this phase, the project leader should decide how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps.

6. Deployment : The knowledge gained must be organized and presented in a way that the customer can use it, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. Even though it is often the customer, not the data analyst, who carries out the deployment steps, it is important for the customer to understand up

front what actions must be taken in order to actually make use of the created models. The key steps here are plan deployment, plan monitoring and maintenance, the production of the final report, and review of the project.

2.3.3 Data Exploration and Analysis

Data Exploration is about exploring the past, it describes the data by means of statistical and visualization techniques (Saed, 2017). Data is explored in order to bring important aspects of that data into focus for further analysis. This exploration if done for an individual feature it is called Univariate Analysis, and if done for two features it is Bivariate Analysis. The top part of figure 2.7 provides summary of data exploration methods, below are some details about these methods.

2.3.3.1 Univariate Analysis

Univariate analysis explores variables one by one. Variables could be either categorical or numerical. There are different statistical and visualization techniques of investigation for each type of variable. Numerical variables can be transformed into categorical counterparts by a process called binning or discretization. It is also possible to transform a categorical variable into its numerical counterpart by a process called encoding. Finally, proper handling of missing values is an important issue in mining data.

1. **Categorical Variables:** For categorical variable either nominal and ordinal only count function could be used. A frequency table is a way of counting how often each category of the variable in question occurs. It may be enhanced by adding percentages that fall into each category. Graphs (bar chart and pie charts) is a good visualization method that is used to give better understanding of the data.
2. **Numerical Variables:** A numerical variable (interval and ratio) is one that may take on any value within a finite or infinite interval. Many statistical functions could be used with numerical attributes to give better understanding about the

problem. These functions like: Count, Minimum, Maximum, Mean, Median, Mode, Quantile, Range, Variance, Standard Deviation, Coefficient of Deviation, Skewness and Kurtosis. Histogram and Box Plot are good visualization methods that could be used with numerical data.

2.3.3.2 Bivariate Analysis

Bivariate analysis is a simultaneous analysis of two variables. It explores the relationship between the two variables, it explores whether there is an association between them and the strength of this association, or whether there are differences between two variables and the significance of these differences. There are three types of bivariate analysis according to data types:

1. Numerical & Numerical
2. Categorical & Categorical
3. Numerical & Categorical

1. **Numerical and Numerical**: If the two attributes are numerical data exploration could be done by: Scatter Plot or Linear Correlation, below are some details about these methods.
   i. **Scatter Plot** : A scatter plot is a graph that represents the relationship between two numerical variables and is usually drawn before doing a linear correlation or fitting a regression line. The resulting pattern indicates the relationship between the two variables. Figure 2.9 shows a sample of a Scattered Plot graph.

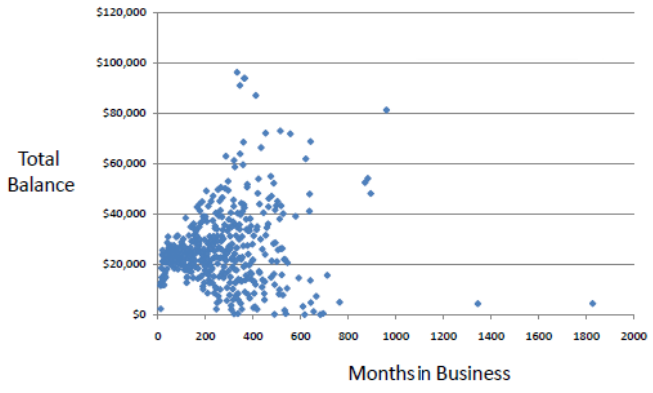


Figure 2.9 Sample of Scattered Plot graph (Saed, 2017)

ii. **Linear Correlation**: Linear correlation quantifies the strength of a linear relationship between two numerical variables. When there is no correlation between two variables, there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity.

$$r = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}} \quad \textbf{(2.3)}$$

$$Covar(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

$$Var(x) = \frac{\sum(x-\bar{x})^2}{n}$$

$$Var(y) = \frac{\sum(y-\bar{y})^2}{n}$$

$r : Linear\ Correlation$

$Covar : Covariance$

$Var : Variance$

In equation (2.3) above, r measures the strength of a linear relationship between attribute x and y. Always r between -1 and 1, where -1 means perfect negative linear correlation, and +1 means perfect positive linear correlation, and zero means no linear correlation.

2. **Categorical and Categorical :** If the two attributes are categorical, data exploration could be done by: Stacked Column Chart Combination Chart, or Chi-square Test, below are some details about these methods.

i. **Stacked Column Chart**: is a useful graph to visualize the relationship between two categorical variables. It compares the percentage that each category from one variable contributes to a total across categories of the second variable. Figure 2.10 shows a sample of Stacked Column Chart.

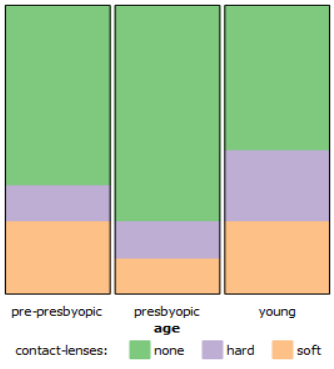


Figure 2.10 Sample of Stacked Column Chart. (Saed, 2017)

ii. **Combination Chart**: A combination chart uses two or more chart types to emphasize that the chart contains different kinds of information. Like using bar chart to show the distribution of one categorical variable and a line chart to show the percentage of the selected category from the second categorical variable. The combination chart is used to demonstrate the predictability power of a predictor (X-axis) against a target (Y-axis). Figure 2.11 shows a sample of a Combination Chart.

iii. **Chi-square Test :** The chi-square test can be used to determine the association between categorical variables. It is based on the difference between the expected frequencies (e) and the observed frequencies (n) in one or more categories in the frequency table. The chi-square distribution returns a probability for the computed chi-square and the degree of freedom. A probability of zero shows a complete dependency between two categorical variables and a probability of one means that two categorical variables are completely independent.

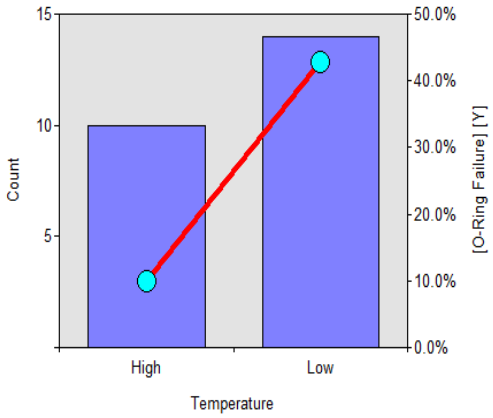


Figure 2.11 Sample of a Combination Chart (Saed, 2017)

3. **Numerical & Categorical:** If the one attribute is numerical and the other is categorical, data exploration could be done by: Line Chart with Error Bars, Combination Chart, Z-test and t-test, or Analysis of Variance (ANOVA), below are some details about these methods.

i. **Line Chart with Error Bars:** this line chart displays information as a series of data points connected by straight line segments. Each data point is average of

the numerical data for the corresponding category of the categorical variable with error bar showing standard error. It is a way to summarize how pieces of information are related and how they vary depending on one another. Figure 2.12 shows a sample of a Line Chart with Error Bars.

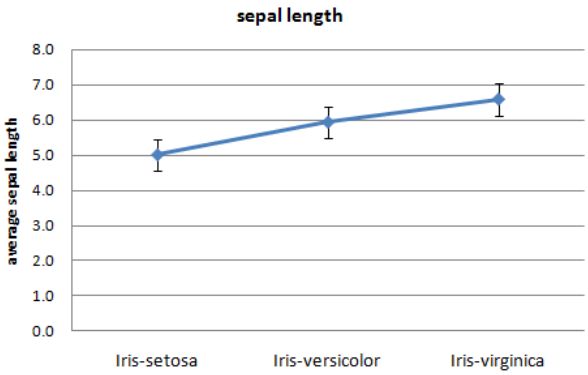


Figure 2.12 Sample of a Line Chart with Error Bars (Saed, 2017)

ii. **Combination Chart:** A combination chart uses two or more chart types to emphasize that the chart contains different kinds of information. Like using bar chart to show the distribution of one categorical variable and a line chart to show the percentage of the selected category from the second categorical variable. The combination chart is used to demonstrate the predictability power of a predictor (X-axis) against a target (Y-axis).

iii. **Z-test and t-test:** These tests are same, they assess whether the averages of two groups are statistically different from each other. This analysis is appropriate for comparing the averages of a numerical variable for two categories of a categorical variable.

iv. **Analysis of Variance (ANOVA):** The ANOVA test assesses whether the averages of more than two groups are statistically different from each other. This analysis is appropriate for comparing the averages of a numerical variable for more than two categories of a categorical variable.

To have a better understanding of the dataset, it is important to precede the creation of data mining model by a data exploration phase. As seen from this section this exploration is statistical analysis, which is done by specific methods according to

the data type and the required analysis, whether it is univariate or bivariate analysis. A known problem in practical data mining, is the poor quality of the data. In real world datasets, errors like unreliable, missing and corrupted data are extremely common. These errors could be noticed by finding odd values like high standard deviation, or observing a point in a graph that is far from the majority of points. Such errors in data will lead to performance degradation, and decrease the accuracy of data mining techniques (Witten and Frank and Hall, 2011). The data exploration will assist to painstakingly check the data quality and get rid of these problems.

2.3.4 Building Prediction Models

After having basic understanding about the datasets by data exploration and analysis phase, and reducing the feature set either by feature extraction or feature selection, the dataset will be ready for developing the prediction model to solve a data mining problem. Predictive modeling is the process by which a model is created to predict an outcome. If the outcome is categorical it is called classification and if the outcome is numerical it is called regression. Descriptive modeling or clustering is the assignment of observations into clusters so that observations in the same cluster are similar. Finally, association rules can find interesting associations amongst observations. According to the data mining task and the data type of the class, a suitable method could be used. Figure 2.7 depicts the main tasks that could be handled by data mining.

1. **Classification**: Classification is a data mining task of predicting the value of a categorical variable (target or class) by building a model based on one or more numerical and/or categorical variables (predictors or attributes). As shown in figure 2.7, classification algorithms are grouped into the following four main groups:
   i. Frequency Table: this group contains ZeroR, OneR, Naive Bayesian, and Decision Tree algorithms.
   ii. Covariance Matrix: this group contains Linear Discriminant Analysis and Logistic Regression algorithms.
   iii. Similarity Functions : this group contains K Nearest Neighbors algorithm.
   iv. Others : this group contains algorithms like Artificial Neural Network and Support Vector Machine.

2. **Regression:** is a data mining task of predicting the value of target (numerical variable) by building a model based on one or more predictors (numerical and categorical variables). As shown in figure 2.7, regression algorithms are also grouped into four main groups, but the algorithms are different:
   i. Frequency Table: this group contains Decision Tree algorithm.
   ii. Covariance Matrix: this group contains Multiple Linear Regression algorithm.
   iii. Similarity Function: this group contains K Nearest Neighbors algorithm.
   iv. Others: this group contains Artificial Neural Network and Support Vector Machine algorithms.

3. **Clustering:** A cluster is a subset of data which are similar. Clustering (also called unsupervised learning) is the process of dividing a dataset into groups such that the members of each group are as similar (close) as possible to one another, and different groups are as dissimilar (far) as possible from one another. Clustering can uncover previously undetected relationships in a dataset. There are many applications for cluster analysis. For example, in business, cluster analysis can be used to discover and characterize customer segments for marketing purposes and in biology, it can be used for classification of plants and animals given their features. Two main groups of clustering algorithms as shown in figure 2.7 are:
   i. Hierarchical: this group contains Agglomerative and Divisive algorithms.
   ii. Partitive: this group contains K Means and Self-Organizing Map algorithms.

4. **Association:** Association Rules find all sets of items (itemsets) that have support greater than the minimum support and then using the large itemsets to generate the desired rules that have confidence greater than the minimum confidence. The lift of a rule is the ratio of the observed support to that expected if X and Y were independent. A typical and widely used example of association rules application is market basket analysis. The main algorithms of association are: AIS, Apriori, AprioriTid and AprioriHybrid algorithms. More details about data mining algorithms could be found in (Wu et al., 2008).

2.3.5 Some Regression Algorithms

Predictors and classes of all datasets in this research are numerical, so only regression could be used to predict the classes. Below is a brief review about common algorithms in regression that were used in similar researches.

. **Linear regression (LR):** Linear regression is a well known method of mathematical modeling of the relationship between a dependent variable and one or more independent variables. Regression uses existing (or known) values to forecast the required parameters. In the simplest case, regression employs standard statistical techniques such as linear regression. Unfortunately, many real world problems are not simply linear projections of previous values. So, more complex techniques (e.g., logistic regression, decision trees or neural networks) may be necessary to forecast future values (Zhou, 2003)

. **Pace regression (PR):** Pace regression improves the classical ordinary least squares regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their contribution to the overall regressions. Under regularity conditions, pace regression is provably optimal when the number of coefficients tends to infinity. It consists of a group of estimators that are either overall optimal or optimal under certain conditions (Witten and Frank and Hall, 2011).

. **Multi layer Regression (MLR):** Multiple linear regression (**MLR**) is a method used to model the linear relationship between a dependent variable (target) and one or more independent variables (predictors). Equation (2.4)

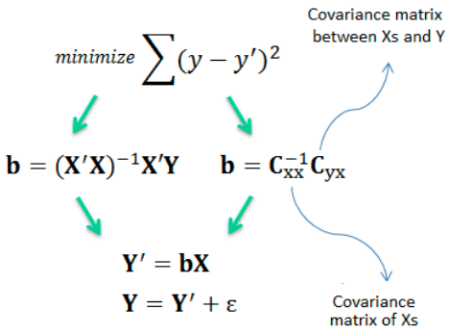
$$\begin{aligned} \textit{observed data} &\rightarrow y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon \\ \textit{predicted data} &\rightarrow y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \\ \textit{error} &\rightarrow \varepsilon = y - y' \end{aligned} \qquad \textbf{(2.4)}$$

Where:

- The subscript *i* refers to the *i*th individual. *X are the variables.* -variables, the subscript following *i* simply denotes which *x*-variable it is.

- The word "linear" in "multiple linear regression" refers to the fact that the model is *linear in the parameters*, $b_0, b_1, \ldots$.

MLR is based on ordinary least squares (OLS), the model is fit such that the sum-of-squares of differences of observed and predicted values is minimized.



The MLR model is based on several assumptions (e.g., errors are normally distributed with zero mean and constant variance). Provided the assumptions are satisfied, the regression estimators are optimal in the sense that they are unbiased, efficient, and consistent. Unbiased means that the expected value of the estimator is equal to the true value of the parameter. Efficient means that the estimator has a smaller variance than any other estimator. Consistent means that the bias and variance of the estimator approach zero as the sample size approaches infinity (Saed, 2017).

. **SMOReg:** SMO (Self-Organizing Maps (SMO) algorithm for regression) implements a non-linear method for sequential minimal optimization to train a support vector regression using polynomial or radial basis function (RBF) kernels. Multi-class problems are solved using pair wise classification. To obtain the proper probability estimates, we use the option that fits logistic regression models to the outputs of the support vector machine (Chiu, 2008).

. **SVMReg (Support Vector Machine Regression):** The SVR model maps data nonlinearly into a higher-dimensional feature space, in which it undertakes linear regression. Rather than obtaining empirical errors, SVR aims to minimize the upper limit of the generalization error (Saed, 2017)

K Star: Is an instance-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. The underlying assumption of instance-based classifiers is such as K Star (Saed, 2017).

. **M5 Model Tree Algorithm:** An algorithm for generating M5 model trees. M5 builds a tree to predict numeric values for a given instance. The algorithm requires the output attribute to be numeric while the input attributes can be either discrete or continuous. For a given instance the tree is traversed from top to bottom until a leaf node is reached. At each node in the tree a decision is made to follow a particular branch based on a test condition on the attribute associated with that node. Each leaf has a linear regression model associated with it. As the leaf nodes contain a linear regression model to obtain the predicted output, the tree is called a model tree. To build a model tree, using the M5 algorithm, we start with a set of training instances. The tree is built using a divide-and-conquer method. At a node, starting with the root node, the instance set that reaches it is either associated with a leaf or a test condition is chosen that splits the instances into subsets based on the test outcome. A test is based on an attributes value, which is used to decide which branch to follow (Witten, Frank and Hall, 2011).

. **REP Tree:** Quinlan first introduced Reduced Error Pruning (REP) as a method to prune decision trees. REP is a simple pruning method though it is sometimes considered to over prune the tree. A separate pruning dataset is required, which is considered a downfall of this method because data is normally scarce. However, REP can be extremely powerful when it is used with either a large number of examples or in combination with boosting. The pruning method that is used is the replacement of a subtree by a leaf representing the majority of all examples reaching it in the pruning set. This replacement is done if this modification reduces the error, i.e. if the new tree would give an equal or fewer numbers of misclassifications (Witten, Frank and Hall, 2011).

. **Decision Table (DT):** Decision table summarizes the dataset with a 'decision table', a decision table contains the same number of attributes as the original dataset, and a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. This implementation employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller, more condensed decision table (Saed, 2017).

2.4. Feature Reduction Techniques

The abundance of data in contemporary datasets requires efficient data mining models to discover the information. Because of the negative effect of irrelevant attributes, it is common to precede the learning process with an attribute selection phase, to select only the relevant attributes. The best way to select these relevant attributes is manually, based on a deep understanding of the learning problem and the attributes themselves. However, automatic methods can also be useful. Hence, the development of these models is preceded by data preparation phase for two reasons: to reduce the dataset size, and to adopt the dataset to the best suit selected analysis method (Witten, Frank and Hall, 2011).

The size of the datasets is more important, because data keeps growing in term of both the number of features and samples. Dataset size reduction can be performed in one of two ways: feature set reduction or sample set reduction (Jović, Brkić and ,2015). In this research the focus will be on feature set reduction. If the number of features is higher than the number of samples in a dataset, this will lead to overfitting, which causes poor results when validating the datasets. Moreover building models using high number of features is more computationally demanding (Korn, Pagel and Faloutsos, 2001). The feature set reduction is performed through the processes of feature extraction and feature selection. In this review, the focus will be on feature selection.

2.4.1 Feature Extraction

Feature extraction creates new features from functions of the original features. Feature extraction works by transforming the original features into a new feature set. feature extraction starts from an initial set of measured data, and builds the derived feature set that is intended to be more informative and non-redundant, to be used in the subsequent steps of learning and generalization (Alpaydin, 2010). The new set is constructed from the original one based on their combinations, with the aim of discovering more meaningful information in the new set (Tang, Alelyani and Liu, 2014).

2.4.2 Feature Selection

Feature selection also known as variable selection and attribute selection, is the process of selecting a subset of features from the original set without transformation, and validating it according to the analysis goal. Feature selection techniques are often used in domains where there are many features and relatively few samples (Witten, Frank and Hall, 2011). Feature selection techniques are used for the following reasons:

. Simplify the models to make them interpretable by users, focus on the target concept and direct the user's attention to the most relevant variables. (Witten, Frank and Hall, 2011) , (James et al, 2013).
. Reduce training time (Witten, Frank and Hall, 2011).
. Improve generalization by reducing overfitting, which in turn leads to poor results on validating datasets (Bermingham et al, 2015), (Jović and Brkić, 2015).
. Reduce the size of the datasets in order to achieve more efficient analysis (Jović and Brkić, 2015).

2.4.2.1 Relevance and redundancy

Feature set reduction is based on the terms of feature relevance and redundancy. More specifically, a feature is usually categorized as: strongly relevant, weakly relevant, irrelevant, and redundant (Yu and Liu, 2004), (S. Alelyani et al, 2013).

- Strongly relevant : is always necessary for an optimal feature subset; if a strongly relevant attribute is removed from the dataset, this will affect the original conditional target distribution (Yu and Liu, 2004).
- Weakly relevant but not redundant: this attribute may not always be necessary for an optimal subset, this may depend on certain conditions.
- Irrelevant : Irrelevant features are not necessary to include in the dataset, and not relevant to the model.
- Redundant: are those that are weakly relevant but can be completely replaced with a set of other features such that the target distribution is not disturbed (L Yu; H. Liu, 2004), (Tang, Alelyani and Liu, 2013).

Redundancy is thus considered in multivariate case, whereas relevance is established for individual features. The target of feature selection process is to maximize relevance and minimize redundancy. It usually includes finding a compact feature subset consisting of only relevant features.

2.4.2.2 Steps to find Sub set

The whole process of finding the feature subset typically consists of four basic steps (Liu and Yu, 2005):

1. Subset generation: a subset is generated according to the state space search strategy.
2. Subset evaluation: after a strategy selects a candidate subset, it will be evaluated using an evaluation criterion to evaluate the performance of the subset. In order to ensure that the optimal feature subset with respect to the goal concept has been found, feature selection method should evaluate $(2^m - 1)$ subsets, where m is the total number of features in the dataset. This is computationally infeasible specially for large m. Therefore, many heuristic methods have been proposed to find a sufficiently good subset.
3. Stopping criterion: after repeating steps 1 and 2 for a number of times depending on the process stopping criterion, the best candidate feature subset is selected.
4. Result validation: the selected subset is then validated on an independent dataset or using domain knowledge, while considering the data mining task.

2.4.2.3 Categories of Feature Selection Algorithms

A feature selection algorithm is a combination of a search technique that proposes the new feature subsets, and an evaluation measure that scores these subsets. The simplest algorithm is to test each possible subset of features, to find the one that minimizes the error rate. This is called the exhaustive search, whish requires high computation power and big memory for big feature sets. The selection of evaluation metric influences the algorithm. The feature selection methods are typically presented in three classes, based on how they combine the selection algorithm and the model building According to evaluation metrics; feature selection algorithms can be distinguished into three main categories: wrappers, filters and embedded methods (Guyon and Elisseeff, 2003), (Witten, Frank and Hall, 2011).

(1) Filter Methods

Filter methods select attributes based on a performance measure regardless of the modeling algorithm, they are based only on general features like the correlation with the variable class. Only after selecting the best set of attributes, the modeling algorithm can use them. Filter methods use a proxy measure instead of the error rate to rank a feature subset. Mainly these measures could be classified into : statistical, information, distance, consistency and similarity measures. Common measures are the mutual information (Guyon and Elisseeff, 2003), the pointwise mutual information, Pearson product-moment correlation coefficient, inter/intra class distance or the scores of significance tests for each class/feature combinations (Yiming and Jan, 1997) .

Filters are less computationally intensive than wrappers, the selected features are not tuned to a specific predictive model. That means a feature set that is selected using a filter method, is more general and gives lower prediction performance than the set which is selected using wrapper method. Additionally Filter methods may select redundant attributes, therefore they are used as a pre-process method. On the other hand because feature set has no assumptions about a prediction model, so it is more useful for exploring the relationships between the features. Many filters provide a feature ranking rather than an explicit best feature subset, and the cut off point in the ranking is chosen via cross-validation.

There are many filter methods described in literature, however not all of them could be used for any data mining task. Therefore, the filter methods could be classified according to data mining tasks: classification, regression or clustering. Some filter methods are only applicable for classification tasks, like Information gain (Hoque, Bhattacharyya and Kalita, 2014), Gain ratio, Chi-square (Witten, Frank and Hall, 2011), Symmetrical (Yu and Liu, 2004), Inconsistency criterion (Liu and Setiono, 1996), Fast correlation-based ,filter (FCBF), and Fisher score (Duda, Hart and Stork, 2012). While others like Correlation Symmetrical are applicable only for regression (Yu and Liu, 2004). Some other filters could be used with both regression and classification like: Minimum redundancy maximum relevance (mRmR) (Tang, Alelyani and Liu, 2014). Correlation-based feature selection (CFS) (Witten, Frank and Hall, 2011), Relief and ReliefF (Sikonja and Kononenko, 2003), Spectral feature selection (SPEC) and Laplacian Score (LS) (Tang, Alelyani and Liu, 2013). And some others only used for clustering, like: Feature selection for sparse clustering (James et al, 2013). Localized Feature Selection Based on Scatter Separability (LFSBSS) (Li, Dong and Hua, 2008), Multi-Cluster Feature Selection (MCFS) (S. Alelyani et al, 2013), Feature weighting Kmeans (Modha and Spangler, 2003), and ReliefC (Dash and Ong, 2011).

Univariate feature filters evaluate a single feature, while multivariate filters evaluate an entire feature subset. The feature subset generation depends on the search strategy, and there are four usual starting points for subset generation: forward selection, backward elimination, bidirectional selection, and heuristic feature subset selection.

1. Forward selection: starts with an empty feature set, then starts adding one or more features to the set.
2. Backward elimination: starts with the full set, then starts removing one or more features from the set.
3. Bidirectional search: starts from both sides simultaneously considering larger and smaller feature subsets.
4. Heuristic selection: generates a starting subset based on a heuristic (e.g. a genetic algorithm), and then explores it further.

Common search strategies which are used with multivariate filters can be categorized into exponential, sequential and randomized algorithms.

1. Exponential algorithms: evaluate a number of subsets that grows exponentially with the feature space size. Sample algorithms of this category are : Exhaustive search and Branch-and-bound.
2. Sequential algorithms: add or remove features sequentially. Algorithms of this category like: Greedy forward selection or backward elimination, Best-first, Linear forward selection, Floating forward or backward selection, and Race search.
3. Random algorithms: incorporate randomness into their search procedure. Algorithms of this category like: Random generation, Simulated annealing, Evolutionary computation algorithms (e.g. genetic, ant colony optimization) and Scatter search (Liu and Motoda, 1998).

(2) Wrapper methods

Wrapper methods use predictive models to evaluate feature subsets. The evaluation is repeated for each subset using the algorithm that is used to develop the predictive model. Thus, for classification tasks a wrapper methods evaluate subsets based on the classifier performance (e.g. Naïve Bayes or SVM) (Bradley and Mangasarian, 1998), (Maldonado, Weber and Famili, 2014). For regression a wrapper will evaluate the subsets based on the performance of a regression algorithm (e.g. Linear regression). While for clustering, a wrapper will evaluate the subsets based on the performance of a clustering algorithm (e.g. K-means) (Kim, Street and Menczer, 2002). Each new subset is used to train del, which is tested on a hold-out set. The score of the subset is given by counting the number of mistakes made on that hold-out set.

Because wrapper methods train a new model for each subset, they provide best performing feature set for that specific model because the subsets are evaluated using a real modelling algorithm (Witten, Frank and Hall, 2011). On the other hand wrappers are computationally intensive, they are much slower than filters, because they run a modelling algorithm for each subset. The feature subsets are also biased towards the modelling algorithm on which they were evaluated. Therefore, for a reliable generalization error estimate, it is necessary that both an independent validation sample and another modelling algorithm are used after the final subset is found. In this research wrapper method is used, because of this there is no separate step for feature selection, because it is wrapped with the prediction model in one step.

(3) Embedded Methods

Embedded methods perform feature selection as part of the model construction process. First, a filter method is used to reduce the feature space (Das, 2001), then a wrapper is employed to find the best candidate subset. Therefore, these methods are embedded in the algorithm, either as its normal or extended functionality. Common embedded methods include various types of decision tree algorithms: CART, C4.5, random forest (Sandri and Zuccolotto, 2006), and other algorithms like: multinomial logistic regression and its variants (Cawley, Talbot and Girolami, 2007). Some embedded methods perform feature weighting based on regularization models with objective functions that minimize fitting errors and in the mean time force the feature coefficients to be small or to be exact zero. These methods which are based on Lasso (Bach and Francis, 2008) or Elastic Net (Zou and Hastie, 2005), usually work with linear classifiers (SVM or others) and stimulate penalties upon features that do not contribute to the model.

(4) Hybrid Methods

Hybrid methods combines the best characteristics of filters and wrappers. Hybrid methods achieve high accuracy of wrappers and high efficiency of filters. Any combination of filter and wrapper can be considered as hybrid methodology. Several other methodologies were recently proposed, such as: hybrid genetic algorithms (Oh, Lee and Moon, 2004), hybrid ant colony optimization (Ali and Shahzad, 2012), fuzzy random forest based feature selection (Cadenas, Garrido and Martínez, 2013).

2.5. Prediction used for Power Systems and power Plants

The availability of real time data in the electric power industry encourages the adoption of data mining techniques. Many papers were found in the literature, each is focusing on one area of the power system. Some are focusing on the **Distribution System** like (Ramos, 2008) who used decision tree to classify the consumers. (Saibal, 2008) used WN (Wavelet Networking is an extension of perceptron networks) for the Classification of transients. (Figueiredo, 2005) used Decision tree for the Classification Electric energy consumer. (Dola, 2005) used Decision tree and neural network for Faults classification in distribution system. (Mori, 2002) used Regression tree and neural network for Load forecasting. Other researcher focused on the **Transmission Line** Problems like: (Hagh, 2007), (Silva, 2006), (Costa, 2006), (Vasilic, 2002) all of them used Neural network to study Faults detection, classification and locations in Transmission Lines. Dash, 2007 used Support Vector Machine for the classification and identification of series compensated. (Vasilic, 2005) and (Huisheng, 1998) used Fuzzy/ neural network for faults classification. Some other researchers focused on **Power Generation** part (power plants) like (Kusiak, Zhang and Li, 2011) who used a multi objective optimization model to optimize wind turbine performance. Other researchers focused on Work Process Optimization and Performance Monitoring. Water and Power Plant Fujairah (FWPP) in the United Arabic Emirates is a true success story of data mining usage, where more than 4% of of the total consumption have been achieved (Himani Tyagi and Rajat Kumar, 2014). Tomas Hills showed the superiority of data mining tools to traditional approaches like DOE (design of experiments), CFD (computational fluid dynamics). In his research they started by feature selection then apply DM algorithms to get better performance of Flame temperature. Finally recommendations from the model were deployed.

This section first presents some researches about the application of data mining methods in power systems as general. Then the various types of problems in the different types of power plants.

2.5.1 Applications of data mining methods in the Power Systems

Various data mining methods have been used by many researchers for different types of problems in power systems like: Energy efficiency, HVAC systems, Energy demand modeling, Electricity price forecast, Prediction of properties of refrigerants, Cluster of load profiles, Modeling of absorption heat transformer and more. Table (2.1) shows a summary of some researches that used data mining techniques in different power systems. This section gives an overview of some of these researches grouped by the categories of data mining tasks.

i. Classification

Classification is a data mining task of predicting the value of a categorical variable (target or class) using one or more numerical and/or categorical variables (predictors or attributes). Classification is intensively used to solve different types of problems in Power Systems, like: price forecast, energy demand, consumer characterization, fault diagnoses and detection, and many other types of problems.

(Amooee, M-Bidgoli, and B-Dehnavi, 2011) used many classification algorithms to predict the failure of industrial machinery and minimize the consequences of such failures. Among these many algorithms C5.0 is found to achieve the highest classification accuracy. An example of generated prediction rules by C5 model is as follows:

1) Mold temperature =< 325.500 and hardness <= 82 and distance between sensitive point and umbilical =< 23.95 then the part is normal.
2) Mold temperature > 325.500 and hardness > 82 and distance between sensitive point and umbilical =< 23.200 then the part is defective

Because of the complexity of problems in power systems, it is common to build a model that is composed of many modules, each of which is using a different data mining method. Like (Figueiredo, Rodrigues and Gouveia, 2005) who combined clustering and classification to present an electricity consumer characterization framework.

Table 2.1 Summary of Data Mining used for Power Systems and Engineering

DM Task Category	Data mining Method	What is Predicted	Accuracy	Researcher
Classification	C 5.0	Predict the failure of industrial machinery.	92%	(G. Amooee, et al, 2011)
		Inference of a rule set to characterize each class, and support the the classes obtained by the load profiling module.	81%	(Figueiredo et al, 2005)
	Bayesian	Predict the normal price and the price spikes	Less than 2–5% forecast error	(Lu et al , 2005)
	Decision tree	Predict energy demand model using	92%	(Yu et al. , 2010)
Regression	Decision tree and neural networks	Predict Electricity consumption		(Tso and Yau, 2007)
	LR, PR, sequential minimal optimization (SMO), M5 tree, M5'Rules and back propagation neural network (BPNN)	Predict Volume values of methanol/LiBr and methanol/LiCl		(A. Şencan, 2007)
	Data mining approach	Predict Boiler efficiency and analyze relationships between parameters of a circulating fluidized-bed boiler.		(Kusiak et al, 2005)
	Many data mining algorithms	Prediction of thermodynamic properties of alternative refrigerants		(Küçüksille et al. , 2011)
	Multiple-linear perceptron (MLP)	Minimize the energy of air condition		(Kusiak et al.,2010)
Clustering	k-means algorithm	Creates a set of consumer classes		(Figueiredo et al, 2005)
Others	Rough set and an artificial neural network	Detects and diagnose sensor faults based on the past running performance data in heating, ventilating and air conditioning (HVAC) systems		(Hou et al. , 2006)

This framework consists of two modules: The first one is the load profiling module which creates a set of consumer classes using a clustering with k-means algorithm to represent the load profiles for each class. The second is the classification module which used the C5.0 algorithm. This algorithm was selected because it creates robust models and does not require long training times to estimate so it presents good performances with large data sets.

(Lu, Dong and Li, 2005) carried out electricity price forecast framework, they used Bayesian algorithm to build a classification model to predict the normal price and the price spikes. The model is based on a mining database including market clearing price, trading hour, electricity demand, electricity supply and reserve. The model can generate forecasted price spike, level of spike and associated forecast confidence level.

(Yu et al, 2010) built energy demand model using decision tree. This model applied to Japanese residential buildings for predicting and classifying building levels. The results have demonstrated that the use of decision tree method can classify and predict building energy demand levels accurately (93% for training data and 92% for test data).

ii. Regression

Regression is a data mining task for predicting the value of numerical variable (target or class), by building a model based on one or more numerical and/or categorical variables (predictors or attributes). Most of the data (both targets and predictors) generated by power systems sensors are numerical. Because of this regression is intensively used in prediction problems in power systems when the target is numeric, like: electricity consumption, volume, pressure, temperature, amount of power, efficiency and many others.

(Tso and Yau, 2007) have used regression analysis, decision tree and neural networks models to predict electricity consumption. Model with least squared errors were selected. In an electricity energy consumption study, the decision tree and neural network models outperformed the stepwise regression model in understanding energy consumption patterns and predicting energy consumption levels. Using data mining approach for predictive modeling, different types of models can be built in a unified platform: to assess, select and implement the most appropriate model. **(Şencan, 2007)** applied DM process to determine specific volume values of methanol/LiBr and methanol/LiCl used in absorption heat pump systems. Six

algorithms were used: Linear regression (LR), pace regression (PR), sequential minimal optimization (SMO), M5 model tree, M5'Rules and back propagation neural network (BPNN). (Kusiak, Burns and Milster 2005) applied data mining approach to analyze relationships between parameters of a circulating fluidized-bed boiler. The model can predict efficiency to the same degree of accuracy with and without the data describing the fuel composition or boiler demand levels. It is proved that data mining is applicable to different types of burners and fuel types.

(Küçüksille, Selbas, and Sencan, 2011) used data mining to predict the *thermodynamic properties of refrigerants*, they followed the CRISP-DM to build their models, their results presentation is very clear. In their research they studied four alternative refrigerants R134a, R404a, R407c and R410a. The results obtained from data mining have been compared to actual data from the literature. In their study they showed that, data mining is successfully applied to determine enthalpy, entropy and specific volume values of refrigerants when using temperature and pressure as predictors. They used 12 algorithms to predict 3 different targets (enthalpy, entropy and volume) . So the total number of models is (4 X 2 X 12 X 3= 288 models). Selected algorithms are LR, MLP, PR, SMO, SVM, KStar, AR, RD, M5 model tree, RepTree, DT, M5'Rules. To compare between all these different modeling techniques they used three measures: the correlation coefficient (R2-value), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). In addition, mathematical equations in order to calculate enthalpy, entropy and specific volume values of each refrigerant were presented. The values calculated from obtained formulations were found to be good compared to actual values. The results showed that data mining is suitable for predicting thermodynamic properties of refrigerants for every temperature and pressure.

(Kusiak, Li and Tang 2010) presented data-driven approach to minimize the energy of air condition. Eight algorithms were applied to model the nonlinear relationship among controllable parameters (supply air temperature and supply air static pressure), and uncontrollable parameters. The multiple-linear perceptron (MLP) outperforms other models, so it is selected to model a chiller, a pump, a fan, and a reheat device. These four models are integrated into an energy optimization model with two decision variables. The optimization results have demonstrated the

total energy consumed by the heating, ventilation, and air-conditioning system is reduced by over 7%.

iii. Clustering

A cluster is a subset of data which are similar. Clustering (also called unsupervised learning) is the process of dividing a dataset into groups such that the members of each group are as similar (close) as possible to one another. Clustering is used to solve special types of problems in power systems, like: clustering electricity consumers according to their behavior. (Figueiredo, Rodrigues and Gouveia, 2005) presented an electricity consumer characterization framework based on a knowledge discovery in databases, supported by clustering data mining techniques. This framework consists of two modules: The first one is the load profiling module which creates a set of consumer classes using clustering and the representative load profiles for each class. The second is the classification module which uses this knowledge to build a classification model to assign different consumers to the existing classes.

iv. Others

Other researchers used different Statistical or Artificial Intelligence methods to solve power systems' problems, like fault diagnostic or detection. (Hou et al., 2006) combined rough set and an artificial neural network to developed a model that detects and diagnose sensor faults based on the past running performance data in heating, ventilating and air conditioning (HVAC) systems. The reduced information is used to develop classification rules and train the neural network to infer appropriate parameters. Results from a real HVAC system showed that only the temperature and humidity measurements can work very well as the to distinguish simultaneous temperature sensor faults of the supply chilled water (SCW) and return chilled water (RCW).

It is very clear from these researches; there are so many types of problems of power systems that could be better solved by data mining. Different prediction algorithms could be used, there is no restriction for selecting an algorithm, while it gives accurate results. Some researchers like (Küçüksille, Selbas, and Sencan, 2011) used many algorithms then selected the one that showed better results, while others selected only one algorithm for their problems like (Hou et al, 2006) who used neural network.

2.5.2 Data-Driven Performance Optimization of Wind Farms

Wind energy is a green energy source and does not cause pollution, due to this fact it received much attention recently. One of the weakest points in wind power generation is the low predictive accuracy of the energy output. Many research in the literature were reviewed, some of them are modeling the performance of individual wind turbines, while others look at their collection (a wind farm). A solution for prediction of wind farm performance should be able to predict the amount of energy to be produced on different time scales, e.g., 15 min, 2 hour, a day, and so on. The basic methodology used is this research is data mining**,** because this new industry generates huge amount of data that have not yet been explored. Table 2.2 provides a summary of some researches that used data mining techniques to solve different problems in wind power plants. More similar researches could be found in the research project final report at (Iowa Energy Center, 2017). This section gives an overview of some of these researches grouped by the categories of data mining tasks.

i. Regression

Regression is properly used for prediction problems in wind power plants. (Kusiak, A., Zheng, A. and Song, H., 2009) developed time series models for predicting the power of a wind farm at different time scales, i.e., 10-min and hour-long intervals have been developed with data mining algorithms. Five different data mining algorithms have been tested on various wind farm datasets. Two of the five algorithms performed particularly well. The support vector machine regression algorithm provides accurate predictions of wind power and wind speed at 10-min intervals up to 1 h into the future, while the multilayer perceptron algorithm is accurate in predicting power over hour-long intervals up to 4 h ahead. Wind speed can be predicted fairly accurately based on its historical values; however, the power cannot be accurately determined given a power curve model and the predicted wind speed. Test computational results of all time series models and data mining algorithms are discussed. The tests were performed on data generated at a wind farm of 100 turbines.

Regression is also used for wind power turbine optimization. In optimization problems it is common to divide the work into sub tasks. (Kusiak, Zhang and Li, 2011) developed a regression model to optimize wind turbine performance. This is done by three objectives, maximization of the power produced by a wind turbine, and minimization of vibrations of the turbine's drive train and tower. Data for this research was obtained from a 150 MW wind farm, more than 120 parameters were available, however, only parameters related to wind turbine vibrations and their power output were selected according to the domain experts advice. Some of parameter like vibrations due to the air passing through the wind turbine were non controllable, while Others like vibrations caused by forces originating with the control system that affect the torque and the blade pitch angle are controllable. Because they are regression models, evaluation is done using MAE, SD of MAE, MAPE and SD of MAPE. More over a histogram is used to present results, that makes the model evaluation easier and meaningful.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{5}$$

$$\mathrm{SDofMAE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(|\hat{y}_i - y_i| - \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i|\right)^2} \tag{6}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{\hat{y}_i - y_i}{y_i}\right|\right) \times 100\% \tag{7}$$

$$\mathrm{SDofMAPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{\hat{y}_i - y_i}{y_i}\right| - \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_i}{y_i}\right|\right)^2} \times 100\% \tag{8}$$

Where:

$\hat{\mathbf{y}}_\mathbf{i}$ and $\mathbf{y}_\mathbf{i}$ are the predicted and observed **i**th instance respectively, and n is the total number of instances. **MAE** is Mean Absolute Error, **SD of MAE**: Standard Deviation of Mean Absolute Error, **MAPE**: Mean Absolute Percentage Error, **SD of MAPE**: Standard Deviation of Mean Absolute Percentage Error.

Table 2.2 Summary of Data Mining techniques used for Wind Power Plants

DM Task Category	Data mining Method	What is Predicted	Part	Accuracy	Researcher
Regression	SVM, MLP and others	Wind Speed and Power Prediction	Wind farm		A. Kusiak, H.-Y. Zheng, and Z. Song, 2009
	NN	Optimize performance of wind turbine by predicting: drive train acceleration ,Tower acceleration, and Generated power.	Tower vibration, Wind turbine	Prediction of drive train acceleration 99% Prediction of Tower acceleration 97% Prediction of generated power 97%	Kusiak et al, 2010
Classification/ Regression	Various data-mining algorithms	Condition Monitoring and Fault detection ((1) fault and no-fault prediction; (2) fault severity; and (3) the specific fault prediction)	Wind turbine subsystems		A. Kusiak and W.Y. Li, 2011
Other	Evolutionary strategy algorithm, least squares method , k-NN	Monitoring Power Curves	Wind turbines		A. Kusiak, H.-Y. Zheng, and Z. Song, 2009

ii. Classification & Regression

To achieve research objectives and develop a robust solution, many different methods and algorithms may be used together. **(A. Kusiak and W.Y. Li, 2011)** used Classification and Regression to developed a condition monitoring solutions to detect and diagnose abnormalities of various wind turbine subsystems with the goal of reducing operations and maintenance costs. Various data-mining algorithms have been applied to develop models predicting possible faults. This research explores fault data provided by the supervisory control and data acquisition system, and offers fault prediction at three levels: (1) fault and no-fault prediction; (2) fault category (severity); and (3) the specific fault prediction, the first two models are classification, the last one I a regression model. For each level, the emerging faults are predicted 5-60 min before they occur.

iii. Others

Other Artificial Intelligence algorithms and statistics methods were also used in wind turbine power plants to monitoring Power Curves. (Kusiak, Zheng and Song, 2009) used Evolutionary strategy algorithm, least squares method and k-NN algorithms to analyze the performance of wind turbines. Turbine performance is captured with a power curve. The power curves are constructed using historical wind turbine data. Three power curve models are developed, one by the least squares method and the other by the maximum likelihood estimation method. The models are solved by an evolutionary strategy algorithm. The power curve model constructed by the least squares method outperforms the one built by the maximum likelihood approach. The third model is non-parametric and is built with the k-nearest neighbor (k-NN) algorithm. The least squares (parametric) model and the non-parametric model are used for on-line monitoring of the power curve and their performance is analyzed.

2.5.3 Data mining for Power Plants

This section presents the use of data mining to solve many types of problems in different types of power plants. Some researches have used data mining for power plant optimization, others used classification to detect and classify failure types, some are using regression to predict Nox emission and other targets.

(Huang, Qi and Liu, 2006) indicated that the efficiency and availability depend on reliability and maintainability, to raise efficiency, the equipment of thermal power plants is becoming larger and more complex. However, due to lack of manpower and information resources, the diagnosis and repair of failed equipment cannot be done immediately. Identifying the failure types of steam turbines and their root causes is time consuming, and requires professional knowledge in materials and mechanical engineering. Actually, thermal power plant engineers can only handle routine and direct maintenance tasks. Complex faults require intervention from technical support and equipment manufacturers. These types of tasks are very expensive and require special experiments, which leads to long downtime and causes production losses (Yang and Liu, 2004). The concept of e-maintenance has been introduced to mitigate all these problems and easily identify the root cause of failures, also it reduces the failures of production systems, eliminate unscheduled shutdown maintenances, and improves productivity (Iung and Marquez 2006). Data mining techniques are the core of such intelligent systems and can greatly enhance their performance. Recently, several data mining techniques such as artificial neural networks, fuzzy logic systems, genetic algorithms, and rough set theory have all been employed to assist the detection and condition monitoring tasks in power plants.

This section begins by researches that used Regression, then classification, a new method of visual data mining which is used by (Fazullula, Praveen and Reddy, 2014) and (Prasad, Swidenbank, Hogg, 1999) is also presented. Then some researches that combined Inference system and Classification together to solve one problem. Moreover models used Statistical Analysis and clustering methods are presented. Some researchers used one algorithm to solve the problem, while others compare the performance of many algorithms then select the one that shows the best test results. Table 2.3 provides a summary of these researches.

i. Regression

Regression is used in thermal power plants to predict different types of numeric targets. (Ilamathi, Selladurai and Balamurugan, 2012) used ANN to build a regression model that predicts nitrogen oxides emission from a 210 MW coal fired thermal power plant. The coal combustion parameters (oxygen concentration in flue gas, coal properties, coal flow, boiler load, air distribution scheme, flue gas outlet temperature and nozzle tilt) were used as inputs and nitrogen oxides as output of the model. Values predicted by ANN model were verified with the actual values.

ii. Classification

Classification is used in thermal power plants to predict different types of categorical targets, so it is suitable to solve many types of problems like fault detection and diagnosis. (Chen et al, 2011) proposed a SVM based model to predict failures of turbines in a thermal power plant. In order to handle the huge amount of collected data, they started by feature selection techniques to eliminate irrelevant and noisy data, then they built their model based on the new clean dataset. To evaluate the effectiveness of their model, they used a real-world data from a thermal power company. Their SVM model can successfully detect the types of turbine faults with a high degree of accuracy (greater than 90%). Their method can assist the power plant engineers to find failure types without referring to the manufacturers.

SVM is used by many researcher as it showed higher accuracy in classification. (Mahadevan and Shah, 2009) used one-class SVM for fault detection and diagnosis, the claimed that their approach outperformed principal components analysis (PCA) and dynamic principal components analysis (DPCA).

Table 2.3 Summary of Data Mining used for Power Plants

DM Task Category	Data mining Method	What is Predicted	Researcher
Regression	ANN	Predicts nitrogen oxides emission	Ilamathi P, et al, 2012
Classification	SVM,BPN, LDA	Detects types failures of turbines in a thermal power plant	Kai-Ying Chen et al, 2011
	SVM	Fault detection and diagnosis	Mahadevan and Shah, 2009
Visual Data Mining	AUTORULE, a visual analytics software	Identify some of the possibilities where Nox could be high.	Md Fazullula, et al, 2014
	Histogram based method	Maximize & monitor the performance of thermal power plants.	Prasad et al., 1999
Inference system, Classification	Interactive data mining approach , NN	Failure inspections	Shu, 2007
Statistical Analysis	(PCA) and T2 statistics	Inspect different types of faults	Huang et al., 2006
Statistical, Clustering	Kernel independent component analysis (KICA, for non-Gaussian distribution) and Kernel principal component analysis (KPCA, for Gaussian distribution)	Fault detection	Y. Zhang, 2009
Other	Partial least squares (PLS), SVM	Increase the performance of on-line fault detection in batch processes	Li et al., 2006

iii. Visual Data Mining

The basic idea of Visual Data Mining is to present the data in a visual format, to get better understanding of data, draw conclusions, and directly interact with the data. Visual data mining techniques have proven to be of high value in exploratory data analysis, they also have a high potential for exploring large databases. Moreover these techniques provide a much higher degree of confidence in exploration findings. (Fazullula, Praveen and Reddy, 2014) applied in their research a visual data mining technique with parallel coordinates to a Thermal Power Plant data. AUTORULE, a visual analytics software, was used to process dataset which was collected at different loads. Sixteen predictors were considered for this research *(Coal Speed , Coal Flow, Primary Air Flow, Secondary Air Temperature, Burner Tilt, Induced Draught, Forced Draught, Flue Gas Temperature, Un burnt Oxygen, Sulphur Di Oxide, Nitrogen Di Oxide, Carbon Di Oxide , Date, Time, Total Air Flow, Primary Air Temperature)* to identify some of the possibilities where Nox could be high. Any record that had a Nox greater than 250 was considered to be as higher class and below 250 as lower class. As a result of their research, they found that when the total airflow is low, NOX is high. This was surprising as the NOX production is possible when O_2 and N presence in air is high. i.e. when the total airflow is high. So, these results were submitted to domain expert for further investigation to identify the possible causes for this phenomenon.

(Prasad, Swidenbank, Hogg, 1999) proposed a histogram based method to monitor and maximize the performance of thermal power plants. Therefore, building an intelligent system for the fault prediction of turbines in thermal power plants.

iv. Inference system, Classification

Inference systems are also used for complicated problems in nuclear power plants. (Shu, 2007) established an interactive data mining approach based inference system to solve the basic technical challenge and speed up the discovery of knowledge in nuclear power plant. An artificial neural network method, is evolved by adding a detecting and retraining function for handling complicated nuclear power plant quake-proof data. Based on proposed approach, an information inference system has been developed. To demonstrate how the proposed technique can be used as a

powerful tool for inferring of structural health status in unclear power plant, quakeproof testing data have been applied.

v. Statistical Analysis

(Huang, Qi and Liu, 2006) used principle component analysis (PCA) and T2 statistics to inspect different types of faults in a thermal power plant.

vi. Statistical, Clustering

In the work of (Zhang, 2009), both kernel independent component analysis (KICA, for non-Gaussian distribution) and kernel principal component analysis (KPCA, for Gaussian distribution) are used for fault detection in, named Tennessee Eastman process, which is a complex non-linear process created by Eastman Chemical Company.

vii. Other

(Li, Wang and Yuan, 2006) combined another dimension reduction method, partial least squares (PLS), with SVM to increase the performance of on-line fault detection in batch processes.

2.5.4 Boiler Efficiency

Boiler is the primary source of energy in the power plants. Therefore, the efficiency of combustion is crucial to the performance of boilers; many researchers focused in this area. The intelligent control approaches in combustion can be grouped into three categories: rule-based expert systems, soft computing, (i.e., neural networks, fuzzy logic, and evolutionary computation) and the the third one is hybrid systems which combines analytical modeling with the soft-computing methods. The intelligent control concept can be extended by incorporating data-mining algorithms. Below is summary of some work related to the efficiency of boilers using data mining. Table 2.4 shows a summary of reviewed literature related to Boiler Efficiency. This section begins by presenting Regression models, then researches that combined Clustering and regression. Some Association models are then presented, after that some models that used Genetic data mining, then Optimization-Based Approach.

Finally other researches that used different methods like Genetic wrapper approach, Expert system using fuzzy logic and rough set theory are presented.

i. Regression

(Zhou et al., 2012) Showed a modelling Nox emission from coal fired utility boiler is critical to develop predictive emissions monitoring system (PEMS) and to implement combustion optimization software package for low Nox combustion. Hao Zhou presented an efficient NOx emissions model based on support vector Regression (SVR), and compares its performance with traditional modelling techniques ,like back propagation (BPNN) and generalized regression (GRNN) neural networks. Hao used NOx emissions data from an actual power plant, to train and validate the SVR model. Moreover ,an ant colony optimization(ACO)based technique was proposed to select the generalization parameter C and Gaussian kernel parameter g. The focus is on the predictive accuracy and time response characteristics of the SVR model.

ii. Clustering / Regression

(Song and Kusiak, 2007) applied a data-mining approach to develop a model for optimizing the efficiency of an electric utility boiler subject to operating constraints. Selection of process variables to optimize combustion efficiency is discussed. The selection of variables of a coal fired boiler, is critical to control of combustion efficiency. Two schemes of generating control settings and updating control variables were evaluated. The first is based on both controllable and non controllable variables. While the second scheme merged response variables in clustering process. The process control scheme based on the response variables produces the smallest variance of the target variable due to reduced coupling among the process variables. 37 input variables were used out of 76 were selected to build the regression model. In their work they used Clustering to construct clusters of parameters. Decision tree to predict the boiler efficiency. Neural networks to predict the megawatt load and unit heat rate.(Chu et al., 2003) applied a neural network to predict the performance index and some non-analytical constraints, thus speeding up the trial-and-error process of finding the optimal operating points, thereby optimizing the boiler's combustion process.

Table 2.4 Summary of Data Mining used for Boilers' Efficiency in Power Plants

DM Task Category	Data mining Method	What is Predicted	Researcher
Regression	Support vector Regression (SVR), back propagation (BPNN) , generalized regression (GRNN)	Implement combustion optimization software package for low Nox combustion	Hao Zhou, 2012
Clustering/ Regression	Clustering to construct clusters of parameters. Decision tree to predict the boiler efficiency. Neural networks to predict the megawatt load and unit heat rate.	Boiler efficiency megawatt load, and unit heat rate. 37 input variables were used.	Zhe Song and Andrew Kusiak, 2007
	NN	Optimize the boiler's combustion process	J. Z. Chu et al., 2003
Association	Expert system using association-rule	Mine relationships among the parameters of a power plant.	Ogilvie et al., 1991
Genetic data mining	Evolutionary computation algorithm	Determine the optimal design of the burner reducing the emissions of NOx as well as the pressure fluctuation of the flame.	Büche et al, 2002
	NN and evolutionary computation techniques	determine the optimal fuel/air ratio.	Cass et al., 1997
Optimization-Based Approach	NN	Optimize the boiler's operations and thus reduce the emission of NOx and improve the boiler's performance.	Booth and Roland, 1998
Others	NN	Identify the dynamic process of the nitrogen oxides and carbon monoxide emissions	Chong et al., 2002
	Genetic wrapper approach	Select a subset of parameters for mining boiler data to improve combustion efficiency	Burns et al.,2004
	Expert system using fuzzy logic	Combustion control	Miyayama et al., 1991
	rough set theory	Diagnose the faults of boilers	Yang and Liu , 2004

iii. Association

(Ogilvie, Swidenbank and Hogg, 1991) applied the association-rule algorithm to mine relationships among the parameters of a power plant. The inducted rules were intended for an expert system.

iv. Genetic data mining

(Büche, Stoll, Dornberger and Koumoutsakos, 2002) applied an evolutionary computation algorithm to determine the optimal design of the burner reducing the emissions of NOx as well as the pressure fluctuation of the flame. (Cass and Radl, 1997) combined the neural network and evolutionary computation techniques to determine the optimal fuel/air ratio.

v. Optimization-Based Approach

(Booth and Roland, 1998) developed a neural network model to optimize the boiler's operations and thus reduce the emission of NOx and improve the boiler's performance.

vi. Others

(Chong, Wilcox, and Ward, 2002) applied a neural network model to identify the dynamic process of the nitrogen oxides and carbon monoxide emissions. (Burns, Kusiak and Letsche, 2004) used a genetic wrapper approach to select a subset of parameters for mining boiler data to improve combustion efficiency. (Miyayama et al., 1991) developed an expert system for combustion control using fuzzy logic and applied it to the coal-fired power plant. (Yang and Liu, 2004) presented a hybrid-intelligence data mining framework which involves an attribute reduction technique and rough set theory to diagnose the faults of boilers.

2.6. Discussion

This chapter reviewed the literature in the domain of data mining techniques which were applied to solve different problems in power plants. The review shows the variety of problems in power plants like: power plant yield, failure detection and diagnoses, emission of Nox, power curves and wind speed, and boilers' efficiency. Even some economical problems related to power plants were reviewed like: prediction of power demand, and price forecasting. Solving these problems by traditional ways require expensive, complicated, and long investigations. Therefor, the purpose of using data mining in power plants problems, is to provide reliable, accurate and effective solution for these problems

In the literature many data mining techniques were applied. However, some of them were dominant like Regression and Classification, because they are more suitable for power plants' problems. Power plants are equipped with huge number of sensors that generates thousands of features. Therefore, building prediction models is normally commenced by a feature reduction phase (either feature extraction or feature selection). Feature selection techniques is important to get rid of redundant and irrelevant features, which require high computational resources, and lead to overfitting. Many researchers only depend on domain expert to select relevant features, although feature selection techniques (Filter, Wrapper, Embedded and Hybrid) can provide not only optimum set of features, but also indications about the power plant problems. Using wrapper method in feature selection, which wraps the feature selection, and prediction in one phase provides better results, because the algorithm used for building the prediction mode, is the same that is used to select the features.

Different regression algorithms are designed to predict the amount of generated power in wind power plants, and to optimize performance of boilers in thermal power plants. However, the amount of power generated by thermal power plants, and the efficiency of power plant still need a lot of work, and can get a lot of benefits from data mining and prediction techniques.

2.7. Summary

This chapter presented a review of the available literature related to the application of data mining techniques to solve power plant problems, specially the prediction of amount of generated power. A high emphasis has been paid to the methods used for feature selection, and the accuracy of the applied prediction algorithms. The chapter begins with an overview of thermal power plants, followed by data mining taxonomy and application of data mining to solve power plant problems. The review showed the benefits of data mining techniques, especially feature selection and regression, in predicting the behavior of power plants. The next chapter presents the research methodology followed to achieve the desired goals.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1. Introduction

This chapter presents the methodology followed in this research. It shows the steps followed to achieve the research objectives. One of the research objectives is to design a feature selection technique, that can determine the best set of features to predict the amount of generated power from a thermal power plant. Another objective is to Design a regression technique that can accurately predicts the amount of generated power, using only the controllable parameters.

CRISP-DM model (Shearer, 2000) is followed in this research. Therefore, this chapter is organized in six sections, the first one is this introduction. The second section shows the research operational framework. This framework starts with understanding the domain, then each one of the two sub problems uses its own datasets, but both of them are following identical steps. The third section is about algorithms selection and initial comparison. The forth section is about discussion and results evaluation. The fifth section is about papers to be submitted and thesis writing. The sixth and last section gives a summary about this chapter.

3.2. Operational Framework

Operational framework is a structured guide that helps the researcher to achieve the research objectives. The framework should be well prepared and organized to describe the exact steps that followed, research phases, experiments, and results evaluation methods. Figure 3.1 illustrate the operational framework for this research. The operation framework is composed of seven phases:

1. Understanding the Domain
2. Literature Review
3. Data Preparation
4. Feature Reduction & Selection
5. Prediction Model Development (for the two objectives)
6. Writing the Thesis.

Next sections provides more details about each phase.

3.2.1 Understanding the Domain

This phase focuses on understanding the theory of power generation, specially thermal power plants. This phase is composed of three main parts. The first part is to understand the theory of power generation, it started by studying power systems in general, then two important topics of thermal power plants were studied:

1. Carnot Cycle: this cycle is a pure theoretical steam power generation cycle (Martínez et al, 2016).
2. Rankine Cycle: is the application of Carnot cycle, that is implemented by all thermal power plants (Kapooria et al, 2008).

The second part of this section is more about practice. It is about a real world case study, which is KNPP (**K**hartoum **N**orth **T**hermal **P**ower **P**lant). This power plant was studied by :

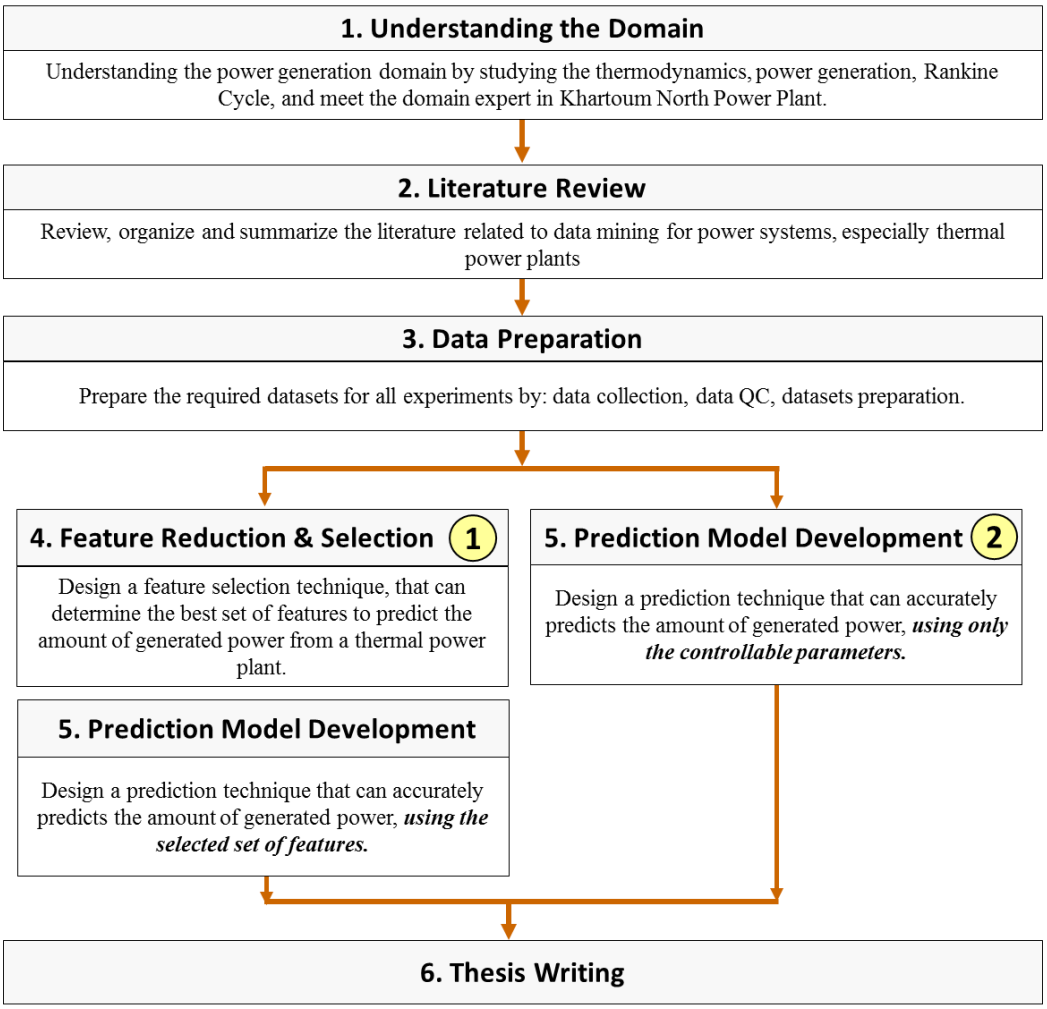


Figure 3.1 Operational Framework

1. Meeting the domain experts (Power plant operation and efficiency engineers) to understand the **design** of KNPP, and defining the most important factors that **influence** power generation.
2. Collecting **Manufacturers** documentation. Many documents are collected like: Steam Consumption Graph in figure 2.4, Unit 3 and Unit 4 components diagrams (Appendix C, D).

This phase is the base of the research. By understanding the domain, the road map of the research becomes very clear.

3.2.2 Literature Review

This Phase presents and summarizes many articles and researches related to application of data mining in power plants. The chapter starts by providing an introduction about thermal power plants and power generation, then an introduction about data mining methods and algorithm by presenting a taxonomy of data mining methods. After that the CRISP-DM model is presented in more details. Feature Reduction and Selection techniques are then presented, which are followed by developing and evaluating Prediction Models. Finally an intensive review about application of data mining techniques in power systems, and specially in thermal power plants was shown. The literature review is a continuous phase, the research starts with literature review, then during all research phases any interesting information that is related to the research is summarized and added to the references list. The literature review presents the concepts and the related works in the last 15 years.

3.2.3 Data Preparation

Preparing input for a data mining usually consumes the bulk of the effort invested in the entire data mining process. There are some simple practical points to be aware of when preparing the data. Best practice shows that real data, like the one that is captured instantly by SCADA, is always low in quality, so, to increase this data quality; careful checking - or *data cleaning* - is needed. Also data needs to be prepared in a relational format to be suitable for machine learning tools. So the goal of this phase is to prepare the required data in a usable format, that could be loaded directly to data mining tools. This part will show how data is collected, and prepared for machine learning tools.

3.2.3.1 Data Collection

As shown in chapter one, the case study for this research is unit 3 and 4 from Khartoum North Power Plant. The data is collected instantly via SCADA system and saved in a central database. Figure 3.2 shows an illustrative diagram of KNPP that

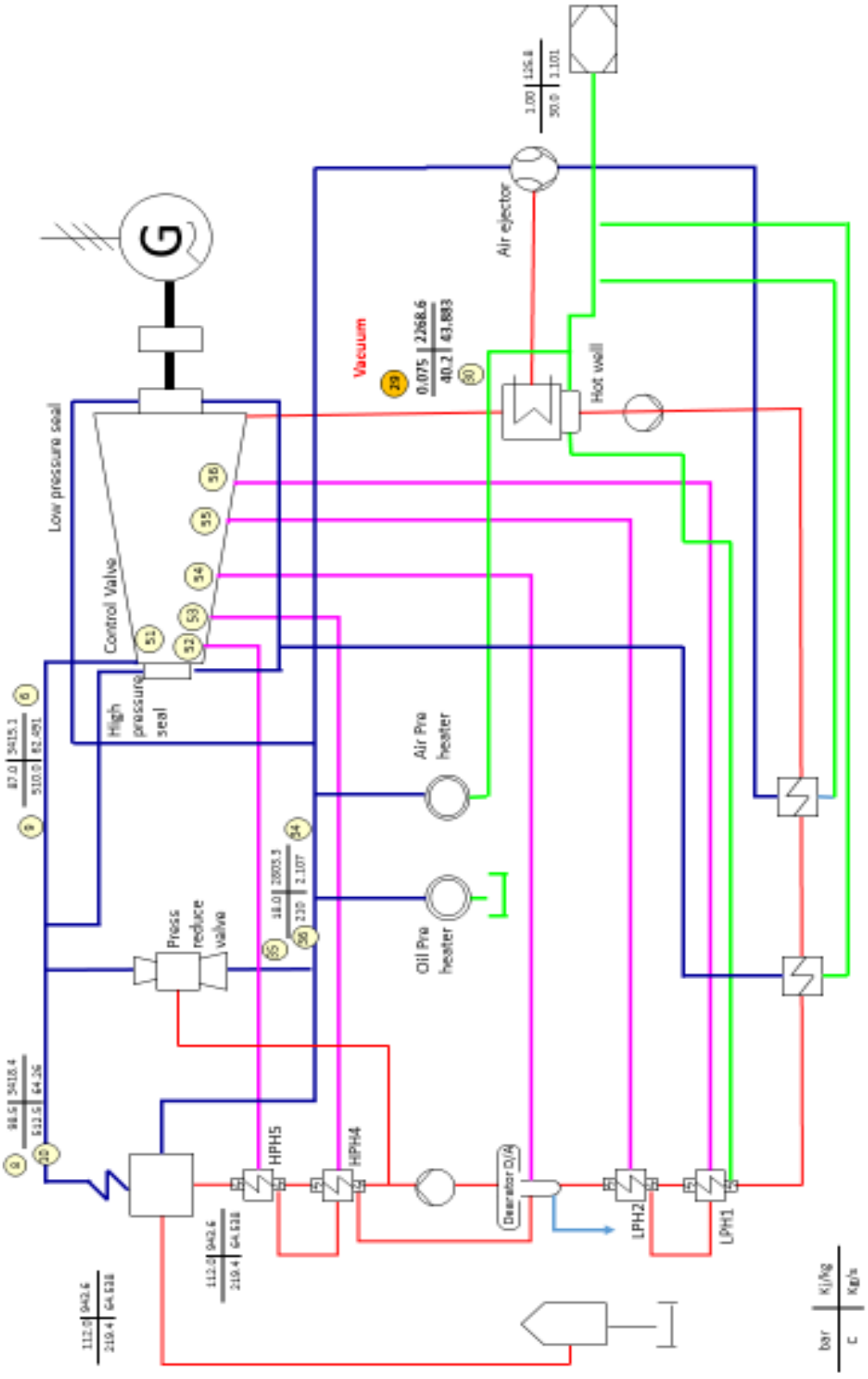


Figure (3.2) Illustrative diagram of KNPP

Table (3.1) All parameters collected from KNPP

Serial	Code In Diagram	Circuit	Name	Tag Number	Engineer Classification
1	***5***	***T/A***	***Generated power (MW)***	***24CFA01CE033XQ01:av***	***Target***
2	6	T/A	Main steam flow (kg/s)	24LBA10CF903:av	Direct
3	7	Boiler	Total steam flow (kg/s)	24LBA10CF902:av	Direct
4	8	Boiler	Main steam header pressure (bar)	24LBA10CP003XQ01:av	Direct
5	9	T/A	T/A inlet steam temperature (deg C)	24LBA20CT002XQ01:av	Direct
6	10	Boiler	Main steam header steam temperature (deg C)	24LBA10CT003XQ01:av	Direct
7	11	Gas	Flue gas temperature after RAH (deg C)	24HNA40CT003XQ01:av	Irrelevant
8	12	Gas	Flue gas temperature after RAH (deg C)	24HNA40CT004XQ01:av	Irrelevant
9	13	Gas	Flue gas O2% at RAH in (%)	24HNA40CQ001XQ01:av	Irrelevant
10	14	Gas	Flue gas temperature at economiser inlet (deg C)	24HNA40CT001XQ01:av	Irrelevant
11	15	Gas	Flue gas temperature at RAH inlet(deg C)	24HNA40CT002XQ01:av	Irrelevant
12	16	Gas	Flue gas pressure at RAH in (mbar)	24HNA40CP001XQ01:av	Irrelevant
13	17	Gas	Flue gas pressure at stack in (mbar)	24HNA40CP002XQ01:av	Irrelevant
14	18	Gas	Combustion champer pressure	24HLA42CP001XQ01:av	Irrelevant
15	19	Fuel Oil	Fuel oil flow (kg/s)	24END40CF901:av	Irrelevant
16	20	Fuel Oil	Fuel oil temp. supplied to burners (deg C)	24END40CT001XQ01:av	Irrelevant
17	21	Fuel Oil	Fuel oil pressure befor burners	24END40CP005XQ01:av	Irrelevant
18	22	Feedwater	HPH5 discharge feedwater flow (kg/s)	24LAB50CF902:av	?
19	23	Feedwater	Feedwater temperature at economiser inlet (deg C)	24LAB50CT001XQ01:av	?
20	24	Feedwater	Feedwater pressure at economiser inlet (bar)	24LAB50CP002XQ01:av	?
21	25	Coolingwater	Condenser right inlet temperarure (deg C)	24PAB31CT001XQ01:av	Important
22	26	Coolingwater	Condenser left inlet temperarure (deg C)	24PAB32CT001XQ01:av	Important
23	27	Coolingwater	Condenser right outlet temperarure (deg C)	24PAB31CT003XQ01:av	Important
24	28	Coolingwater	Condenser left outlet temperarure (deg C)	24PAB32CT003XQ01:av	Important
25	30	Condensate	Condenser inlet exhaust steam temperature (deg C)	24MAG10CT001XQ01:av	Target
26	31	Condensate	Condensate water flow (kg/s)	24LCA03CF001XQ01:av	?
27	32	Condensate	Condenser hot well temperature (deg C) a	24MAG10CT003:av	?
28	33	Condensate	Condenser hot well temperature (deg C) b	24MAG10CT004:av	?
29	34	Aux. steam	Auxiliary steam flow (kg/s)	24LBG10CF901:av	Important
30	35	Aux. steam	Auxiliary steam pressure (bar)	24LBG10CP002XQ01:av	Important
31	36	Aux. steam	Auxiliary steam temperature (deg C)	24LBG10CT001XQ01:av	Important
32	37	Air	Combustion air flow (Nm3/s)	24HLA30CF905:av	?
33	38	Air	Air temperature at FDF inlet (deg C)	24HLA10CT001XQ01:av	Indirect
34	39	Air	Air temperature at FDF inlet (deg C)	24HLA10CT002XQ01:av	Indirect
35	40	Air	Air temperature after RAH (deg C) (1)	24HLA30CT003XQ01:av	?
36	41	Air	Air temperature after RAH (deg C) (2)	24HLA30CT004XQ01:av	?
37	42	Air	FDF discharge air pressure (mbar)	24HLA13CP001XQ01:av	?
38	43	Air	FDF A speed (rpm)	24HLB11CS001XQ01:av	?
39	44	Air	FDF B speed (rpm)	24HLB12CS001XQ01:av	?
40	45	Air	Air temperature after SAH (deg C) (1)	24HLA30CT001XQ01:av	?

Table (3.1) All parameters collected from KNPP - Continued

Serial	Code In Diagram	Circuit	Name	Tag Number	Engineer Classification
41	46	Air	Air temperature after SAH (deg C) (2)	24HLA30CT002XQ01:av	?
42	47	Feedwater	HPH4 inlet feedwater temperature (deg C)	24LAB30CT001XQ01:av	?
43	48	Feedwater	HPH4 outlet feedwater temperature (deg C)	24LAB40CT001XQ01:av	?
44	49	Feedwater	HPH5 outlet feedwater temperature (deg C)	24LAB50CT002XQ01:av	?
45	50	Feedwater	HPH5 inlet feedwater temperature (deg C)	24LAB40CT003XQ01:av	?
46	51	T/A	T/A wheel champer steam pressure (bar)	24MAL05CP001XQ01:av	?
47	52	T/A	T/A bleeder (1) pressure (bar)	24LBQ40CP001XQ01:av	?
48	53	T/A	T/A bleeder (2) pressure (bar)	24LBQ30CP001XQ01:av	?
49	54	T/A	T/A bleeder (3) pressure (bar)	24LBS30CP001XQ01:av	?
50	55	T/A	T/A bleeder (4) pressure (bar)	24LBS20CP001XQ01:av	?
51	56	T/A	T/A bleeder (5) pressure (bar)	24LBS10CP001XQ01:av	?
52	57	T/A	T/A Lub oil pressure (bar)	24MAV30CP001XQ01:av	Irrelevant
53	58	T/A	T/A lub oil temperature after strainer (deg C)	24MAV30CT001XQ01:av	Irrelevant
54	59	T/A	Control oil pressure (bar)	24MAV23CP001XQ01:av	Irrelevant
55	60	T/A	T/A differential expansion (mm)	24MAA10CY041XQ01:av	?
56	61	T/A	T/A axial displacement A (mm)	24MAA10CG001XQ01:av	Important
57	62	T/A	T/A axial displacement B (mm)	24MAA10CG002XQ01:av	Important
58	63	T/A	T/A axial displacement C (mm)	24MAA10CG003XQ01:av	Important
59	64	T/A	T/A bearing 3 vibration (mm/s)	24MKD10CY051XQ01:av	?
60	65	T/A	T/A bearing 4 vibration (mm/s)	24MKD10CY061XQ01:av	?
61	66	T/A	T/A bearing 1 vibration (1) (mic)	24MAD10CY011XQ01:av	?
62	67	T/A	T/A bearing 1 vibration (2) (mic)	24MAD10CY012XQ01:av	?
63	68	T/A	T/A bearing 2 vibration (1) (mic)	24MAD10CY031XQ01:av	?
64	69	T/A	T/A bearing 2 vibration (2) (mic)	24MAD10CY032XQ01:av	?
65	70	Generator	TBN side cold air (deg C)	24MKA10CT001:av	?
66	71	Generator	TBN side warm air (deg C)	24MKA10CT002:av	?
67	72	Generator	Exciter side cold air (deg C)	24MKA10CT003:av	?
68	73	Generator	Exciter side warm air (deg C)	24MKA10CT004:av	?
69	74	Generator	PMG side cold air (deg C)	24MKA10CT005:av	?
70	75	Generator	PMG side warm air (deg C)	24MKA10CT006:av	?
71	76	Generator	Generator winding temperature (deg C) 1	24MKA10CT007:av	?
72	77	Generator	generator winding temperature (deg C) 2	24MKA10CT008:av	?
73	78	Generator	Generator winding temperature (deg C) 3	24MKA10CT009:av	?
74	79	Generator	Generator winding temperature (deg C) 4	24MKA10CT010:av	?
75	80	Generator	Generator winding temperature (deg C) 5	24MKA10CT011:av	?
76	81	Generator	Generator winding temperature (deg C) 6	24MKA10CT012:av	?
77	82	Boiler	Economizer inlet FW pressure (bar)	24LAB50CP002XQ01:av	Dublicated
78	83	Boiler	economizer inlet FW temperature (deg C)	24LAB50CT001XQ01:av	Dublicated
79	29	Condensate	Condenser inlet exhaust steam pressure (bar a)	24MAG10CP001XQ01:av	?

shows location of some sensors in the power plant, the numbers in circles are parameters numbers. Table 3.1 shows a list of all parameters collected from KNPP, the table contains the following 6 attributes:

1. **Serial**: just a serial for the parameters.
2. **Code In Diagram**: Each parameter has a unique number, this number is used to reference the parameter location in figure 3.1 the illustrative diagram.
3. **Part in Power Plant**: This is the part of power plant of this parameter.
4. **Name**: the name of the parameter is full name that gives a clear description of the parameter.
5. **Tag Number**: it is a unique code of the parameter across the whole power plant Unit.
6. **Engineer Classification**: this is the domain expert comments about the parameter. The values in this field are as follows:
    i. **Target**: This is the amount of generated power which will be used as the Target of prediction algorithms.
    ii. **Direct**: means this parameter has a direct effect in the amount of the generated power.
    iii. **Irrelevant**: means this parameter has no effect in the amount of the generated power.
    iv. **?**: the domain expert are not sure about the parameter relevancy.
    v. **Important**: this parameter is important in the amount of the generated power

To study the status of the power plant, and to be able to calculate the generated power, we need pressure, temperature and steam flow at turbine inlet and outlet, hoever in this research all parameters were studied. We have to measure these three variables at specific locations in the power plant like turbine inlet, and turbine outlet. So, the required datasets for this research are numeric, specifically pressure, temperature and steam flow captured from different locations.

i. Existing System in Power Plant

As mentioned in chapter (1) KNPP was constructed in three phases, each phase consists of two identical units. Each phase has a different monitoring and control system. Phase 1 uses a manual traditional control system, while phase 2 uses Metso monitoring and control system. In this research we focus only on phase 2 which consists of unit 3 and 4. The existing monitoring and control system for unit 3 and 4 uses advanced sensors to automatically collect data that shows the current status of all components of power plant. Appendix C and D show a snapshot of the monitoring screens of Metso system.

ii. Sensors

Many types of sensors are used to collect real time data from the power plant. For the purpose of this research we focus on pressure, temperature and mass follow sensors.

Pressure sensors: are used in controlling and monitoring in thousands of applications. A pressure sensor measures pressure of gases or liquids. Pressure is an expression of the force required to stop a fluid from expanding, and is usually stated in terms of force per unit area. A pressure sensor generates a signal as a function of the pressure imposed. Pressure sensors can vary in technology, design, performance, application and cost (Sensorland, 2017).

Temperature Sensors: One of the most critical factors in power plants is the temperature, so accurate thermometers are connected to all parts of Rankine Cycle. According to the expected temperature a suitable thermometer is used. These thermometer are connected to the required component to instantly measure the temperature and send theses reading to the database. Figure 3.3 shows a sample of a thermometer that is used in steam turbine and can measure up to 5300 C.(Series 176 - On-Turbine Instability Sensor (OTIS))

Figure 3.3 Series 176 - On-Turbine Instability Sensor (OTIS) (Flokal, 2017)

Flow meter: Turbine flow meters have a widespread use for accurate liquid and gas measurement applications. The unit consists of a multiple-bladed rotor mounted with a pipe, perpendicular to the liquid flow. The rotor spins as the liquid passes through the blades. The rotational speed is a direct function of flow rate, finally electrical pulses counted and totalized. To ensure the accuracy of the meter, manufacturer calibrates and tests meters before shipping to customers. Figure 3.4 shows a turbine mass flow meter [50].

Figure 3.4 Turbine mass flow meter (Flokal, 2017)

iii. Database

Real time data that is collected from all sensors in the power plant is sent directly to a central database. In KNPP the central database is a single SQL server database, but because of huge amount of real time data collected, the system in KNPP is configured with auto purge to retain data for two months only. So, at any time two months data is available for analysis. From SQL server you can directly export data to any text or spread sheet format.

3.2.3.2 Data preprocessing

Real time data as collected from the monitoring and control system can not directly be loaded to the data mining tool. The attributes in the file are represented as rows rather than columns, while the columns represents the intervals for each reading. To prepare the dataset in a usable format to the machine learning tools, the following steps were done:

i. Collect the raw data.
ii. Remove and add some rows and columns from raw excel file.
iii. Transpose the cells.
iv. Remove redundant fields.
v. Create separate file for each unit (Unit 3 and Unit 4)
vi. Create one file that contains all data.

Below is some details about what have been done and the final layout for the data set. The output from each of the following steps, becomes and input to the next step.

i. Collect the Raw Data

The input of this step is the central database, and the out put is an excel file that contains the raw data as-is from the database. MS excel is used to pull required data from the central database. Table 3.2 shows a sample of a raw data as taken from the system. First two rows shows the period for data collected in this file is between 2 to 4 PM on the 1st of May 2014.

1. The first column is **Circuit:** contains the component of the power plant, this could be one of ten options which are: Air, Aux. steam, Boiler, Cooling water, Feedwater, Fuel Oil, Gas, Generator, Generator or T/A (turbine). This is same as Circuit in table 3.1.
2. The second is **Description**: this contains a description of this parameter, this column actually contains the names of attributes of the data set, it is same as Name field in table 3.1.
3. The third is **Tag Number**: this is a unique identifier for this parameter, it uniquely identifies the parameter among all parameters in the power plant. Same field exist in table 3.1.

4. The rest of columns contain the actual data collected. The headers of these columns represents the time.

ii. Remove and add some rows and columns from raw excel file.

This steps takes as input the raw file (which is output of the previous step), its output is an updated raw file. The first two rows of excel raw data file contains start and end date of this data file. These rows where removed and new columns were added to represent date (Year, Month and Day). New column is added to represent the Unit, as shown in table (3.3).

iii. Transpose the cells.

The rows in the raw file represents attributes, while columns represents instances. So the table was transposed to swap rows and columns. There are more than eighty attributes in the data set, all of them shown in table 3.1. The output of this step is a transposed file which is ready to be used by the machine learning tool. The new file still contains the eighty attributes.

iv. Remove redundant fields.

The output from the previous step contains redundant attributes like (*Flue gas temperature after RAH (deg C)*). Those attributes were collected using different sensors, in most of them the data is almost typical. So the redundant attributes were removed. The output of this step is a clean dataset that contains only 63 attributes out of the 83 attributes of table 5.1. To refer to each attribute from this step we need a unique identifier, so a new serial from 1 to 63 is assigned for each attribute. Table 3.4 shows the final attribute set which will be used for feature selection and prediction models.

v. Create separate file for each unit.

Because each unit in the power plant is an independent unit, separate file is prepared for each unit. This is done because we are going to study each unit separately.

vi. Create one file that contains all data.

A combined file that contains both unit 3 and 4 is prepared, to check whether a generic model for a thermal power plant could be obtained.

Table 3.2 Structure of Raw data in excel as taken from the Central Database

Start Date	5/1/2014 14:00			
End Date	5/1/2014 16:00			
Circuit	**Description**	**Tag Name**	**14:00:00**	**14:02:00**
T/A	Generated power (MW)	24CFA01CE033XQ01:av	49.973	50.286
T/A	Main steam flow (kg/s)	24LBA10CF903:av	51.354	51.553
Boiler	Total steam flow (kg/s)	24LBA10CF902:av	52.452	52.68
Boiler	Main steam header pressure (bar)	24LBA10CP003XQ01:av	87.223	87.956
T/A	T/A inlet steam temperature (deg C)	24LBA20CT002XQ01:av	516.372	517.59
Boiler	Main steam header steam temperature (deg C)	24LBA10CT003XQ01:av	516.418	515.594
Gas	Flue gas temperature after RAH (deg C)	24HNA40CT003XQ01:av	148.249	148.391
Gas	Flue gas temperature after RAH (deg C)	24HNA40CT004XQ01:av	150.801	150.935

Table 3.3 Part of Unit 3 data set

Unit	**Year**	**Month**	**Day**	**Time**	**Generator power MW**	**Main Steam Flow kg/s**	**Total Steam Flow kg/ s**	**Main Steam Header Pressure bar**	**T_A Inlet Steam Temperature deg_C**	**Main Steam Header Steam Temperature deg_C**
3	2012	8	22	10:24:00	40.42	47.971	49.456	88.914	510.847	513.34
3	2012	8	22	10:26:00	40.166	48.446	49.944	89.911	503.362	502.873
3	2012	8	22	10:28:00	39.717	48.462	49.939	88.421	498.064	500.57
3	2012	8	22	10:30:00	39.56	47.768	49.226	86.162	503.087	509.349
3	2012	8	22	10:32:00	39.834	47.581	49.071	85.782	511.273	517.941
3	2012	8	22	10:34:00	40.166	47.574	49.096	86.669	514.592	519.463

3.2.3.3 The Datasets

Each problem has its own dataset. Below is the description of these datasets:

i. Problem (1) Dataset

Three datasets were prepared for this problem (Unit 3, Unit 4, and Unit 3 & 4). All these datasets have the same structure, Table (5.4) shows the names of all 63 attributes, the last one (attribute No. 63) is the target. Number of instances for each dataset as follows:

1. Unit 3 dataset: 300 clean instances, the total set size was 1083.
2. Unit 4 dataset: 720 clean instances, the total set size was 1083.
3. Unit 3 & 4 dataset: 300+720 =1020 clean instances.

ii. Problem (2) Dataset

The second problem focuses on controllable parameters, (*pressure, temperature and steam flow*) at turbine inlet. The target field for these datasets is also the amount of generated power in mega watts. Tables (5.5) and (5.6) shows samples of unit 3 and unit 4 datasets respectively.

Table 3.4 Attributes of problem (1) dataset

#	Feature Name
1	***Main steam flow (kg/s)***
2	Total steam flow (kg/s)
3	***Main steam header pressure (bar)***
4	***T/A inlet steam temperature (deg C)***
5	Main steam header steam temperature (deg C)
6	HPH5 discharge feedwater flow (kg/s)
7	Feedwater temperature at economiser inlet (deg C)
8	Feedwater pressure at economiser inlet (bar)
9	Condenser right inlet temperature (deg C)
10	Condenser left inlet temperature (deg C)
11	Condenser right outlet temperature (deg C)
12	Condenser left outlet temperature (deg C)
13	Condensate water flow (kg/s)
14	Condenser hot well temperature (deg C) a
15	Condenser hot well temperature (deg C) b
16	Auxiliary steam flow (kg/s)
17	Auxiliary steam pressure (bar)
18	Auxiliary steam temperature (deg C)
19	Combustion air flow (Nm3/s)
20	Air temperature at FDF inlet (deg C)
21	Air temperature at FDF inlet (deg C)
22	Air temperature after RAH (deg C) (1)
23	Air temperature after RAH (deg C) (2)
24	FDF discharge air pressure (mbar)
25	FDF A speed (rpm)
26	FDF B speed (rpm)
27	Air temperature after SAH (deg C) (1)
28	Air temperature after SAH (deg C) (2)
29	HPH4 inlet feedwater temperature (deg C)
30	HPH4 outlet feedwater temperature (deg C)
31	HPH5 outlet feedwater temperature (deg C)
32	HPH5 inlet feedwater temperature (deg C)
33	T/A wheel champer steam pressure (bar)
34	T/A bleeder (1) pressure (bar)
35	T/A bleeder (2) pressure (bar)
36	T/A bleeder (3) pressure (bar)
37	T/A bleeder (4) pressure (bar)
38	T/A bleeder (5) pressure (bar)
39	T/A differential expansion (mm)
40	T/A axial displacement A (mm)
41	T/A axial displacement B (mm)
42	T/A axial displacement C (mm)
43	T/A bearing 3 vibration (mm/s)
44	T/A bearing 4 vibration (mm/s)
45	T/A bearing 1 vibration (1) (mic)
46	T/A bearing 1 vibration (2) (mic)
47	T/A bearing 2 vibration (1) (mic)
48	T/A bearing 2 vibration (2) (mic)
49	TBN side cold air (deg C)
50	TBN side warm air (deg C)
51	Exciter side cold air (deg C)
52	Exciter side warm air (deg C)
53	PMG side cold air (deg C)
54	PMG side warm air (deg C)
55	Generator winding temperature (deg C) 1
56	generator winding temperature (deg C) 2
57	Generator winding temperature (deg C) 3
58	Generator winding temperature (deg C) 4
59	Generator winding temperature (deg C) 5
60	Generator winding temperature (deg C) 6
61	***Condenser inlet exhaust steam pressure (bar a)***
62	***Condenser inlet exhaust steam temperature (deg C)***
63	**Generated power (MW)**

Table 3.5 Sample dataset of Unit 3

Steam Flow	Pressure	Temperature	Amount of Power in MW
36.213	85.932	508.91	30.125
36.277	87.256	504.597	29.89
36.17	86.822	507.631	30.144
47.706	86.907	511.416	40.147
47.971	88.914	510.847	40.42

Table 3.6 Sample dataset of Unit 4

Steam Flow	Pressure	Temperature	Amount of Power in MW
30.994	87.291	512.103	29.91
30.701	86.939	509.371	29.988
30.888	87	508.326	30.066
44.886	87.496	508.675	45.011
44.458	86.864	518.263	44.835

3.2.4 Feature Reduction and Selection

As shown in figure Figure 4.1 Operational Framework, after data preparation there are two branches, each for one research objective. The first branch to achieve the first research objective (To design a feature selection technique, that can determine the best set of features to predict the amount of generated power from a thermal power plant) and it is consist of two parts:

- **Feature Reduction and Selection:** The goal of this part is to design a feature selection technique, that can determine the best set of features to predict the amount of generated power from a thermal power plant. Hence, sets will be selected and evaluated to select the best set, which will be used to develop the prediction model using the selected set of features.
- **Prediction Model Development:** The goal of this part is design a prediction technique that can accurately predicts the amount of generated power, *using the selected set of features* which were selected in the previous step. Many regression algorithms were used to develop the models. The algorithms which will show better performance will be selected to build the prediction mode.

  To get better performance wrapper method will be used to achieve this objective.

Wrapper method uses the same algorithm to select the best set of features, and to build the prediction model. Chapter 4 is dedicated for this branch, more details will be provided there. To design a prediction technique that can accurately predicts the amount of generated power, using only the controllable parameters.

3.2.5 Prediction Model Development

After data preparation, the second branch is to achieve the second research objective (*To design a prediction technique that can accurately predicts the amount of generated power, using only the controllable parameters*). Unlike the first branch, the dataset is ready, it consists of three predictors (*Steam flow, Steam Pressure, Steam Temperature*) which were collected from the turbine inlet, and the target (*Amount of Generated Power*). The goal here is to develop a prediction model using only the

controllable parameters. Also many regression algorithms were tested, the one that shows the minimum errors will be selected to develop the model. Chapter 5 of this thesis is dedicated for this branch, more details will be provided there.

3.2.6 Discussion and Results Evaluation:

All dataset in this research are numerical, so all models will be regression models. According to the dataset size test method will be used. For small datasets Cross validation will be used, while Separate test set, (normally 1/3 of the data set) will be used with bigger datasets. To evaluate each model two methods will be followed:

(1) The accuracy of prediction model is measured by :

1. Correlation coefficient,
2. Mean absolute error,
3. Root mean squared error,
4. Relative absolute error,
5. Root relative squared error

Equations of these factors are shown in Figure 3.5 (Witten, Frank and Hall, 2011).

Mean-squared error	$\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}$
Root mean-squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$
Mean-absolute error	$\frac{\lvert p_1 - a_1\rvert + \ldots + \lvert p_n - a_n\rvert}{n}$
Relative-squared error*	$\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \ldots + (a_n - \bar{a})^2}$
Root relative-squared error*	$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \ldots + (a_n - \bar{a})^2}}$
Relative-absolute error*	$\frac{\lvert p_1 - a_1\rvert + \ldots + \lvert p_n - a_n\rvert}{\lvert a_1 - \bar{a}\rvert + \ldots + \lvert a_n - \bar{a}\rvert}$
Correlation coefficient**	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1}$, $S_P = \frac{\sum_i (p_i - \bar{p})^2}{n-1}$, $S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$

*Here, $\bar{a}$ is the mean value over the training data.
**Here, $\bar{a}$ is the mean value over the test data.

[Data mining – Ian Wittan]

Figure 3.5 Equations of the Evaluation methods

(2) Graphical comparison between actual vs predicted values for each model will be presented, to give visual representation of results. This makes the comparison between actual and predicted values much easier.

Each model has its own **target** and **predictors**, therefore, each model is discussed separately to interpret the results according to power plant status and thermodynamics laws. Model evaluation will be used to select the best performance model, but the model **will not be accepted unless** it matches the thermodynamics laws.

3.2.7 Writing the Thesis.

After building the prediction models and evaluating results as described above, three papers were published. The first one is a review paper about the using data mining in power plants, the other two papers are research papers, one for each problem. Then this thesis was reviewed and submitted.

3.3. Summary

This chapter presents the methodology which is followed by the researcher to achieve the research goals. CRISP-DM model was used to prepare the operational framework for this research, this framework is a comprehensive one that describes clearly each phase. Sixteen algorithms will be used for each dataset, the algorithm that shows the minimum errors will be selected to build the prediction model for that dataset. After building models testing will be done using cross validation or separate test set.

The expected outcomes of this research is a list of parameters that influence power prediction, and a prediction model to predict the amount of generated power from a thermal power plant using the controllable parameter.

CHAPTER FOUR

Feature Selection and Prediction Models using the Full Features

4.1.Introduction

The target of this chapter is " *To design a feature selection technique, that can determine the best set of features to predict the amount of generated power from a thermal power plant*." So, the research target could be divided into two parts: the first is to filter dataset parameters to select the most influencing parameters, and the second is to build a prediction model that uses the selected parameters to predict the amount of generated power from a thermal power plant. To achieve these goals, three different datasets were used: one for unit 3 , the other for unit 4 and the last one is a combined dataset that contains both unit 3 and 4. The reason behind that is to study each unit separately, and to check whether a generic model could be used.

This chapter starts by describing the datasets, then basic statistical analysis about these datasets is shown. After that an initial comparison between the prediction algorithms is done for each dataset, to select the most appropriate algorithm for each dataset to build the prediction model. Then for each dataset three main tasks are done: the first is the attributes' selection, the second is the power prediction model using the selected features, the third is the model evaluation. After that a summary and discussion about feature selection is done. Finally results discussion and models Comparison is done to depict the knowledge behind these numbers. The methodology followed to achieve the goal is shown in figure (4.1)

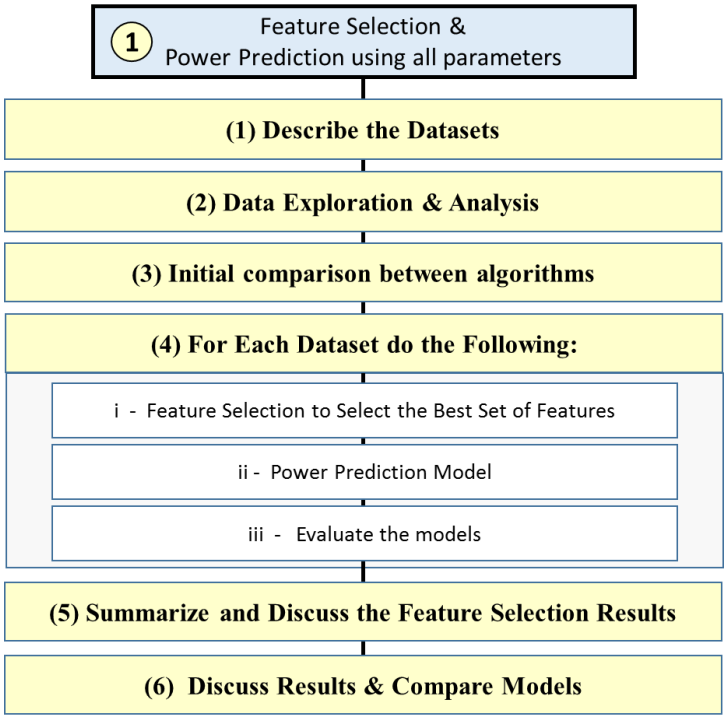


Figure 4.1 Methodology for Feature Selection and Power Prediction Models (1)

4.2.Datasets Description

As described in Chapter 1, Khartoum North Thermal Power Plant KNTPP was commissioned in three phases. Each phase is composed of two identical units, each unit is a separate power generation unit that follows Rankine Cycle . In this research we are focusing on Phase 2 which is composed of Unit 3 and Unit 4. As shown in chapter 3 Research Methodology (3.2.3 Data Preparation), after excluding useless and duplicated features we have only 63 features to be used for datasets preparation. Table 4.1 shows the 63 features of the Power Prediction Datasets. A unique identifier (from 1 to 63) has been assigned for each one of these 63 features. As per the thermodynamic laws and Rankine Cycle (Learn Engineering, 2013); the most important features to calculate the amount of generated power from a thermal power plant are: *Steam Flow, pressure* and *temperature* at turbine inlet, and *steam pressure* and *temperature* at turbine outlet. These are features number 1,3,4, 61 and 62 respectively in Table 4.1 (Attributes of problem (1) dataset). Feature number 63 is the class (*Generated_Power_MW*), it is the amount of the generated power in Mega Watts.

Table (4.1) All Features of Power Prediction Datasets

#	Feature Name
1	***Main steam flow (kg/s)***
2	Total steam flow (kg/s)
3	***Main steam header pressure (bar)***
4	***T/A inlet steam temperature (deg C)***
5	Main steam header steam temperature (deg C)
6	HPH5 discharge feedwater flow (kg/s)
7	Feedwater temperature at economiser inlet (deg C)
8	Feedwater pressure at economiser inlet (bar)
9	Condenser right inlet temperature (deg C)
10	Condenser left inlet temperature (deg C)
11	Condenser right outlet temperature (deg C)
12	Condenser left outlet temperature (deg C)
13	Condensate water flow (kg/s)
14	Condenser hot well temperature (deg C) a
15	Condenser hot well temperature (deg C) b
16	Auxiliary steam flow (kg/s)
17	Auxiliary steam pressure (bar)
18	Auxiliary steam temperature (deg C)
19	Combustion air flow (Nm3/s)
20	Air temperature at FDF inlet (deg C)
21	Air temperature at FDF inlet (deg C)
22	Air temperature after RAH (deg C) (1)
23	Air temperature after RAH (deg C) (2)
24	FDF discharge air pressure (mbar)
25	FDF A speed (rpm)
26	FDF B speed (rpm)
27	Air temperature after SAH (deg C) (1)
28	Air temperature after SAH (deg C) (2)
29	HPH4 inlet feedwater temperature (deg C)
30	HPH4 outlet feedwater temperature (deg C)
31	HPH5 outlet feedwater temperature (deg C)
32	HPH5 inlet feedwater temperature (deg C)
33	T/A wheel champer steam pressure (bar)
34	T/A bleeder (1) pressure (bar)
35	T/A bleeder (2) pressure (bar)
36	T/A bleeder (3) pressure (bar)
37	T/A bleeder (4) pressure (bar)
38	T/A bleeder (5) pressure (bar)
39	T/A differential expansion (mm)
40	T/A axial displacement A (mm)
41	T/A axial displacement B (mm)
42	T/A axial displacement C (mm)
43	T/A bearing 3 vibration (mm/s)
44	T/A bearing 4 vibration (mm/s)
45	T/A bearing 1 vibration (1) (mic)
46	T/A bearing 1 vibration (2) (mic)
47	T/A bearing 2 vibration (1) (mic)
48	T/A bearing 2 vibration (2) (mic)
49	TBN side cold air (deg C)
50	TBN side warm air (deg C)
51	Exciter side cold air (deg C)
52	Exciter side warm air (deg C)
53	PMG side cold air (deg C)
54	PMG side warm air (deg C)
55	Generator winding temperature (deg C) 1
56	generator winding temperature (deg C) 2
57	Generator winding temperature (deg C) 3
58	Generator winding temperature (deg C) 4
59	Generator winding temperature (deg C) 5
60	Generator winding temperature (deg C) 6
61	***Condenser inlet exhaust steam pressure (bar a)***
62	***Condenser inlet exhaust steam temperature (deg C)***
63	**Generated power (MW)**

Three datasets were prepared to build the Power Prediction Model, one for each unit and the third one is a generic dataset that combines unit 3 and unit 4 data in one set. Below is some description about these three datasets:

1. **Unit 3 dataset**: The original dataset which was initially prepared is composed of 1083 instances. After the data quality check, more the 700 instance were found with null values in different attributes, this is the first indication that unit 3 has some problem in many sensors. So, the final dataset which is used in this model contains only 300 instances. All features including the class are numeric. As shows in Table 4.1 the class is attribute number 63 (GeneratedPowerMW).

2. **Unit 4 dataset**: The original dataset which was initially prepared is composed of 1083 instances. After the data quality check, the instances with null values in different attributes is less than their counterparts in unit 3, so unit 4 dataset is bigger and much better

than unit 3 dataset. The final dataset which I used in this model contains 720 instances. All features including the class are numeric. As shows in Table 4.1 the class is attribute number 63 (Generated Power MW).

3. **Unit 3&4 dataset**: This dataset is just a new file which combines both unit 3 and unit 4 datasets. So, the total number of incenses of this new dataset is 1020. The class is also attribute number 63 (Generated Power MW).

4.3.Data Exploration and Analysis:

Some statistical analysis is required to get more deep understanding about the datasets. Tables (4.2), (4.3) and (4.4) show basic statistics (Minimum, Maximum, Mean, and Standard Deviation) for the most five important attributes of Unit 3, Unit 4 and Unit 3&4 datasets respectively. The attributes as shown in tables are:

- Steam_Flow_Inlet : Steam flow at turbine inlet
- Pressure_Inlet: Pressure at turbine Inlet
- Temperature_Inlet: Temperature at turbine Inlet
- Temperature_Outlet: Temperature at turbine Outlet
- Pressure_Outlet: Pressure at turbine Outlet
- Power_MW: Generated Power in Mega Watts (the class of all these datasets).

Simple comparison between unit 3 and 4 through these statistics can give some indications about the status of the unit itself. The standard deviation of Pressure_Inlet is 15.059 in unit 3 compared to 0.909 in unit 4, this is caused by the maximum value of pressure at unit 3 which is 116.963 bar. This will make direct impact about generated power values. Because the steam flow is fully controlled through valves, so the difference in standard deviation is not important. Regarding Temperature_Outlet, in unit 3 it is very high compared to unit 4, the maximum value in unit 3 is 84.893, while its counterpart in unit 4 is 66.628. Even the mean value is higher, in unit 3 it is 67.473, while in unit 4 it is 51.65. There is big difference between minimum values of Temperature_Outlet, it is 47.99 in unit 3 and 40.434 in unit 4.

Table (4.2) Unit 3 Dataset Analysis

Statistic	Steam_Flow	Pressure_Inlet	Temperature_Inlet	TemperatureOutlet	Pressure_Outlet	Power_MW
Minimum	46.879	34.686	498.064	47.99	0.085	39.482
Maximum	56.431	116.963	516.95	84.893	0.324	51.048
Mean	50.992	84.288	508.766	67.473	0.173	43.123
StdDev	2.429	15.059	2.83	10.715	0.068	3.746

Table (4.3) Unit 4 Dataset Analysis

Statistic	Steam_Flow	Pressure_Inlet	Temperature_Inlet	TemperatureOutlet	Pressure_Outlet	Power_MW
Minimum	29.981	84.769	485.746	40.434	0.059	28.073
Maximum	60.601	95.13	524.338	66.628	0.225	57.358
Mean	46.06	87.866	508.304	51.65	0.113	43.498
StdDev	8.185	0.909	4.758	5.797	0.036	7.676

Table (4.4) Unit 3&4 Dataset Analysis

Statistic	Steam_Flow	Pressure_Inlet	Temperature_Inlet	TemperatureOutlet	Pressure_Outlet	Power_MW
Minimum	29.981	34.686	485.746	40.434	0.059	28.073
Maximum	60.601	116.963	524.338	84.893	0.324	57.358
Mean	47.51	86.814	508.44	56.304	0.131	43.388
StdDev	7.353	8.354	4.286	10.461	0.055	6.762

4.4.Initial Comparison between Prediction Algorithms

The data types of all predictors and classes for all datasets used in this research are numeric. Therefore, according to data mining map shown in figure 2.7, the prediction models should be of regression type. Table 4.5 shows the list of algorithms that will be used for feature selection and prediction models for this research.

The purpose of this step is to do an initial comparison between all algorithms shown in Table 4.5, to select the best one for each dataset. Because there are three different datasets; three experiments were created “one for each dataset”. Each experiment uses 5 evaluation factors to rank the results:

- Mean_absolute_error,
- Root_mean_squared_error,
- Relative_absolute_error,
- Root_relative_squared_error
- Correlation coefficient.

Equations of these evaluation factors are shown in Figure 3.2.

Table (4.5) List of Prediction Algorithm Used in Research

Serial	Algorithm Name	Algorithm Name in Weka
1	Gaussian Processes for regression	Gaussian Processes
2	Isotonic Regression	Isotonic Regression
3	Least median squared linear regression	LeastMedSq
4	LinearRegression	Linear Regression
5	Neural Network	Multilayer Perceptron
6	Pace Regression linear models	PaceRegression
7	Partial Least Square Regression	PLSClassifier
8	Support Vector Machine for regression	SMOreg
9	K-nearest neighbors	IBk
10	Instance-based learner	KStar
11	Decision Table	Decision Table
12	M5Rules	M5Rules
13	M5 Model Tree	M5P
14	Decision tree learner C4.5	REPTree

Tables 4.6, 4.7, 4.8 shows the results of the initial comparison between the 14 algorithms for the three datasets respectively. Correlation coefficient is used to order the results of these algorithms in descending order, so the highest row in each table is the best performance algorithm, and the lowest is the worst one.

Tables 4.6, The Initial comparison between algorithms' accuracy for Unit 3 Dataset

Key	Algorithm	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Correlation coefficient.
4	LinearRegression	0.083879885	0.1115372	2.608101934	3.071199418	0.999534882
6	Pace Regression linear models	0.07947486	0.11195299	2.469352962	3.07517815	0.999514781
8	Support Vector Machine for regression	0.082229219	0.117748562	2.557921126	3.240612468	0.999469539
5	Neural Network	0.115456088	0.148388137	3.594696577	4.089735028	0.999381067
7	Partial Least Square Regression	0.110332296	0.157711502	3.431032399	4.346813619	0.999052812
9	K-nearest neighbors	0.153646078	0.224506865	4.778769266	6.162721846	0.997895543
10	Instance-based learner	0.160303297	0.24983911	4.98988145	6.866841346	0.997449099
12	M5Rules	0.22995974	0.343635614	7.150378626	9.431496547	0.995492844
1	Gaussian Processes for regression	0.364022514	0.56187269	11.32518817	15.4619605	0.992079159
13	M5 Model Tree	0.299358474	0.463936948	9.314740302	12.76900029	0.991104012
11	Decision Table	0.224577407	0.458042229	6.947455703	12.51621632	0.988891025
2	Isotonic Regression	0.7055718	1.217407346	21.99535882	33.60839564	0.941871625
3	Least median squared linear Reg.	0.655188556	1.28448858	20.46661714	35.50088054	0.941009965
14	Decision tree learner C4.5	0.456793077	1.242572655	14.23723128	34.21835838	0.933242152

Tables 4.7, The Initial comparison between algorithms' accuracy for Unit 4 Dataset

Key	Algorithm	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Correlation coefficient.
4	LinearRegression	0.090778698	0.132007423	1.462268216	1.716466237	0.999853405
6	Pace Regression linear models	0.09201714	0.133458434	1.482270317	1.735237895	0.999850051
8	Support Vector Machine for regression	0.109740875	0.153986194	1.767739042	2.002920995	0.999801781
5	Neural Network	0.16954587	0.214277192	2.736522943	2.789813196	0.999741401
7	Partial Least Square Regression	0.175533164	0.227428035	2.831145006	2.959030114	0.999565556
12	M5Rules	0.20758291	0.346568306	3.344873725	4.49947247	0.998899321
9	K-nearest neighbors	0.307747787	0.537363476	4.958080767	6.986130814	0.99743058
10	Instance-based learner	0.268448633	0.535891649	4.326801241	6.972246538	0.997424571
13	M5 Model Tree	0.288593321	0.521511918	4.653159421	6.768346851	0.997380193
1	Gaussian Processes for regression	0.551194325	0.828172041	8.882973574	10.76715419	0.995845461
2	Isotonic Regression	0.611010897	0.870444088	9.853861729	11.32469533	0.993554497
14	Decision tree learner C4.5	0.425977315	0.841559224	6.876083802	10.9490532	0.993466611
11	Decision Table	0.465122902	1.221040352	7.507875771	15.88393138	0.986015947
3	Least median squared linear Reg.	1.382632446	4.290088127	22.51536699	56.21080405	0.8411064

Tables 4.8 The Initial comparison between algorithms' accuracy for Unit 3&4 dataset

Key	Algorithm	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Correlation coefficient.
5	Neural Network	0.144291284	0.185629655	2.693371778	2.760036801	0.999713518
8	Support Vector Machine for regression	0.111605955	0.156056349	2.093530156	2.32791079	0.999708854
7	Partial Least Square Regression	0.256861072	0.342506278	4.818978739	5.107127214	0.998699469
9	K-nearest neighbours	0.252548263	0.453208317	4.746676916	6.784910832	0.997617218
15	M5Rules	0.244290404	0.451813449	4.597509927	6.767018336	0.997560331
16	M5 Model Tree	0.290096951	0.465959689	5.440535171	6.944972992	0.997520199
10	Instance-based learner	0.239184945	0.518880069	4.495455016	7.770865527	0.996817978
1	Gaussian Processes for regression	0.419571978	0.665304233	7.865066144	9.925619092	0.996303161
17	decision tree learner C4.5	0.372258075	0.730301402	6.982127768	10.88784093	0.993862586
14	DecisionTable	0.331897638	0.819727419	6.243007879	12.26145415	0.991850713
4	LinearRegression	0.42810000	1.30670000	0.07910800	0.19312000	0.98120000
6	pace regression linear models	0.44550000	1.31940000	0.08232000	0.19500200	0.98080000
3	Least median sqaured linear Reg.	0.543380789	1.739444185	10.03955402	25.82074251	0.939725333
2	Isotonic Regression	1.775236126	2.483967574	33.33455079	37.08256944	0.928511606

4.5. Feature Selection and Prediction Models for All datasets

Referring to chapter one (1.4 Objectives of Study), the first objective is: to *design a feature selection technique, that can determine the best set of features to predict the amount of generated power from a thermal power plant*." This objective could be divided into two parts:

1. Select the best set of parameters that influence the amount of the generated power from the power plant.
2. Build a prediction model to predict the amount of generated power using the selected parameters.

To achieve these goals, three complete datasets were used: unit 3, unit 4 and unit 3&4. These complete datasets contains all attributes shown in table (4.1). Normally, such goals are implemented in simple sequence: first step is to use Feature Selection method (Filter, Wrapper, Hybrid or Embedded) to select the best set of features, using the generated power as the class. Then the second step is to use only the selected features as predictors to build the prediction model. However, According to the used Feature Selection method, this sequence and algorithms change, For example:

- **Filter methods**: select attributes based on a performance measure regardless of the modeling algorithm, they are based only on general features like the correlation with the variable class. Only after selecting the best set of attributes, the modeling algorithm can use them.
- **Wrapper methods**: use predictive models to evaluate feature subsets. The evaluation is repeated for each subset using the algorithm that is used to develop the predictive model. Therefore, in wrapper the feature selection and prediction model development are wrapped together in one step, so, one algorithm will be used for feature selection and prediction model. More details about feature selection methods and algorithms is presented in chapter two (Literature Review).

In this research Wrapper Method will be used, so the feature selection and prediction model will be wrapped together in one step. As shown in figure 4.1, after selecting the best algorithm, the following will be done for each dataset:

i. Feature Selection to Select the Best Set of Features
ii. Design the Prediction Model to predict the Amount of Generated Power.
iii. Model Evaluation.

4.5.1 Feature Selection and Prediction Models for Unit 3 Dataset

According to the results of the initial comparison between models evaluation in table 4.6; the algorithm that shows the highest correlation co-efficient, and minimum errors in Unit 3 dataset is Paces Regression, while Decision tree C4.5 algorithm achieves the worst results. Using Unit 3 dataset, two models were designed; Pace Regression, and Decision Tree C4.5 models. Subsequent parts provide the design, evaluation, and discussion about these models.

4.5.1.1 Pace Regression Model

Wrapper Feature Selection method is used with Pace Regression to Select and evaluate the best set of features. Also Pace regression is used to create the prediction model, because in wrapper method the same algorithm which is used to evaluate the sub sets should be used to develop the prediction mode. Below is the details of Feature Selection, Prediction Model Design and the Model Evaluation.

i. **Feature Selection** : Table 4.9 shows the list of features which were selected and evaluated using Pace Regression algorithm for Unit 3. The table shows the unique feature number, and the feature name. As stated by the domain expert and according to thermodynamic laws, the most important features which are used to calculate the amount of the generated power, are five features (1: Main steam flow (kg/s), 3: Main steam header pressure, 4: T/A inlet steam temperature, 61: Condenser inlet exhaust steam pressure, 62: Condenser inlet exhaust steam temp) which are show in bold italic font in table 4.1. The model succeeded to select four of them, but feature 3 (Main steam header pressure) is missed. The total number of selected features are 28 (out of 62), this number is good compared to Linear Regression model which selected 47 features, this high number of features will lead to overfitting.

ii. **Power Prediction Model**: Figure 4.2 (a) shows the Pace Regression model for Unit 3 dataset. Features are the variables of the regression equation, the factor of each variable in the equation gives indication about the importance and relevancy of this variable to the class (*Amount of Generated Power in MW*). For example the factor of *CondenserInletExhaustSteamPressure_bar* is very high (9.3745), that means the pressure of

steam at condenser inlet (turbine outlet) has very big effect in determining the amount of the generated power, and this is absolutely true according to thermodynamic laws.

Table 4.9 List of Selected Features by Pace Regression Model for Unit 3

Feature No.	Feature Name	
1	***Main steam flow (kg/s)***	
2	Total steam flow (kg/s)	
4	***T/A inlet steam temperature***	
5	Main steam header steam temperature	
7	Feedwater temperature at economiser	
13	Condensate water flow (kg/s)	
14	Condenser hot well temperature a	
16	Auxiliary steam flow (kg/s)	
18	Auxiliary steam temperature	
19	Combustion air flow (Nm3/s)	
20	Air temperature at FDF inlet	
21	Air temperature at FDF inlet	
22	Air temperature after RAH (1)	
33	T/A wheel champer steam pressure	
36	T/A bleeder (3) pressure	
37	T/A bleeder (4) pressure	
40	T/A axial displacement A (mm)	
41	T/A axial displacement B (mm)	
46	T/A bearing 1 vibration (2) (mic)	
47	T/A bearing 2 vibration (1) (mic)	
48	T/A bearing 2 vibration (2) (mic)	
49	TBN side cold air	
51	Exciter side cold air	
52	Exciter side warm air	
57	Generator winding temperature 3	
59	Generator winding temperature 5	
61	***Condenser inlet exhaust steam pressure***	
62	***Condenser inlet exhaust steam temp***	
Total Number of Selected Features		28

Some features which were not been considered as important ones, appear in the regression equation with high factor like *T_AaxialdisplacementA_mm*, this high factor draw the attention of the domain expert, who took this as important input to be considered as a starting point for their investigation about the power plant problems.

Figure 4.2 (b) gives some basic information about the model and the used dataset. The model used 198 instances for training, while 102 different instances were used as a separate test set. The algorithm which is used to build the model is the

PaceRegression, of course it is the same one which was used to evaluate features' sub sets. The time required to build the model is 5.89 seconds.

GeneratedPower_MW =
-35 +
1.2611 * MainSteamFlow_kg_s +
-0.2399 * TotalSteamFlow_kg_s +
0.0748 * T_A_InletSteamTemperature_deg_C +
-0.0021 * MainSteamHeaderSteamTemperature_deg_C +
0.1272 * FeedwaterTemperatureAtEconomiserInlet_deg_ C +
-0.0026 * CondensateWaterFlow_kg_s +
-0.0375 * CondenserHotWellTemperature_deg_C_a +
0.0188 * CombustionAirFlow_Nm3_s +
-0.0706 * AirTemperatureAtFDFinlet_deg_C_1 +
0.0943 * AirTemperatureAtFDFinlet_deg_C_2 +
0.0103 * AirTemperatureAfterRAH_deg_C_1 +
-0.4966 * T_AwheelChamperSteamPressure_bar +
1.6736 * T_Ableeder_3_pressure_bar +
1.9366 * T_Ableeder_4_pressure_bar +
-1.0237 * T_AaxialdisplacementA_mm +
-0.0083 * T_Abearing1vibration_2_mic +
-0.0404 * T_Abearing2vibration_1_mic +
0.0178 * T_Abearing2vibration_2_mic +
-0.0003 * TBNsideColdAir_deg_C +
-0.03 * ExciterSideColdAir_deg_C +
-0.0897 * ExciterSideWarmAir_deg_C +
-0.0562 * GeneratorWindingTemperature_deg_C_3 +
0.0936 * GeneratorWindingTemperature_deg_C_5 +
-9.3745 * CondenserInletExhaustSteamPressure_bar_a +
-0.1648 * CondenserInletExhaustSteamTemperature_deg_C

Figure 4.2 Pace Regression Model for Unit

(a) Pace Regression Model fro Unit 3

Data set : Unit3
Algorithm : PaceRegression
Total Number of instances: 300
Training set: 198
Test set : 102
Time (s) : 5.89

(b) General Information about Model

Figure 4.2 Pace Regression Model for Unit 3

Table 4.10 Sample of comparison between Actual and Predicted values Of the Generated Power, using Pace Regression, for Test Dataset of Unit 3

inst No	actual	predicted	error
1	42.178	42.287	0.109
2	40.361	40.202	-0.159
3	45.245	45.212	-0.033
4	49.778	49.859	0.081
5	49.172	49.237	0.065
6	44.757	44.867	0.110
7	39.717	39.794	0.077
8	40.166	40.419	0.253
9	50.501	50.501	0.000
10	40.166	40.118	-0.048
11	45.109	45.148	0.039
12	39.619	39.731	0.112
13	40.186	40.194	0.008
14	39.893	39.722	-0.171
15	40.010	39.999	-0.011
16	40.029	40.023	-0.006
17	39.912	39.810	-0.102
18	49.074	49.052	-0.022
19	42.198	42.110	-0.088
20	45.148	45.057	-0.091
21	45.070	45.008	-0.062
22	40.166	40.109	-0.057
23	39.697	39.830	0.133
24	42.803	42.984	0.181
25	41.885	41.890	0.005
26	42.471	42.551	0.080
27	49.035	49.003	-0.032
28	50.071	50.013	-0.058
29	44.777	44.797	0.020
30	40.381	40.353	-0.028

Table 4.11 Pace Regression Model Accuracy for Unit 3 Data set

Correlation coefficient	0.9997
Mean absolute error	0.0711
Root mean squared error	0.0949
Relative absolute error	2.23%
Root relative squared error	2.62%
Total Number of Instances	102

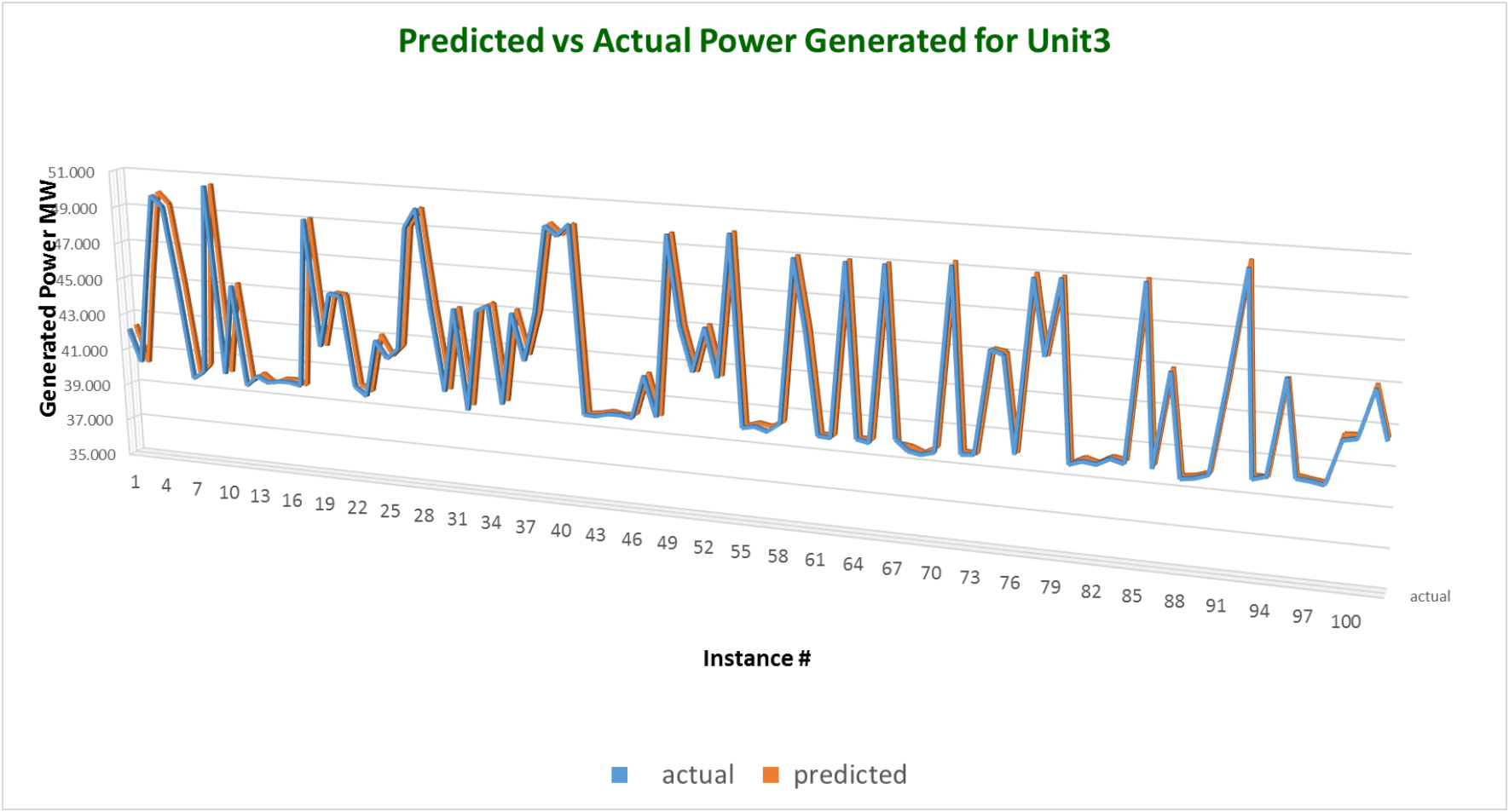


Figure 4.3 Graph for comparison between Actual and Predicted values Of the Generated Power,using Pace Regression, for Test Dataset of Unit 3

iii. **Model Evaluation**: Model testing is done by a separate test set, where 66.0% of the dataset is used for training, and the remainder for testing. The model evaluation is done using the equations in Figure 3.2 (Equations of the Evaluation methods). The model is considered to be more accurate if the Correlation coefficient is high (near to one), and the errors are low. Table 4.11 give the details about the model accuracy, as shown in the table, the correlation coefficient is high (0.9997) , and error factors are low (Mean absolute error is 0.0711, Root mean squared error is 0.0949, Relative absolute error is 2.23%, and the Root relative squared error is 2.62%). So, the the model accuracy is very high.

Table 4.10 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Pace Regression, for 30 instances of the Test Dataset of Unit 3, the difference column shows how much the predicted values are very near to the actual ones. Figure 4.2 gives more clear vision about the model accuracy, the graph is comparing between the actual and predicted values of the amount of the generated power, for the test dataset.

4.5.1.2 Decision tree learner C4.5 Model

Another model was built using Decision tree learner C4.5 algorithm, which showed very poor accuracy, this model was built only to compare its results with the first one which is very accurate.

i. **Feature Selection** : Table 4.12 shows the list of features which were selected and evaluated using Decision tree learner C4.5 algorithm for Unit 3. The table shows the unique feature number, and the feature name. The total number of selected features are 10 (out of 62). It is clear that this model failed to select the most five important features, it can select only one of them which is feature number 4 (*T/A inlet steam temperature).*

ii. **Power Prediction Model**: Figure 4.4 (a) shows the Decision tree learner C4.5 model for Unit 3 dataset. The figure shows the tree structure with the branching condition. The ten selected features are the tree nodes.

Table 4.12 List of Selected Features by Decision tree learner C4.5 Model for Unit 3

iii. Feature No.	Feature Name	
2	Total steam flow (kg/s)	
4	***T/A inlet steam temperature***	
8	Feedwater pressure at economiser inlet	
9	Condenser right inlet temperarure	
29	HPH4 inlet feedwater temperature	
35	T/A bleeder (2) pressure	
38	T/A bleeder (5) pressure	
39	T/A differential expansion (mm)	
43	T/A bearing 3 vibration (mm/s)	
60	Generator winding temperature 6	
Total Number of Selected Features		10

Figure 4.4 (b) gives some basic information about the model and the used dataset. The model used 198 instances for training, while 102 different instances were used as a separate test set. The algorithm which is used to build the model is the Decision tree learner C4.5, of course it is the same one which was used to evaluate features' sub sets. The time required to build the model is 5.89 seconds.

iv. **Model Evaluation**: The model evaluation is done using the equations in Figure 3.2 (Equations of the Evaluation methods). Table 4.14 gives the details about the model accuracy, as shown in the table, the correlation coefficient is high (0.9984), compared to (0.9997) for Pace Regression model. Error factors are higher than errors in Pace Regression model:

- Mean absolute error is 1.1606 compared to 0.0711 in Pace Regression,
- Root mean squared error is 0.2062 compared to 0.0949 in Pace Regression,
- Relative absolute error is 5.03 % compared to 2.23% in Pace Regression,
- Root relative squared error is 5.70% compared to 2.62% in Pace Regression.

So, it is clear that the model accuracy is lower than Pace Regression model. Table 4.13 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Decision tree learner C4.5, for 30 instances of the Test Dataset of Unit 3. The Error column shows how much the predicted values are very near to the actual ones. Figure 4.5 gives more clear vision about the model accuracy, the graph compares between the actual and predicted values of the amount of the generated power, for the test dataset. The difference between the actual and predicted values is clear from the graph.

GeneratedPower_MW =
==========
T_AdifferentialExpansion_mm < -0.27
| TotalSteamFlow_kg_s < 55.88
| | T_AdifferentialExpansion_mm < -0.81
| | | T_A_InletSteamTemperature_deg_C < 508.19
| | | | T_Abearing3vibration_mm_s < 5.85
| | | | | HPH4inletFeedwaterTemperature_deg_C < 142.25 : 39.93 (4/0) [2/0]
| | | | | HPH4inletFeedwaterTemperature_deg_C >= 142.25 : 39.75 (11/0.02) [3/0.01]
| | | | T_Abearing3vibration_mm_s >= 5.85 : 39.99 (6/0.01) [6/0.02]
| | | T_A_InletSteamTemperature_deg_C >= 508.19
| | | | CondenserRightInletTemperarure_deg_C < 29.32
| | | | | TotalSteamFlow_kg_s < 48.95 : 39.89 (6/0) [1/0.08]
| | | | | TotalSteamFlow_kg_s >= 48.95
| | | | | | GeneratorWindingTemperature_deg_C_6 < 112.75 : 39.76 (4/0.03) [3/0.03]
| | | | | | GeneratorWindingTemperature_deg_C_6 >= 112.75 : 40.03 (18/0) [2/0.01]
| | | | CondenserRightInletTemperarure_deg_C >= 29.32
| | | | | FeedwaterPressureAtEconomiserInlet_bar < 95.01 : 40.04 (7/0.01) [7/0.06]
| | | | | FeedwaterPressureAtEconomiserInlet_bar >= 95.01 : 40.3 (6/0.01) [4/0.01]
| | T_AdifferentialExpansion_mm >= -0.81
| | | T_AdifferentialExpansion_mm < -0.74
| | | | T_Ableeder_2_pressure_bar < 9.54
| | | | | T_A_InletSteamTemperature_deg_C < 509.94 : 44.81 (9/0.01) [3/0.01]
| | | | | T_A_InletSteamTemperature_deg_C >= 509.94 : 44.99 (4/0) [3/0.01]
| | | | T_Ableeder_2_pressure_bar >= 9.54 : 45.18 (8/0) [3/0.01]
| | | T_AdifferentialExpansion_mm >= -0.74
| | | | CondenserRightInletTemperarure_deg_C < 29.75 : 40 (37/0.03) [25/0.03]
| | | | CondenserRightInletTemperarure_deg_C >= 29.75
| | | | | TotalSteamFlow_kg_s < 51.38 : 42.14 (7/0.02) [1/0.02]
| | | | | TotalSteamFlow_kg_s >= 51.38
| | | | | | T_Ableeder_2_pressure_bar < 9.4 : 42.46 (8/0) [3/0.01]
| | | | | | T_Ableeder_2_pressure_bar >= 9.4 : 42.77 (6/0.02) [3/0.07]
| TotalSteamFlow_kg_s >= 55.88 : 45.01 (22/0.01) [8/0.02]
T_AdifferentialExpansion_mm >= -0.27
| CondenserRightInletTemperarure_deg_C < 26.74
| | T_Ableeder_5_pressure_bar < 0.05 : 49.8 (12/0.01) [9/0]
| | T_Ableeder_5_pressure_bar >= 0.05 : 50.5 (6/0.11) [3/0.05]
| CondenserRightInletTemperarure_deg_C >= 26.74 : 49.01 (19/0.01) [11/0.02]

(a) Decision tree learner C4.5 Model for Unit 3

Data set : Unit 3
Total Number of instances: 300
Training set: 198
Test set : 102
Algorithm : Decision tree learner C4.5
Time (s) : 5.22

(b) General Information about Model

Figure 4.4 Decision tree learner C4.5 Model for Unit 3

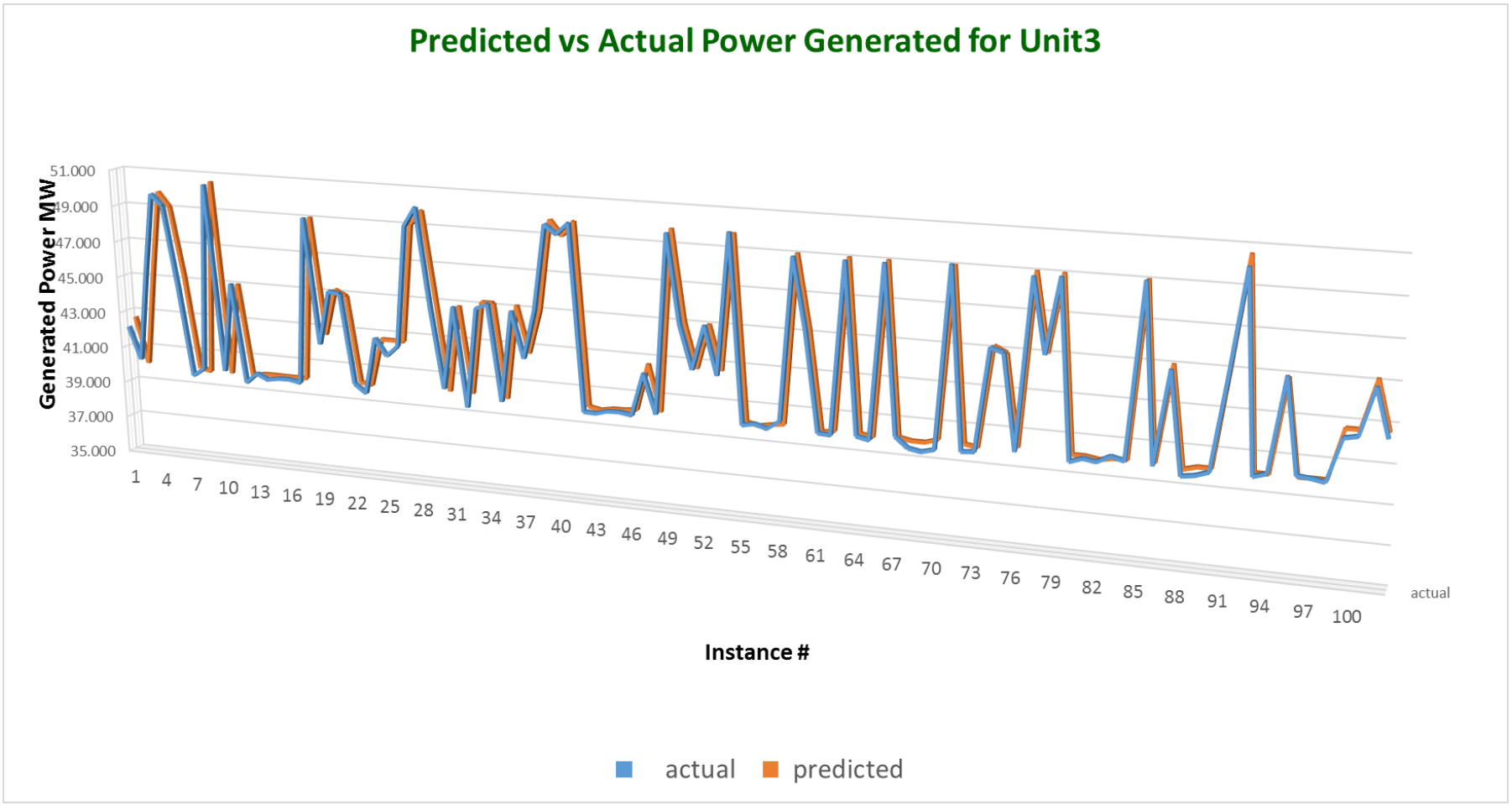


Figure 4.5 Graph for comparison between Actual and Predicted values of the Generated Power, using Decision tree learner C4.5, for Test Dataset of Unit 3

Table 4.13 Sample of comparison between Actual and Predicted values Of the Generated Power, using Decision tree learner C4.5, for Test Dataset of Unit 3

inst No	actual	predicted	error
1	42.178	42.580	0.402
2	40.361	39.993	-0.368
3	45.245	45.103	-0.142
4	49.778	49.786	0.008
5	49.172	48.989	-0.183
6	44.757	45.103	0.346
7	39.717	39.985	0.268
8	40.166	39.866	-0.300
9	50.501	50.559	0.058
10	40.166	39.866	-0.300
11	45.109	44.964	-0.145
12	39.619	39.866	0.247
13	40.186	39.985	-0.201
14	39.893	39.993	0.100
15	40.010	39.984	-0.026
16	40.029	39.984	-0.045
17	39.912	39.993	0.081
18	49.074	48.989	-0.085
19	42.198	42.580	0.382
20	45.148	45.103	-0.045
21	45.070	44.796	-0.274
22	40.166	40.076	-0.090
23	39.697	39.993	0.296
24	42.803	42.580	-0.223
25	41.885	42.580	0.695
26	42.471	42.580	0.109
27	49.035	48.989	-0.046
28	50.071	49.786	-0.285
29	44.777	44.964	0.187
30	40.381	40.076	-0.305

Table 4.14 Decision tree learner C4.5 Model Accuracy for Unit 3 Data set

Correlation coefficient	0.9984
Mean absolute error	0.1606
Root mean squared error	0.2063
Relative absolute error	5.03%
Root relative squared error	5.70%
Total Number of Instances	102

4.5.2 Feature Selection and Prediction Models for Unit 4 Dataset

According to results of the initial comparison between models evaluation in table 4.7; the algorithm that shows the highest correlation co-efficient, and minimum errors in Unit 4 dataset is Linear Regression, while Decision Table algorithm achieves the worst results. Using Unit 4 dataset, two models were designed; Linear Regression, and Decision Table models. Subsequent parts will provide the design, evaluation, and discussion about these models.

4.5.2.1 Linear Regression Model

Wrapper Feature Selection method is used with Linear Regression to Select and evaluate the best set of features. Also Linear regression is used to create the prediction model, because in wrapper method the same algorithm which is used to evaluate the sub sets should be used to develop the prediction mode. Below is the details of Feature Selection, Prediction Model Design and the Model Evaluation.

i. **Feature Selection** : Table 4.15 shows the list of features which were selected and evaluated using Linear Regression algorithm for Unit 4. The model succeeded to select the five features which are used by thermodynamic laws to calculate the amount of the generated power (1: Main steam flow (kg/s), 3: Main steam header pressure, 4: T/A inlet steam temperature, 61: Condenser inlet exhaust steam pressure, 62: Condenser inlet exhaust steam temp). The total number of selected features are 52 (out of 62), this number is high compared to Pace Regression model in Unit 3 which selected 28 features. This high number of features may lead to overfitting, but actually most of these attributes are uncontrollable, therefore, the benefit of this model is to study the hidden features which had not been considered before by the domain expert. Also this list of features can assist the efficiency engineers of the power plant to dig deep in the power plant problems and to study the effect of each feature.

ii. **Power Prediction Model**: the Linear Regression algorithm is represented as linear equation (Kenny and Keeping, 1962) as shown in equation 4.1.

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k \qquad (4.1)$$

where x is the class; $a_1, \ldots, a_k$ are the attribute values; and $w_0, w_1, \ldots, w_k$ are weights.

Figure 4.6 (a) shows the Linear Regression model for Unit 4 dataset. The selected features are the variables of the regression equation, the factor of each variable in the equation gives indication about the importance and relevancy of this variable to the class (Amount of Generated Power in MW). For example the factor of CondenserInletExhaustSteamPressure_bar is very high (14.1767), that means the pressure of steam at condenser inlet (turbine outlet) has very big effect in determining the amount of the generated power, and this is absolutely true according to

Table 4.15 List of Selected Features by Linear Regression Model for Unit 4

Feature No.	Feature Name	
1	***Main steam flow (kg/s)***	
2	Total steam flow (kg/s)	
3	***Main steam header pressure***	
4	***T/A inlet steam temperature***	
5	Main steam header steam temperature	
7	Feedwater temperature at economiser	
8	Feedwater pressure at economiser inlet	
9	Condenser right inlet temperarure	
11	Condenser right outlet temperarure	
12	Condenser left outlet temperarure	
13	Condensate water flow (kg/s)	
14	Condenser hot well temperature a	
15	Condenser hot well temperature b	
16	Auxiliary steam flow (kg/s)	
17	Auxiliary steam pressure	
18	Auxiliary steam temperature	
22	Air temperature after RAH (1)	
23	Air temperature after RAH (2)	
25	FDF A speed (rpm)	
26	FDF B speed (rpm)	
28	Air temperature after SAH (2)	
30	HPH4 outlet feedwater temperature	
31	HPH5 outlet feedwater temperature	
32	HPH5 inlet feedwater temperature	
33	T/A wheel champer steam pressure	
34	T/A bleeder (1) pressure	
35	T/A bleeder (2) pressure	
36	T/A bleeder (3) pressure	
37	T/A bleeder (4) pressure	
38	T/A bleeder (5) pressure	
39	T/A differential expansion (mm)	
40	T/A axial displacement A (mm)	
41	T/A axial displacement B (mm)	
42	T/A axial displacement C (mm)	
45	T/A bearing 1 vibration (1) (mic)	
46	T/A bearing 1 vibration (2) (mic)	
47	T/A bearing 2 vibration (1) (mic)	
48	T/A bearing 2 vibration (2) (mic)	
49	TBN side cold air	
50	TBN side warm air	
51	Exciter side cold air	
52	Exciter side warm air	
53	PMG side cold air	
54	PMG side warm air	
55	Generator winding temperature 1	
56	generator winding temperature 2	
57	Generator winding temperature 3	
58	Generator winding temperature 4	
59	Generator winding temperature 5	
60	Generator winding temperature 6	
61	***Condenser inlet exhaust steam pressure***	
62	***Condenser inlet exhaust steam temp***	
Total Number of Selected Features		52

thermodynamic laws. Also the pressure at turbine bleeders 1,2,3,4 and 5 have high factors, this is because these points are part of turbine outlet. Some features which were not been considered as important ones, appear in the regression equation with high factor like T_Aaxial displacement A, B and C (- 6.1079 * T_AaxialdisplacementA_mm + 9.2302 * T_AaxialdisplacementB_mm - 5.0426 * T_AaxialdisplacementC_mm), these high factor draw the attention of the domain expert because it appears also in Pace Regression model with Unit 3. The efficiency engineers took this as important input to be considered, and to be a starting point for their investigation about the power plant problems.

Figure 4.b (b) gives some basic information about the model and the used dataset. The model used 475 instances for training, while 254 different instances were used as a separate test set. The algorithm which is used to build the model is the Linear Regression, of course it is the same one which was used to evaluate features' sub sets. The time required to build the model is 73.67 seconds.

iii. **Model Evaluation**: Model testing is done by a separate test set, where 66.0% of the dataset is used for training, and the remainder for testing. The model evaluation is done using the equations in Figure 3.2 (Equations of the Evaluation methods). The model is considered to be more accurate if the Correlation coefficient is high (near to one), and the errors are low. Table 4.17 give the details about the model accuracy, as shown in the table, the correlation coefficient is high (0.9998) , and error factors are low (*Mean absolute error is 0.0928, Root mean squared error is 0.137, Relative absolute error is 1.51%, and the Root relative squared error is 1.81%*). So, the the model accuracy is very high.

GeneratedPower_MW =

0.8915 * MainSteamFlow_kg_s +
-0.9942 * TotalSteamFlow_kg_s +
0.0832 * MainSteamHeaderPressure_bar +
0.0117 * T_A_InletSteamTemperature_deg_C +
0.0045 *
MainSteamHeaderSteamTemperature_deg_C +
-0.0316 *
FeedwaterTemperatureAtEconomiserInlet_deg_ C +
-0.0172 * FeedwaterPressureAtEconomiserInlet_bar +
-0.3315 * CondenserRightInletTemperarure_deg_C +
0.0912 * CondenserRightOutletTemperarure_deg_C +
0.0974 * CondenserLeftOutletTemperarure_deg_C +
-0.0091 * CondensateWaterFlow_kg_s +
-0.0275 * CondenserHotWellTemperature_deg_C_a +
-0.0253 * CondenserHotWellTemperature_deg_C_b +
0.8423 * AuxiliarySteamFlow_kg_s +
0.2227 * AuxiliarySteamPressure_bar +
-0.1098 * AuxiliarySteamTemperature_deg_C +
-0.0329 * AirTemperatureAfterRAH_deg_C_1 +
0.0309 * AirTemperatureAfterRAH_deg_C_2 +
0.0052 * FDF_A_speed_rpm +
-0.0051 * FDF_B_speed_rpm +
0.0094 * AirTemperatureAfterSAH_deg_C_2 +
0.1397 * HPH4outletFeedwaterTemperature_deg_C +
0.0085 * HPH5outletFeedwaterTemperature_deg_C +
-0.0004 * HPH5inletFeedwaterTemperature_deg_C +
0.9519 * T_AwheelChamperSteamPressure_bar +
-1.7743 * T_Ableeder_1_pressure_bar +
3.2632 * T_Ableeder_2_pressure_bar +
0.1442 * T_Ableeder_3_pressure_bar +
3.4064 * T_Ableeder_4_pressure_bar +
-2.5871 * T_Ableeder_5_pressure_bar +
0.4359 * T_AdifferentialExpansion_mm +
-6.1079 * T_AaxialdisplacementA_mm +
9.2302 * T_AaxialdisplacementB_mm +
-5.0426 * T_AaxialdisplacementC_mm +
-0.0117 * T_Abearing1vibration_1_mic +
0.0084 * T_Abearing1vibration_2_mic +
-0.0067 * T_Abearing2vibration_1_mic +
0.0039 * T_Abearing2vibration_2_mic +
-0.0189 * TBNsideColdAir_deg_C +
0.0801 * TBNsideWarmAir_deg_C +
0.0892 * ExciterSideColdAir_deg_C +
-0.0444 * ExciterSideWarmAir_deg_C +
0.0001 * PMGsideColdAir_deg_C +
0.0102 * PMGsideWarmAir_deg_C +
-0.0047 * GeneratorWindingTemperature_deg_C_1 +
-0.0088 * generatorWindingTemperature_deg_C_2 +
-0.0103 * GeneratorWindingTemperature_deg_C_3 +
-0.0306 * GeneratorWindingTemperature_deg_C_4 +
-0.0415 * GeneratorWindingTemperature_deg_C_5 +
0.0549 * GeneratorWindingTemperature_deg_C_6 +
-14.1767 *
CondenserInletExhaustSteamPressure_bar_a +

(a) Linear Regression Model Equation for Unit 4

Data set : Unit 4
Total Number of instances: 720
Training set: 475
Test set : 245
Algorithm : Linear

(b) General Information about the Model

Figure 4.6 Pace Regression Model for Unit 4

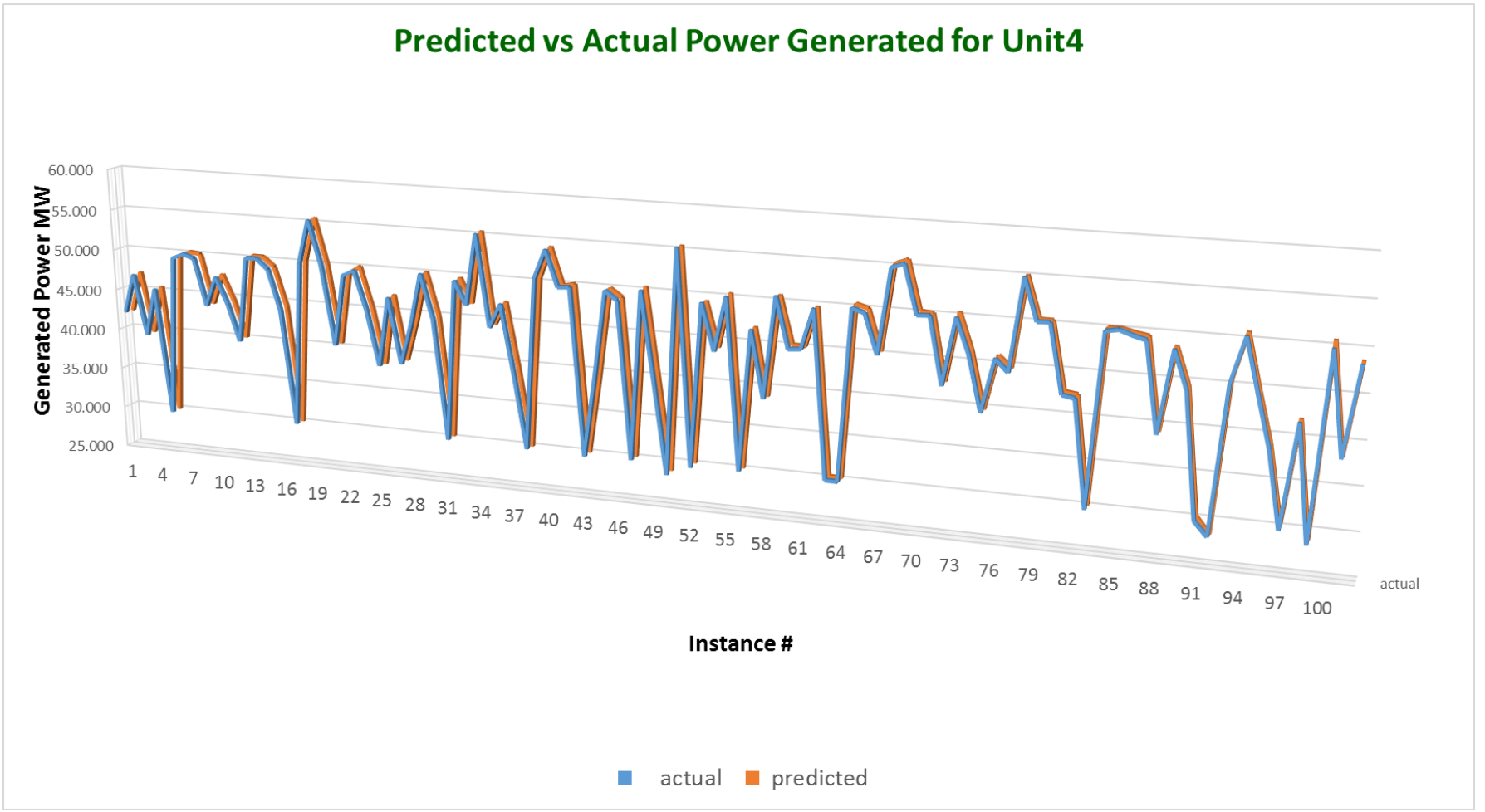


Figure 4.7 Graph for comparison between Actual and Predicted values Of the Generated Power, using Linear Regression, for Test Dataset of Unit 4

Table 4.16 Sample of comparison between Actual and Predicted values Of the Generated Power, using Linear Regression, for Test Dataset of Unit 4

Inst No.	Actual	Predicted	Error
1	42.139	42.084	-0.055
2	46.906	46.963	0.057
3	39.463	39.531	0.068
4	45.304	45.354	0.050
5	29.832	29.850	0.018
6	49.465	49.526	0.061
7	50.051	50.029	-0.022
8	49.602	49.843	0.241
9	43.858	43.840	-0.018
10	47.473	47.556	0.083
11	44.347	44.371	0.024
12	39.756	39.916	0.160
13	50.188	50.170	-0.018
14	50.227	50.149	-0.078
15	49.016	49.032	0.016
16	44.093	44.251	0.158
17	30.027	29.978	-0.049
18	50.149	50.183	0.034
19	55.424	55.451	0.027
20	50.071	49.856	-0.215
21	40.381	40.306	-0.075
22	49.055	49.040	-0.015
23	49.719	49.931	0.212
24	45.089	45.012	-0.077
25	38.349	38.249	-0.100
26	46.769	46.815	0.046
27	38.838	38.970	0.132
28	43.741	43.862	0.121
29	49.973	49.946	-0.027
30	44.620	44.702	0.082

Table 4.17 Linear Regression Model Accuracy for Unit 4 Data set

Correlation coefficient	0.9998
Mean absolute error	0.0928
Root mean squared error	0.137
Relative absolute error	1.51%
Root relative squared error	1.81%
Total Number of Instances	245

Table 4.16 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Linear Regression, for 30 instances of the Test Dataset of Unit 4, the Error column shows how much the predicted values are very near to the actual ones. Figure 4.7 gives more clear vision about the model accuracy, the graph is comparing between the actual and predicted values of the amount of the generated power, for the test dataset.

4.5.2.2 Decision Table Model

Another model was built using Decision Table algorithm (Kohavi, 1995), which showed very poor accuracy with Unit 4 dataset, as shown in table 4.7 which shows the initial comparison between algorithms' accuracy for Unit 4 Dataset. This model was built only to compare its results with the Linear Regression model which achieved very high accuracy.

i. **Feature Selection** : Table 4.18 shows the list of features which were selected and evaluated using Decision Table algorithm for Unit 4. The total number of the selected features are 4 (out of 62). It is clear that this model failed to select the most five important features.

ii. **Power Prediction Model**: Figure 4.8 (a) shows the Decision table model for Unit 4 dataset. The figure shows the number of the generated rules, but the model can not assist in explaining the behavior of the power plant. Figure 4.8 (b) gives some basic information about the model and the used dataset. The model used 475 instances for training, while 245 different instances were used as a separate test set. The algorithm which is used to build the model is the Decision Table, it is the same one which was used to evaluate features' sub sets. The time required to build the model is 121.72 seconds.

iii. **Model Evaluation**: The model evaluation is done using the equations in Figure 3.2 (Equations of the Evaluation methods). Table 4.20 gives the details about the model accuracy, as shown in the table, the correlation coefficient is high (0.9917), compared to (0.9998) for Linear Regression model. Error factors are higher than errors in Linear Regression model:

- Mean absolute error is 0.3645 compared to 0.0928 in Linear Regression,
- Root mean squared error is 0.9718 compared to 0.137 in Linear Regression,
- Relative absolute error is 5.94 % compared to 1.51% in Linear Regression,
- Root relative squared error is 12.85% compared to 1.81% in Linear Regression.

So, it is clear that the model accuracy is lower than Linear Regression model.

Table 4.19 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Decision Table, for 30 instances of the Test Dataset of Unit 4. The Error column shows how much the predicted values are very far from the actual ones. Figure 4.9 gives more clear vision about the model accuracy, the graph compares between the actual and predicted values of the amount of the generated power, for the test dataset. The difference between the actual and predicted values is clear from the graph.

Table 4.18 List of Selected Features by Decision Table Model for Unit 4

Feature No.	Feature Name	
32	HPH5 inlet feedwater temperature	
34	T/A bleeder (1) pressure	
55	Generator winding temperature 1	
59	Generator winding temperature 5	
Total Number of Selected Features		4

GeneratedPower_MW =
Classifier Model
Decision Table:
Number of training instances: 720
Number of Rules : 54
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 11
Merit of best subset found: 0.884
Evaluation (for feature selection): CV (leave one out)
Feature set: 1,2,3,4,5

(1) Decision Table Model for Unit 4

Data set : Unit4
Total Number of instances: 720
Training set: 475
Test set : 245
Algorithm :

(b) General Information about the Model

Figure 4.8 Decision Table Model for Unit 4

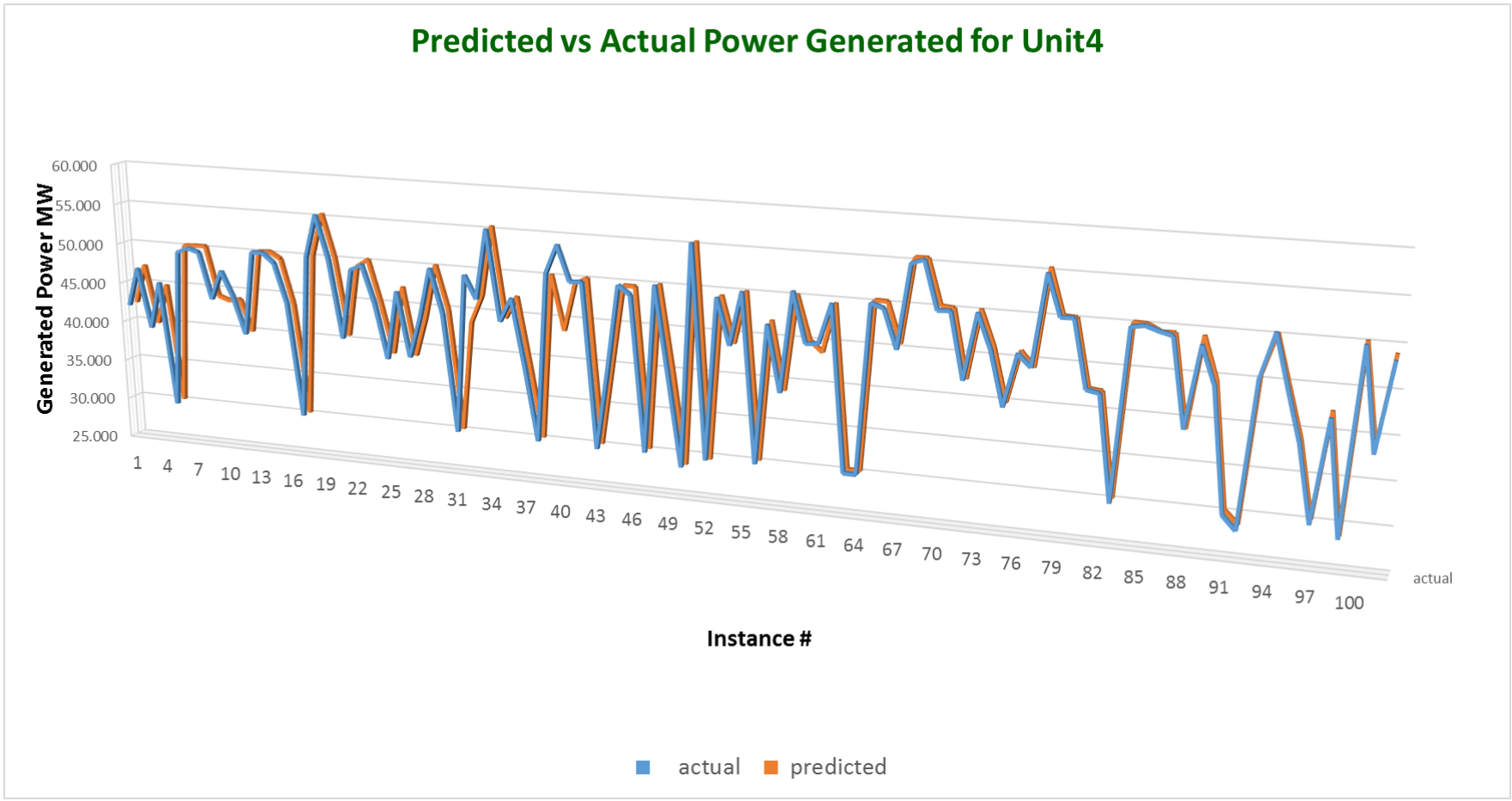


Figure 4.9 Graph for comparison between Actual and Predicted values Of the Generated Power, using Decision Table, for Test Dataset of Unit 4

Table 4.19 Sample of comparison between Actual and Predicted values Of the Generated Power, using Decision Table, for Test Dataset of Unit 4

Instance No	Aactual	Predicted	Error
1	42.139	42.196	0.057
2	46.906	47.043	0.137
3	39.463	39.752	0.289
4	45.304	44.714	-0.590
5	29.832	30.027	0.195
6	49.465	49.995	0.530
7	50.051	50.085	0.034
8	49.602	50.085	0.483
9	43.858	43.936	0.078
10	47.473	43.438	-4.035
11	44.347	43.712	-0.635
12	39.756	39.752	-0.004
13	50.188	49.995	-0.193
14	50.227	50.085	-0.142
15	49.016	49.270	0.254
16	44.093	43.712	-0.381
17	30.027	30.026	-0.001
18	50.149	50.085	-0.064
19	55.424	55.314	-0.110
20	50.071	49.995	-0.076
21	40.381	40.400	0.019
22	49.055	49.177	0.122
23	49.719	50.085	0.366
24	45.089	44.856	-0.233
25	38.349	38.691	0.342
26	46.769	47.043	0.274
27	38.838	38.691	-0.147
28	43.741	43.610	-0.131
29	49.973	50.085	0.112
30	44.620	44.714	0.094

Table 4.20 Decision Table Model Accuracy for Unit 4 Data set

Correlation coefficient	0.9917
Mean absolute error	0.3645
Root mean squared error	0.9718
Relative absolute error	5.94%
Root relative squared error	12.85%
Total Number of Instances	245

4.5.3 Feature Selection and Prediction Models for Unit 3&4 dataset

Unit 3&4 dataset is a new dataset that combines Unit 3 and Unit 4 datasets, the purpose of this model is to check whether there is a generic prediction model regardless the unit, and to check if there is a similarity in selected sets of features. According to results of the initial comparison between models evaluation in table 4.8; the algorithm that shows the highest correlation co-efficient, and minimum errors in Unit 3&4 dataset is Neural Network (Multilayer Perceptron), while Isotonic Regression algorithm achieves the worst results. Using Unit 3&4 dataset, two models were designed; Neural Network, and Isotonic Regression models. Subsequent parts provide the design, evaluation, and discussion about these models.

4.5.3.1 Neural Network Model

Wrapper Feature Selection method is used with Neural Network to Select and evaluate the best set of features. Also Neural Network (Multi Layer Perceptron) is used to create the prediction model (Hastie, Tibshirani and Friedman, 2009). Below is the details of Feature Selection, Prediction Model Design and the Model Evaluation.

i. **Feature Selection** : Table 4.21 shows the list of features which were selected and evaluated using Neural Network for Unit 3&4 dataset. . The total number of the selected features are 18 (out of 62). The model can select only three (1: Main steam flow (kg/s), 4: T/A inlet steam temperature, 62: Condenser inlet exhaust steam temp) of the five features which are used by thermodynamic laws to calculate the amount of the generated power. Therefore, the model failed to select the required features.

ii. **Power Prediction Model**: For power prediction model, three Neural Network models were created. The first used 1 hidden layer with 9 neurons, the second used 3 hidden layers with (9,18,9) neurons, and the third used 3 hidden layers with (9,18,36) neurons. Figure 4.10 shows the design of the first Neural Network, in witch the 18 selected features forms the input layer, and the class (Generated_Power_MW) is the output layer. Weights are learned from the training set. Iteratively minimizing the error using steepest decent. The gradient is determined using the back-propagation algorithm. The change in weight is computed by multiplying the gradient by the learning rate (0.3) and adding the previous change in weight multiplied by the momentum (0.2), equation 4.2 below shows how the weight is calculated:

$$W_{next} = W + \Delta W$$
$$\Delta W = -\,learning\_rate \times gradient \;+\; momentum \times \Delta W_{previous} \qquad (4.2)$$

Where : $\mathbf{W_{next}}$ is the new weight , $\mathbf{W}$ is the weight, $\triangle\mathbf{W}$ is the change in weight.

Although the models shows high accuracy, but it is difficult for the Neural Network model to assist in explaining the behavior of the power plant. Also the model can select only three of the features used to calculate the power, for these two reasons the model is not accepted. The model used 1020 instances for training. The time required to build the model is longer than other methods.

Table 4.21 List of Selected Features by Neural Network (MLP) Model for Unit 3&4 Dataset

Feature No.	Feature Name	
1	***Main steam flow (kg/s)***	
4	***T/A inlet steam temperature***	
6	HPH5 discharge feedwater flow	
9	Condenser right inlet temperarure	
14	Condenser hot well temperature a	
16	Auxiliary steam flow (kg/s)	
23	Air temperature after RAH (2)	
29	HPH4 inlet feedwater temperature	
30	HPH4 outlet feedwater temperature	
35	T/A bleeder (2) pressure	
36	T/A bleeder (3) pressure	
40	T/A axial displacement A (mm)	
42	T/A axial displacement C (mm)	
45	T/A bearing 1 vibration (1) (mic)	
51	Exciter side cold air	
57	Generator winding temperature 3	
60	Generator winding temperature 6	
62	***Condenser inlet exhaust steam temp***	
Total Number of selected Features		18

i. **Model Evaluation**: Model testing is done using 10 fold cross validation. The model evaluation is done using the equations in Figure 3.2 (Equations of the Evaluation methods). Table 4.23 compares between the accuracy of three Neural Network models. The first model which used only one hidden layer achieved higher performance, as shown in the table, the correlation coefficient is high (0.9995) , and error factors are low (*Mean absolute error is 0.1689, Root mean squared error is 0.2244, Relative absolute error is 3.13%, and the Root relative squared error is 3.32%*). So, the model accuracy is higher than it's counterparts. Table 4.22 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Neural Network with 1 hidden layer, for 30 instances of the Test Dataset of Unit 3&4, the Error column shows how much the predicted values are very near to the actual ones. Figure 4.11 gives more clear vision about the model accuracy, the graph is comparing between the actual and predicted values of the amount of the generated power, for the test dataset.

Table 4.22 Sample of comparison between Actual and Predicted values Of the Generated Power, using Neural Network (1 Hidden Layer), for Test Dataset of Unit 3&4 Dataset

instance No	Actual	Predicted	Error
1	45.070	45.043	-0.027
2	48.801	48.865	0.064
3	29.968	29.835	-0.133
4	50.833	50.434	-0.399
5	44.190	44.265	0.075
6	50.012	50.054	0.042
7	44.249	44.384	0.135
8	47.238	46.938	-0.300
9	39.775	39.856	0.081
10	39.502	39.600	0.098
11	38.857	39.033	0.176
12	50.012	49.840	-0.172
13	40.010	40.023	0.013
14	56.928	56.716	-0.212
15	46.750	46.916	0.166
16	48.938	48.874	-0.064
17	49.700	49.658	-0.042
18	30.144	30.018	-0.126
19	49.153	49.179	0.026
20	46.808	46.685	-0.123
21	49.993	50.332	0.339
22	50.012	49.811	-0.201
23	39.990	39.951	-0.039
24	39.756	39.821	0.065
25	28.777	28.690	-0.087
26	39.775	40.185	0.410
27	49.758	49.713	-0.045
28	44.093	44.207	0.114
29	39.853	39.665	-0.188
30	39.893	40.027	0.134

Table 4.23 Comparison between Neural Network Models' Accuracy for Unit 3&4 Dataset

Number of Hidden Layers	**1**	**3**	**3**
Number of Neurons	**9**	**9, 18, 9**	**9, 18, 36**
Correlation coefficient	0.9995	0.9992	0.9994
Mean absolute error	0.1689	0.2047	0.1742
Root mean squared error	0.2244	0.2707	0.2311
Relative absolute error	3.13%	3.80%	3.23%
Root relative squared error	3.32%	4.00%	3.42%
Total Number of Instances	1020	1020	1020

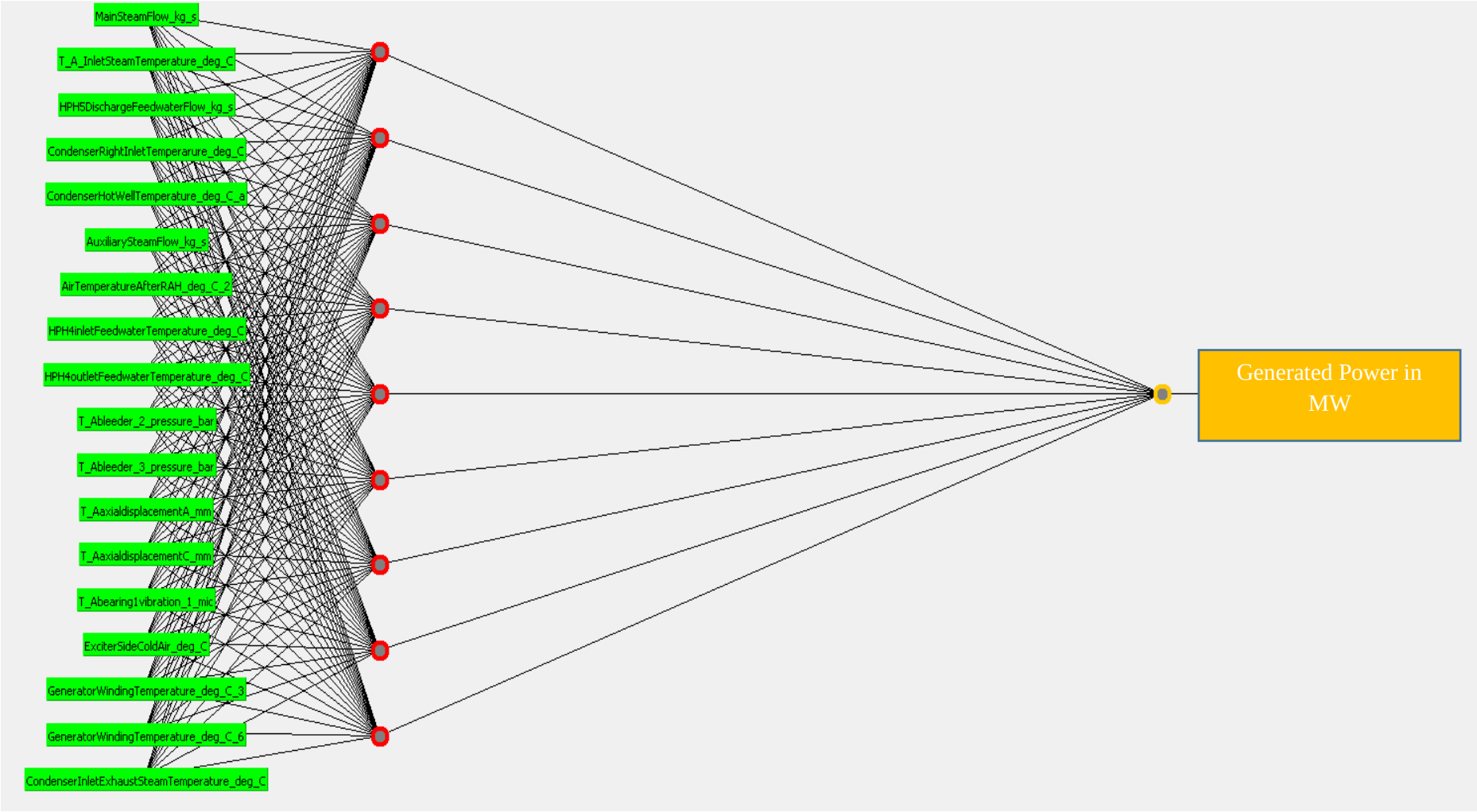


Figure 4.10 Neural Network Model for Unit 4 Data set

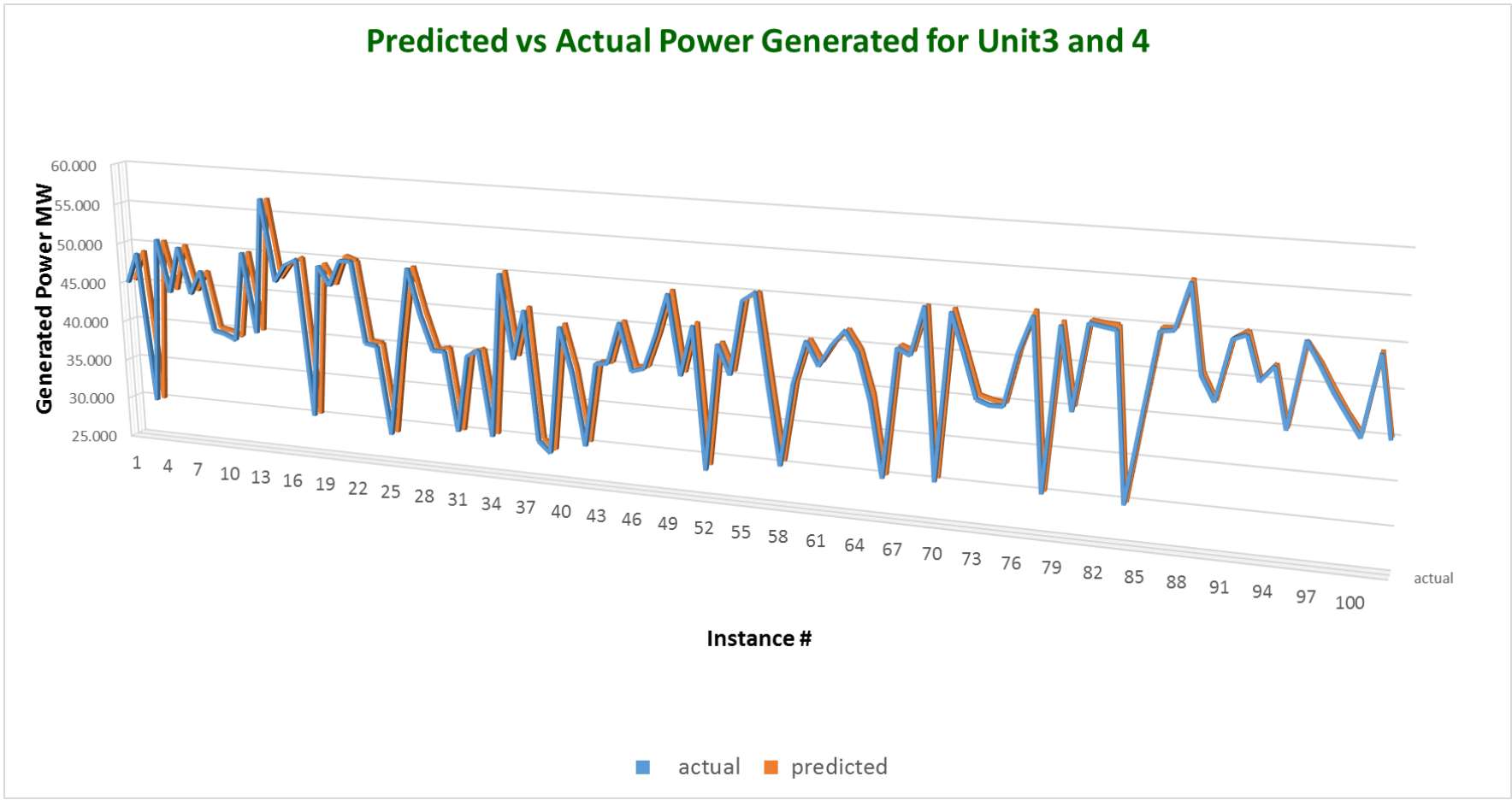


Figure 4.11 Graph for comparison between Actual and Predicted values Of the Generated Power, using Neural Network (1 Hidden Layer), for Test Dataset of Unit 3&4 Dataset

4.5.3.2 Isotonic Regression Model

Following the same practice, another model was built using Isotonic Regression (Jan, Kurt and Patrick, 2009), which is selected from the bottom of the performance comparison table for Unit 3&4 dataset in table 4.8.

i. **Feature Selection** : When wrapper method is applied for Unit 3&4 dataset using Isotonic Regression algorithm, only feature number 30 was selected (HPH4 outlet feedwater temperature). So, so the model failed in feature selection.

ii. **Power Prediction Model**: The model selected used one feature (HPH4 outlet feedwater temperature), non of the most important features were selected and used for prediction, so regardless of the model accuracy, the model is also not accepted.

iii. **Model Evaluation**: Table 4.25 gives the details about the model accuracy, as shown in the table, the correlation coefficient is high (0.9232), compared to (0.9997) for Neural Network model. Error factors are higher than errors in Neural Network model:

- Mean absolute error is 1.774 compared to 0.1294 in Linear Regression,
- Root mean squared error is 2.6102 compared to 0.1694 in Linear Regression,
- Relative absolute error is 32.84 % compared to 2.39% in Linear Regression,
- Root relative squared error is 38.58% compared to 2.50% in Linear Regression.

So, it is clear that the model accuracy is lower than Neural Network model. Table 4.24 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Isotonic Regression, for 30 instances of the Test Dataset of Unit 3&4. The Error column shows how much the predicted values are very far from the actual ones. Figure 4.12 gives more clear vision about the model accuracy, the graph compares between the actual and predicted values of the amount of the generated power, for the test dataset. The difference between the actual and predicted values is clear from the graph.

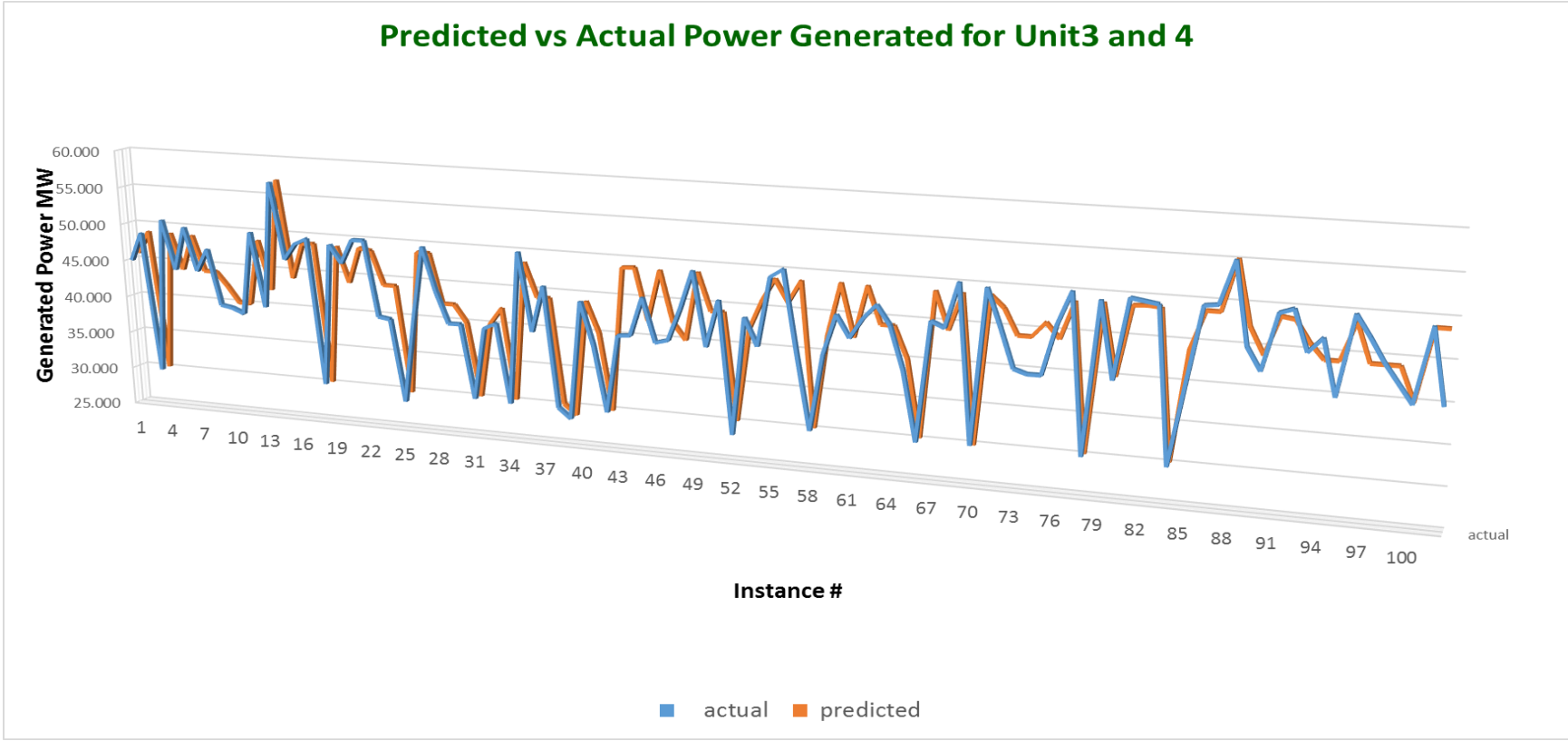


Figure 4.12 Graph for comparison between Actual and Predicted values Of the Generated Power, using Isotonic Regression or Test Dataset of Unit 3&4 Dataset

Table 4.24 Sample of comparison between Actual and Predicted values Of the Generated Power, using Istonic Regression, for Test Dataset of Unit 3&4 Dataset

inst#	actual	predicted	error
1	45.070	45.770	0.700
2	48.801	48.731	-0.070
3	29.968	30.016	0.048
4	50.833	48.731	-2.102
5	44.190	43.889	-0.301
6	50.012	48.619	-1.393
7	44.249	43.889	-0.360
8	47.238	43.889	-3.349
9	39.775	42.039	2.264
10	39.502	39.876	0.374
11	38.857	39.876	1.019
12	50.012	48.619	-1.393
13	40.010	42.039	2.029
14	56.928	56.980	0.052
15	46.750	43.889	-2.861
16	48.938	48.619	-0.319
17	49.700	48.731	-0.969
18	30.144	30.044	-0.100
19	49.153	48.619	-0.534
20	46.808	43.889	-2.919
21	49.993	48.619	-1.374
22	50.012	48.355	-1.657
23	39.990	43.889	3.899
24	39.756	43.889	4.133
25	28.777	29.646	0.868
26	39.775	48.619	8.844
27	49.758	48.619	-1.139
28	44.093	42.039	-2.054
29	39.853	42.039	2.186

Table 4.25 Istonic Regression Accuracy for Unit 3&4 Dataset

Correlation coefficient	0.9232
Mean absolute error	1.7774
Root mean squared error	2.6102
Relative absolute error	32.84%
Root relative squared error	38.58%
Total Number of Instances	347

4.6 Discussion about Features' Selection Results

All results of attribute selection models are collected in table 4.26, the table is ordered by the last column (Rank). The first column in the table is the attribute ID, the second is the feature name. The rest of the columns are organized in three groups, one for each dataset. If the attribute is selected by an algorithm, the number "1" will be written at the corresponding cell; else the cell will be null. The last column of the table (Rank) is the summation of ones, that represents the number of algorithms that had selected this feature, this will give an indication about the influence of this feature in determining the amount of generated power, if the number is high that means this attribute has been selected by many algorithms, and that means this attribute has a high effect in determining the amount of the generated power. That is absolutely true with the *Main steam flow (kg/s)*, and *T/A inlet steam temperature* which are belong to the top 5 features list (features that are used to calculate the amount of power), this could be an evidence to the correctness of the feature selection results. Table 4.26 gave impressive results to domain expert, who spent a lot of time discussing these results. One of these findings is about feature number 41 (*T/A axial displacement B (mm)*) which had never been taken in consideration by the efficiency engineers. So table 4.27 and 4.28 are very important analysis reports that give the Power Plant Efficiency engineers indication about the status of the power plant.

Unit 4 results is also provided in a table 4.27 separately, to give more clear vision about this unit. That is because Unit 4 dataset was is the most accurate data, and the Linear Regression model when used with this dataset succeeded in selecting the top 5 features (features that are used to calculate the power).

Table 4.26 Attribute selection summary of all units

		Unit 3					Uni 4				Unit 3&4						
Feature No.	**Feature Name**	Linear Regression	PaceRegression	SMOreg	Neural Network	REPTree	Linear Regression	Pace Regression	Neural Network	Isotonic Reg.	Pace Regression	Linear Regression	Neural Network	IBK	M5Rules	Isotonic Reg	**Total (Rank)**
1	***Main steam flow (kg/s)***	**1**	**1**	**1**	**1**		**1**	**1**	**1**			**1**	**1**	**1**	**1**		**11**
4	***T/A inlet steam temperature***	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**		**1**	**1**	**1**				**11**
37	T/A bleeder (4) pressure	1	1	1	1		1	1			1	1					9
2	Total steam flow (kg/s)	1	1	1	1	1	1	1				1					8
30	HPH4 outlet feedwater temperature	1					1	1	1		1	1	1		1	1	8
36	T/A bleeder (3) pressure	1	1	1			1	1	1		1	1	1				8
59	Generator winding temperature 5	1	1	1			1	1		1	1	1					8
62	***Condenser inlet exhaust steam temp***	**1**	**1**	**1**			**1**		**1**		**1**	**1**	**1**		**1**		**8**
13	Condensate water flow (kg/s)	1	1	1			1	1				1			1		7
14	Condenser hot well temperature a	1	1				**1**	**1**	**1**	**1**		**1**	**1**				**7**
16	Auxiliary steam flow (kg/s)		1	1			1	1	1			1	1				7
33	T/A wheel champer steam pressure	1	1	1	1		1	1				1					7
34	T/A bleeder (1) pressure	1					1	1		1	1	1			1		7
41	T/A axial displacement B (mm)		1	1	1		1	1			1	1					7
46	T/A bearing 1 vibration (2) (mic)	1	1	1	1		1					1			1		7
48	T/A bearing 2 vibration (2) (mic)	1	1	1	1		1	1				1					7
61	***Condenser inlet exhaust steam press***	**1**	**1**	**1**			**1**	**1**			**1**	**1**					**7**
18	Auxiliary steam temperature		1	1	1		1	1				1					6
22	Air temperature after RAH (1)	1	1	1			1	1				1					6
29	HPH4 inlet feedwater temperature				1	1		1	1		1	1	1				6
32	HPH5 inlet feedwater temperature	1					1	1		1	1	1					6
38	T/A bleeder (5) pressure	1				1	1	1				1					6
47	T/A bearing 2 vibration (1) (mic)	1	1	1			1	1				1					6
50	TBN side warm air						1	1			1	1			1		6
51	Exciter side cold air		1	1			1		1			1	1				6
52	Exciter side warm air	1	1	1	1		1					1					6
57	Generator winding temperature 3	1	1				1		1			1	1		1		6
5	Main steam header steam temperature	1	1				1				1	1					5
9	Condenser right inlet temperarure	1				1	1		1			1	1				5
11	Condenser right outlet temperarure			1			1	1				1					5
12	Condenser left outlet temperarure	1					1	1				1					5
15	Condenser hot well temperature b	1		1			1	1				1					5
23	Air temperature after RAH (2)						1	1	1			1	1	1			5
28	Air temperature after SAH (2)	1		1			1	1									5
31	HPH5 outlet feedwater temperature	1		1			1				1	1					5
35	T/A bleeder (2) pressure					1	1	1	1			1	1				5
42	T/A axial displacement C (mm)	1					1	1	1			1	1				5
49	TBN side cold air	1	1		1		1					1					5
54	PMG side warm air	1					1					1			1		5
55	Generator winding temperature 1	1					1			1		1			1		5
56	generator winding temperature 2	1					1	1				1					5
58	Generator winding temperature 4	1					1	1				1					5
7	Feedwater temperature at economiser	1	1				1					1					4
8	Feedwater pressure at economiser inlet				1	1	1					1					4
17	Auxiliary steam pressure			1			1	1				1					4
19	Combustion air flow (Nm3/s)	1	1					1			1						4
21	Air temperature at FDF inlet	1	1	1				1									4
25	FDF A speed (rpm)	1					1	1				1					4
26	FDF B speed (rpm)	1					1	1				1					4
40	T/A axial displacement A (mm)	1	1				1		1				1				4
43	T/A bearing 3 vibration (mm/s)					1		1				1					4
3	***Main steam header pressure***						**1**	**1**			**1**						**3**
10	Condenser left inlet temperarure	1		1								1					3
27	Air temperature after SAH (1)	1										1					3
39	T/A differential expansion (mm)	1				1	1										3
45	T/A bearing 1 vibration (1) (mic)						1		1			1	1				3
60	Generator winding temperature 6					1	1		1				1				3
6	HPH5 discharge feedwater flow	1							1				1				2
20	Air temperature at FDF inlet	1	1														2
24	FDF discharge air pressure (mbar)	1										1					2
53	PMG side cold air	1					1										2
44	T/A bearing 4 vibration (mm/s)	1															1
Sum		**47**	**28**	**26**	**13**	**10**	**52**	**38**	**18**	**4**	**16**	**51**	**18**	**2**	**10**	**1**	

Table (4.27) Attribute selection summary of Unit 4

Feature No.	Feature Name	Linear Regression	Pace Regression	Neural Network	Isotonic Reg.	Total (Rank)
30	HPH4 outlet feedwater temperature	1	1	1	1	**4**
14	Condenser hot well temperature a	**1**	**1**	**1**		**3**
1	***Main steam flow (kg/s)***	**1**	**1**	**1**		**3**
4	***T/A inlet steam temperature***	**1**	**1**	**1**		**3**
16	Auxiliary steam flow (kg/s)	1	1	1		**3**
23	Air temperature after RAH (2)	1	1	1		**3**
35	T/A bleeder (2) pressure	1	1	1		**3**
36	T/A bleeder (3) pressure	1	1	1		**3**
42	T/A axial displacement C (mm)	1	1	1		**3**
32	HPH5 inlet feedwater temperature	1	1			**2**
34	T/A bleeder (1) pressure	1	1			**2**
59	Generator winding temperature 5	1	1			**2**
2	Total steam flow (kg/s)	1	1			**2**
3	***Main steam header pressure***	**1**	**1**			**2**
9	Condenser right inlet temperarure	1		1		**2**
11	Condenser right outlet temperarure	1	1			**2**
12	Condenser left outlet temperarure	1	1			**2**
13	Condensate water flow (kg/s)	1	1			**2**
15	Condenser hot well temperature b	1	1			**2**
17	Auxiliary steam pressure	1	1			**2**
18	Auxiliary steam temperature	1	1			**2**
22	Air temperature after RAH (1)	1	1			**2**
25	FDF A speed (rpm)	1	1			**2**
26	FDF B speed (rpm)	1	1			**2**
28	Air temperature after SAH (2)	1	1			**2**
29	HPH4 inlet feedwater temperature		1	1		**2**
33	T/A wheel champer steam pressure	1	1			**2**
37	T/A bleeder (4) pressure	1	1			**2**
38	T/A bleeder (5) pressure	1	1			**2**
40	T/A axial displacement A (mm)	1		1		**2**
41	T/A axial displacement B (mm)	1	1			**2**
45	T/A bearing 1 vibration (1) (mic)	1		1		**2**
47	T/A bearing 2 vibration (1) (mic)	1	1			**2**
48	T/A bearing 2 vibration (2) (mic)	1	1			**2**
50	TBN side warm air	1	1			**2**
51	Exciter side cold air	1		1		**2**
56	generator winding temperature 2	1	1			**2**
57	Generator winding temperature 3	1		1		**2**
58	Generator winding temperature 4	1	1			**2**
60	Generator winding temperature 6	1		1		**2**
61	***Condenser inlet exhaust steam pressure***	**1**	**1**			**2**
62	***Condenser inlet exhaust steam temp***	**1**		**1**		**2**
55	Generator winding temperature 1	1				**1**
5	Main steam header steam temperature	1				**1**
6	HPH5 discharge feedwater flow			1		**1**
7	Feedwater temperature at economiser	1				**1**
8	Feedwater pressure at economiser inlet	1				**1**
19	Combustion air flow (Nm3/s)		1			**1**
21	Air temperature at FDF inlet		1			**1**
31	HPH5 outlet feedwater temperature	1				**1**
39	T/A differential expansion (mm)	1				**1**
43	T/A bearing 3 vibration (mm/s)		1			**1**
46	T/A bearing 1 vibration (2) (mic)	1				**1**
49	TBN side cold air	1				**1**
52	Exciter side warm air	1				**1**
53	PMG side cold air	1				**1**
54	PMG side warm air	1				**1**
10	Condenser left inlet temperarure					**0**
20	Air temperature at FDF inlet					**0**
24	FDF discharge air pressure (mbar)					**0**
27	Air temperature after SAH (1)					**0**
44	T/A bearing 4 vibration (mm/s)					**0**
Sum		**52**	**38**	**18**	**1**	

4.7 Discussion about the Power Prediction Models

According to thermodynamic laws, and as stated by the domain experts, there are five features used to calculate the amount of generated power, whether you are using manufacturer consumption graph, or power equation. This top 5 list contains : *1. Main steam flow, 3. Pressure at Turbine Inlet, 4. Temperature at Turbine Inlet, 61. Pressure at Turbine Outlet, 62. Temperature at Turbine Outlet*. So, any prediction model that is not using these features will not be acceptable. As seen from table 4.30, the models that match this constraint is Linear Regression for Unit 4. Also this model attained the highest correlation coefficient (0.9998) and lowest MAE (0.0928) in unit 4, so Linear Regression model for unit 4 is the most acceptable model.

From table 4.29 we can observe that feature 3 (*Main steam header pressure*) is not selected by any of the algorithms in unit 3 dataset, this observation is directly related to the high Standard Deviation (15.059) which was observed for the same feature, see Table 4.2 Unit 3 Dataset Analysis. From this observation we can highlight that there is an issue in the amount of pressure at turbine inlet of unit 3, which is the reason behind the noisy data of Unit 3.

		Unit 3					Unit 4				Unit 3&4						
Feature No.	**Feature Name**	Linear Regression	PaceRegression	SMOreg	Neural Network	REPTree	Linear Regression	Pace Regression	Neural Network	Isotonic Reg.	Pace Regression	Linear Regression	Neural Network	IBK	M5Rules	Isotonic Reg	**Total (Rank)**
1	***Main steam flow (kg/s)***	1	1	1	1		1	1	1			1	1	1	1		**11**
4	***T/A inlet steam temperature***	1	1	1	1	1	1	1	1		1	1	1				**11**
62	***Condenser inlet exhaust steam temp***	1	1	1			1		1		1	1	1		1		**8**
61	***Condenser inlet exhaust steam pressure***	1	1	1			1	1			1	1					**7**
3	***Main steam header pressure***						1	1			1						**3**
Sum		**4**	**4**	**4**	**2**	**1**	**5**	**4**	**3**	**0**	**4**	**4**	**3**	**1**	**2**	**0**	

Table 4.28 Attribute selection summary of for the Top 5 features in Unit 3 and 4

As shown in table 4.28 , no algorithm in Unit 3&4 dataset succeeded to select the Top 5 features. Because this dataset is composed of both unit 3 and unit 4 dataset, it will inherit all problems observed in unit 3. Hence, non of prediction models of Unit 3&4 dataset are not using the top 5 features. So, we can conclude that, Unit 3&4 dataset could not be used as generic dataset, and any unit should be studied separately to predict the generated power.

4.8 Summary

The calculation of the amount of generated power from a thermal power plant is done using only five features which are (*1. Main steam flow, 3. Pressure at Turbine Inlet, 4. Temperature at Turbine Inlet, 61. Pressure at Turbine Outlet, 62. Temperature at Turbine Outlet*). The objective of this chapter is "*To design a feature selection technique, that can determine the best set of features to predict the amount of generated power from a thermal power plant*". To achieve this objective, 3 datasets were used with 14 prediction algorithms and Wrapper method of feature selection. The models that showed higher correlation coefficient and minimum errors were selected to design the prediction models. Results from all these experiments were discussed with the domain expert, who highlighted many interesting findings to be as starting point for further investigation about the power plant status.

Linear Regression which was used with Unit 4 dataset attained the highest Correlation Coefficient (0.9998) and lowest MAE (0.0928) , so Linear Regression model with unit 4 is the most acceptable model. The model used 52 attribute to predict the power, although this number is relatively high, but the model gave accurate prediction and succeeded in determining the other features with their influence (weight) to the amount of generated power.

CHAPTER FIVE

POWER PREDICTION MODELS USING CONTROLLABLE PARAMETERS

5.1 Introduction

The target of this chapter is "*To design a prediction technique that can accurately predicts the amount of the generated power from a thermal power plant, using only the controllable parameters"*. In the previous chapter, Feature Selection techniques were used to select a set of features, then the selected set was used to predict the amount of the generated power. However, most of these selected features are non controllable parameter. The controllable parameters are (*1. Main steam flow, 3. Pressure at Turbine Inlet, 4. Temperature at Turbine Inlet*) , the efficiency engineers can control the amount of these parameters by adding more energy to the boiler and controlling the steam valves at turbine inlet. But, the other two parameters which are needed to calculate the amount of power (*61. Pressure at Turbine Outlet, 62. Temperature at Turbine Outlet*) are non controllable. The problem is that; the actual amount of the generated power is always different from the expected and calculated values, using either the Steam Consumption Graph or the thermodynamic laws. This chapter presents a solution for this problem by designing a prediction model that uses only the controllable parameters. This model could be used as a tool to accurately control the amount of power generated from a thermal power plant.

To achieve this goal, two datasets were used: one for unit 3, the other for unit 4. This chapter starts by describing the datasets, then basic statistical analysis about these datasets is shown. After that an initial comparison between the prediction algorithms is done for each dataset, to select the most appropriate algorithm for each dataset to build the prediction model. Then for each dataset two main tasks were done: the development of the power prediction model and the model evaluation. After that a discussion about models, then a summary is provided. The methodology followed to achieve the goal is shown in figure 5.1.

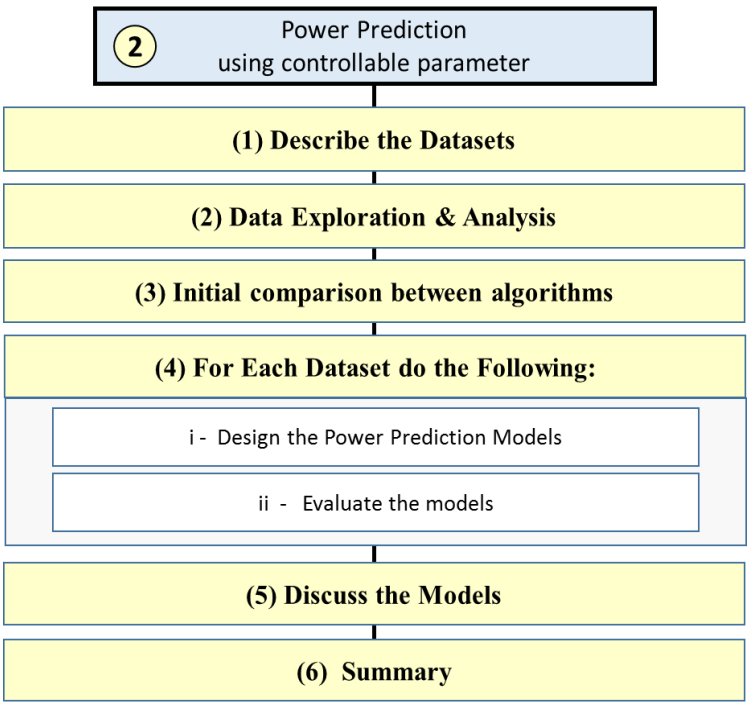


Figure 5.1 Methodology for Problem (2)

5.2 Datasets description

Two datasets were prepared to build Power Prediction Models using controllable parameters, one for each unit. Steam properties at turbine inlet could be controlled by the amount of heat submitted to the boiler. Hence, steam flow, pressure and temperature are the controllable parameters. Only these three controllable parameters were selected to prepare the datasets for this model. Each dataset is prepared by selecting three or more different instances (readings) from each month in the original dataset. To compare the actual amount of the generated power for this selected datasets versus the calculated values, the amount of the generated power is calculated manually using the two methods; the thermodynamic laws and the Steam Consumption Graph. Below is some description about these datasets:

1. Controllable parameters dataset for Unit 3: As shown in Table 5.1 this dataset is composed of three predictors (Steam Flow, Pressure and Temperature) and one class (Power). This dataset is composed of 87 instances.

2. Controllable parameters dataset for Unit 4 : As shows in Table 5.2 this dataset is also composed of three predictors (Steam Flow, Pressure and Temperature)and one class (Power). This dataset is composed of 83 instances.

Table 5.1 Controllable parameters dataset for Unit 3

Steam Flow at Turbine Inlet	Pressure at Turbine Inlet	Temperature at Turbine Inlet	Generated Power in MW
36.213	85.932	508.91	30.125
36.277	87.256	504.597	29.89
36.17	86.822	507.631	30.144
47.706	86.907	511.416	40.147
47.971	88.914	510.847	40.42

Table 5.2 Controllable parameters dataset for Unit 4

Steam Flow at Turbine Inlet	Pressure at Turbine Inlet	Temperature at Turbine Inlet	Power in MW
30.994	87.291	512.103	29.91
30.701	86.939	509.371	29.988
30.888	87	508.326	30.066
44.886	87.496	508.675	45.011
44.458	86.864	518.263	44.835

5.3 Data Exploration and Analysis for Datasets

Some statistical analysis is required to get more deep understanding about the datasets (Saed, 2017) . Tables 5.3 and 5.4 show basic statistics about attributes of Unit 3 and Unit 4 datasets respectively. The class of these datasets is the generated power in MW. Simple comparison between unit 3 and 4 through these statistics can give some indications about the status of the unit itself. The same observation regarding StdDev of Unit 3 pressure appeared here. The standard deviation of Pressure in unit 3 dataset is 28.875 compared to 1.198 in unit 4, this is caused by the maximum value of pressure at unit 3 which is 128.782 bar. This will make direct impact on amount of the generated power value. The mean value of pressure in unit 3 is 75.291 bar, while the optimum value for pressure (as specified by fabricants) should be 87 bar. Unit 4 statistics is normal and very near to proposed values by fabricants.

Table 5.3 Unit 3 Dataset Analysis

Statistic	Steam Flow	Pressure	Temperature	Power
Minimum	24.569	0.4	498.064	19.849
Maximum	59.373	128.782	530.501	50.637
Mean	45.402	75.291	508.93	38.409
StdDev	8.153	28.875	4.203	7.553

Table 5.4 Unit 4 Dataset Analysis

Statistic	Steam Flow	Pressure	Temperature	Power
Minimum	27.266	84.804	493.465	26.022
Maximum	60.021	91.042	524.338	56.928
Mean	45.064	87.675	509.902	42.767
StdDev	9.322	1.198	4.8	8.668

5.4 Initial Comparison Between Algorithms

The data types of all predictors and classes for all datasets used in this research are numeric. Therefore, according to data mining map shown in figure 2.7, the prediction models should be of regression type. Table 5.5 shows the list of algorithms that will be used for power prediction models in this chapter.

Table 5.5 List of Algorithms Used for Power Prediction Models

Serial	Algorithm Name	Algorithm Name in Weka
1	Gaussian Processes for regression	GaussianProcesses
2	Isotonic Regression	IsotonicRegression
3	Least median squared linear regression	LeastMedSq
4	LinearRegression	LinearRegression
5	Neural Network	MultilayerPerceptron
6	Pace Regression linear models	PaceRegression
7	Simple Linear Regression	SimpleLinearRegression
8	Support Vector Machine for regression	SMOreg
9	K-nearest neighbors	IBk
10	Instance-based learner	KStar
11	Locally weighted learning	LWL
12	Conjunctive Rule	ConjunctiveRule
13	Decision Table	DecisionTable
14	M5Rules	M5Rules
15	One-level decision tree	DecisionStump
16	M5 Model Tree	M5P
17	Decision tree learner C4.5	REPTree

The purpose of this step is to do an initial comparison between all algorithms shown in Table 5.5, to select the best one for each dataset. Because there are two different datasets; two experiments were created “one for each dataset”. Each experiment uses 5 evaluation factors to rank the results:

- Mean_absolute_error,
- Root_mean_squared_error,
- Relative_absolute_error,

- Root_relative_squared_error
- Correlation coefficient.

Equations of these evaluation factors are shown in Figure 3.2. Tables 5.6,5.7 shows the results of the initial comparison between the 17 algorithms for the three datasets respectively. Correlation coefficient is used to order the results of these algorithms in descending order, so the highest row in each table is the best performance algorithm, and the lowest is the worst one.

5.5 Prediction Models to Predict the Power Using Controllable Parameters

Referring to chapter one (1.4 Objectives of Study), the second objective is: t*o design a prediction technique that can accurately predicts the amount of generated power, using only the controllable parameters.* This objective could be achieved by building a prediction model to predict the amount of generated power using the controllable parameters. To achieve this goal, two datasets were used: unit 3, unit 4. Each of these datasets compose of three predictors (*1. Main steam flow, 3. Pressure at Turbine Inlet, 4. Temperature at Turbine Inlet*), as shown in tables 5.1 and 5.2.

As shown in figure 5.1, after the initial comparison between algorithms, the following will be done for each dataset:

i. Design the Prediction Model to predict the Amount of Generated Power: The model will be designed using the algorithm that showed the highest Correlation Coefficient in the initial comparison. Another model will be designed using the algorithm that showed the worst Correlation Coefficient, just to compare its results with the best algorithm.
ii. Model Evaluation: the created models will be evaluated using the same factors used in comparing the algorithms (Mean_absolute_error, Root_mean_squared_ error, Relative_absolute_error, Root_relative_squared_error and correlation coefficient). Because the dataset is small, evaluation is done using 10-fold cross validation. Also the predicted results will be compared with the actual values, and the comparison results will be presented as a table and as a graph to give better vision about the models' accuracy.

Tables 5.6 The Initial comparison between algorithms' accuracy for Unit 3 Dataset

No.	Algorithm Name	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
6	**Pace Regression linear models**	**0.9383**	**2.1045**	**2.598**	**32.5636**	**34.2217**
1	Gaussian Processes for regression	0.9381	2.2890	2.8501	35.4179	37.5422
16	M5 Model Tree	0.9379	2.0848	2.6059	32.2593	34.3258
14	M5 Rules	0.9378	2.0973	2.6106	32.4528	34.3874
4	LinearRegression	0.9375	2.1031	2.6128	32.5425	34.4169
8	Support Vector Machine for regression	0.9374	2.1269	2.6592	32.9098	35.0282
2	Isotonic Regression	0.9371	1.9029	2.6219	29.4443	34.5362
7	Simple Linear Regression	0.9241	2.3287	2.8694	36.0325	37.7969
13	Decision Table	0.9216	2.2394	2.9157	34.6513	38.4061
9	K-nearest neighbors	0.9189	1.6682	2.9801	25.8128	39.2545
17	Decision tree learner C4.5	0.9164	2.1505	3.0152	33.2754	39.7166
3	Least median squared linear regression	0.9015	2.6101	3.3594	40.3865	44.2514
5	Neural Network	0.8998	2.6154	3.3355	40.4688	43.9355
10	Instance-based learner	0.8856	2.4211	3.7060	37.4625	48.8163
11	Locally weighted learning	0.8652	3.0524	3.7723	47.2315	49.6903
12	Conjunctive Rule	0.8636	3.0787	3.7870	47.6377	49.8832
15	One-level decision tree	0.8558	3.1496	3.8868	48.7345	51.1977

Tables 5.7 The Initial comparison between algorithms' accuracy for Unit 4 Dataset

No.	**Algorithm Name**	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
2	**Isotonic Regression**	**0.9943**	**0.6440**	**0.9224**	**8.5690**	**10.6071**
8	Support Vector Machine for regression	0.9915	0.8501	1.1267	11.3126	12.9567
3	Least median squared linear regression	0.991	0.9107	1.1893	12.1188	13.6770
4	LinearRegression	0.991	0.9157	1.1534	12.1845	13.2638
6	Pace Regression linear models	0.991	0.9073	1.1562	12.0728	13.2964
7	Simple Linear Regression	0.9896	1.0013	1.2377	13.3236	14.2333
14	M5 Rules	0.9894	0.9533	1.2522	12.6861	14.4000
16	M5 Model Tree	0.9893	0.9648	1.2571	12.8379	14.4567
17	Decision tree learner C4.5	0.9893	0.8271	1.2592	11.0057	14.4813
13	Decision Table	0.9881	0.9644	1.3264	12.8326	15.2538
10	Instance-based learner	0.9864	0.9616	1.4347	12.7954	16.4986
5	Neural Network	0.9857	1.1365	1.4871	15.1233	17.1016
1	Gaussian Processes for regression	0.9854	1.4071	1.6934	18.7246	19.4742
9	K-nearest neighbors	0.981	1.0184	1.6711	13.5519	19.2180
11	Locally weighted learning	0.921	2.5788	3.3608	34.3166	38.6498
15	One-level decision tree	0.9006	3.0655	3.7443	40.7924	43.0597
12	Conjunctive Rule	0.8902	3.1135	3.9343	41.4317	45.2443

5.5.1 Power Prediction Models Using Controllable Parameters for Unit 3

According to the results of the initial comparison between algorithms in table 5.6; the algorithm that shows the highest correlation co-efficient, and minimum errors in Unit 3 dataset is Paces Regression, while One-level decision tree algorithm achieves the worst results. Using Unit 3 dataset, two models were designed; Pace Regression, and One-level decision tree. Subsequent parts provide the design, evaluation, and discussion about these models.

5.5.1.1 Pace Regression Model

Pace regression improves the classical ordinary least squares regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their contribution to the overall regressions (Wang, 2000). Pace regression is selected to develop the prediction model for Unit 3 because it showed the highest correlation coefficient, and minimum errors. Below is the details about the model and its evaluation.

i. **Power Prediction Model**: Figure 5.2 (a) shows the Pace Regression model for Unit 3 dataset. As seen from the figure the features are the variables of the linear equation. Figure 5.2 (b) gives some basic information about the model and the used dataset. The model used 10-fold cross validation for evaluation, because the number is only 87 which is relatively low. The algorithm which is used to build the model is the Pace. The time required to build the model is less than a second.

Pace Regression Model
GeneratedPower_MW =
-111.0877 +
0.8909 * MainSteamFlow +
0.0294 * PressureInlet +
0.2099 * TemperatureInlet

(a) Pace Regression Model for Unit 3

Data set : Unit3
Total Number of instances: 87
Training set: 87
Evaluation : 10-fold Cross-validation
Algorithm : Pace Regression
Time (s) : 0

(b) General Information about the model

Figure 5.2 Pace Regression Model for Unit 3

Table 5.9 Sample of comparison between Actual and Predicted values Of the Generated Power, using Pace Regression, for Test Dataset of Unit 3

Instance No.	Actual	Predicted	Error
1	41.885	39.928	-1.957
2	40.127	40.306	0.179
3	49.172	45.25	-3.922
4	39.912	43.453	3.541
5	32.625	34.189	1.564
6	42.081	42.939	0.858
7	34.09	33.719	-0.371
8	39.736	44.393	4.657
9	36.005	36.58	0.575
10	42.12	43.947	1.827
11	47.902	43.865	-4.037
12	48.938	44.309	-4.629
13	30.066	26.192	-3.874
14	31.785	33.635	1.85
15	48.098	43.969	-4.129
16	44.757	44.757	0
17	44.991	48.379	3.388
18	29.949	25.798	-4.151
19	40.127	39.571	-0.556
20	40.303	43.369	3.066
21	40.029	45.151	5.122
22	34.95	34.913	-0.037
23	40.166	40.511	0.345
24	40.147	41.502	1.355
25	30.144	29.538	-0.606
26	42.295	40.917	-1.378
27	39.658	41.908	2.25
28	40.42	41.368	0.948
29	35.126	35.121	-0.005
30	45.773	46.464	0.691

Table 5.8 Pace Regression Model Accuracy for Unit 3 Data set

Correlation coefficient	0.9383
Mean absolute error	2.1045
Root mean squared error	2.598
Relative absolute error	32.56%
Root relative squared error	34.22%

i. **Model Evaluation**: Model evaluation is done using 10-fold cross validation. Each fold is done using the equations in Figure 3.2 (Equations of the Evaluation methods). The model is considered to be more accurate if the Correlation coefficient is high (near to one), and the errors are low. Table 5.8 give the details about the model accuracy, as shown in the table, the correlation coefficient is high (0.9383) , and error factors are low (Mean absolute error is 2.1045, Root mean squared error is 2.598, Relative absolute error is 32.56%, and the Root relative squared error is 34.22%). The model accuracy is low, this is due to the

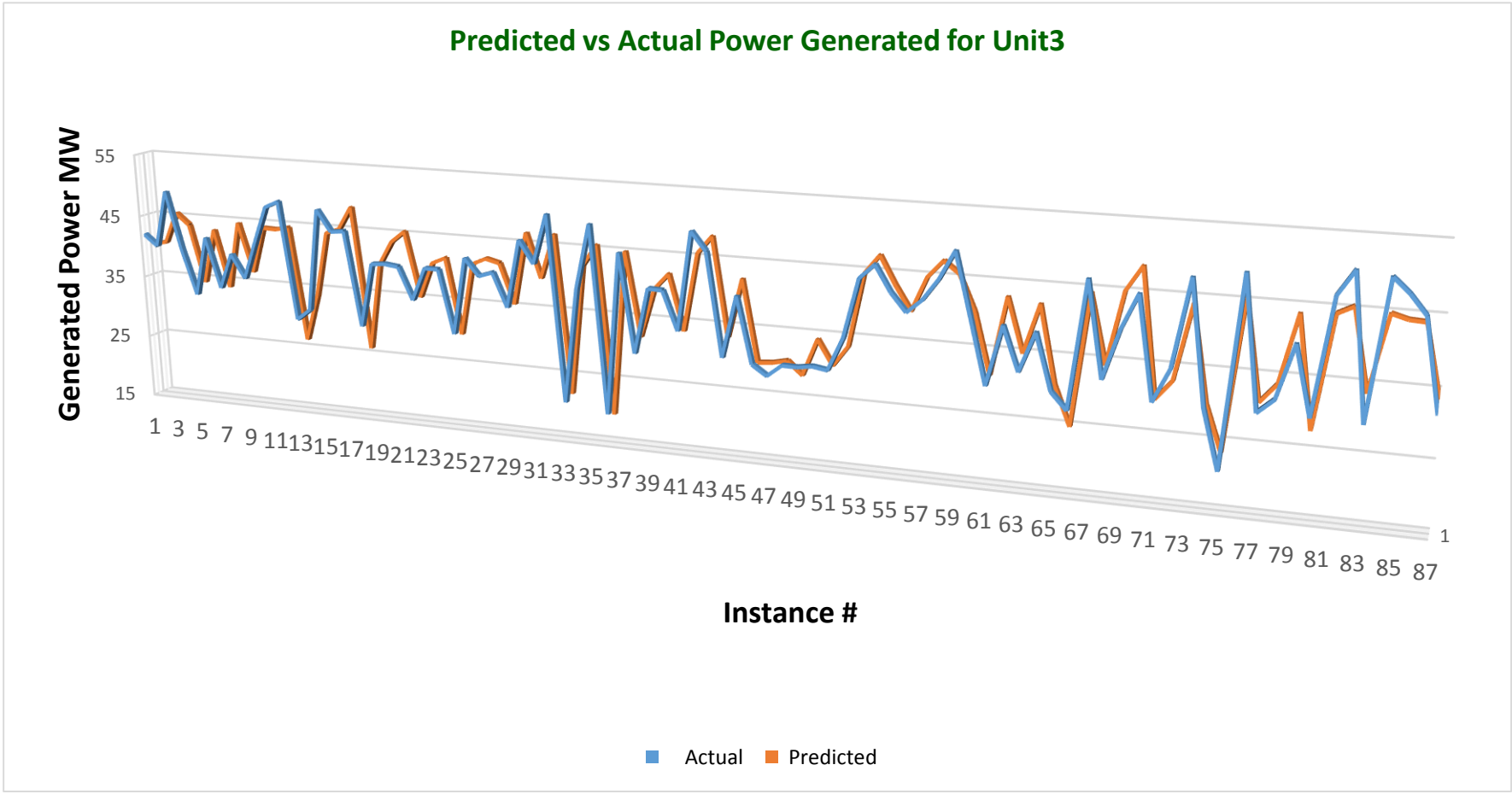


Figure 5.3 Graph for comparison between Actual and Predicted values Of the Generated Power, using Pace Regression, for Test Dataset of Unit 3

instability of sensors' readings of unit 3 which was observed in chapter 4. Table 5.9 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Pace Regression, for 30 instances of Unit 3, the Error column shows how much the predicted values are far from from the actual ones. Figure 4.3 gives more clear vision about the model accuracy, the graph compares between the actual and predicted values of the amount of the generated power.

5.5.1.2 One-level Decision Tree Model

Another model was built using One-level Decision Tree algorithm, which is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves) (Iba et al, 1992). The model showed very poor accuracy, this model was built only to compare its results with the first one which is the most accurate model in Unit 3. Below is the details about the One-level Decision Tree model and it evaluation.

i. **Power Prediction Model**: Figure 5.4 (a) shows the One-level Decision Tree model for Unit 3 dataset. As seen from the figure, the One-level Decision Tree algorithm selected only the *Main Steam Flow* as its root node. The model calculates the power as an *if-then-else* statement:

> If the (MainSteamFlow <= 44.4925) Then Generated Power = 30.47;
> Elsif (MainSteamFlow > 44.4925) Then Generated Power =: 43.75;
> Elsif (MainSteamFlow is missing) Then Generated Power =38.40;

The model doesn't assist in depicting the behavior of the power plant, and can not give a reliable prediction method.

Decision Stump Model
Classifications
MainSteamFlow <= 44.4925 : 30.470600000000008
MainSteamFlow > 44.4925 : 43.751346153846164
MainSteamFlow is missing : 38.40851724137932

(a) One-level Decision Tree Model for Unit 3

Data set : Unit3
Total Number of instances: 87
Training set: 87
Evaluation : 10-fold Cross-validation
Algorithm : DecisionStump
Time (s) : 0

(b) General Information about the model

Figure 5.4 One-level Decision Tree Model for Unit 3

ii. **Model Evaluation**: The model evaluation is done using the equations in Figure 3.2 (Equations of the Evaluation methods). Table 5.10 gives the details about the model accuracy, as shown in the table, the correlation coefficient is low (0.8558), compared to (0.9383) for Pace Regression model for this dataset. Error factors are higher than errors in Pace Regression model:

- Mean absolute error is 3.1496 compared to 2.1045 in Pace Regression,
- Root mean squared error is 3.8868 compared to 2.598 in Pace Regression,
- Relative absolute error is 48.73 % compared to 32.56% in Pace Regression,
- Root relative squared error is 51.20% compared to 34.22% in Pace Regression.

So, it is clear that the model accuracy is very low. Table 5.11 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using One-level Decision Tree, for 30 instances of the Dataset of Unit 3. The Error column shows how much the predicted values are very far from the actual ones. Also the graph at Figure 5.5 gives more clear vision about the model accuracy, the graph compares between the actual and predicted values of the amount of the generated power. The big difference between the actual and predicted values is very clear from the graph.

5.5.2 Power Prediction Models Using Controllable Parameters for Unit 4

According to the results of the initial comparison between algorithms in table 5.7; the algorithm that shows the highest correlation co-efficient, and minimum errors in Unit 4 dataset is Isotonic Regression, while Conjunctive Rule algorithm achieves the worst results. Using Unit 4 dataset, two models were designed; Isotonic Regression, and Conjunctive Rule. Subsequent parts provide the design, evaluation, and discussion about these models.

Table 5.11 Sample of comparison between Actual and Predicted values Of the Generated Power, using One-level Decision Tree , for Test Dataset of Unit 3

inst#	actual	predicted	error
1	41.885	43.96	2.075
2	40.127	43.96	3.833
3	49.172	43.96	-5.212
4	39.912	43.96	4.048
5	32.625	30.117	-2.508
6	42.081	43.96	1.879
7	34.09	30.117	-3.973
8	39.736	43.96	4.224
9	36.005	30.117	-5.888
10	42.12	43.441	1.321
11	47.902	43.441	-4.461
12	48.938	43.441	-5.497
13	30.066	30.458	0.392
14	31.785	30.458	-1.327
15	48.098	43.441	-4.657
16	44.757	43.441	-1.316
17	44.991	43.441	-1.55
18	29.949	30.458	0.509
19	40.127	44.274	4.147
20	40.303	44.274	3.971
21	40.029	44.274	4.245
22	34.95	30.345	-4.605
23	40.166	44.274	4.108
24	40.147	44.274	4.127
25	30.144	30.345	0.201
26	42.295	44.274	1.979
27	39.658	44.274	4.616
28	40.42	43.663	3.243
29	35.126	30.952	-4.174
30	45.773	43.663	-2.11

Table 5.10 One-level Decision Tree Model Accuracy for Unit 3 Data set

Correlation coefficient	0.8558
Mean absolute error	3.1496
Root mean squared error	3.8868
Relative absolute error	48.73%
Root relative squared error	51.20%

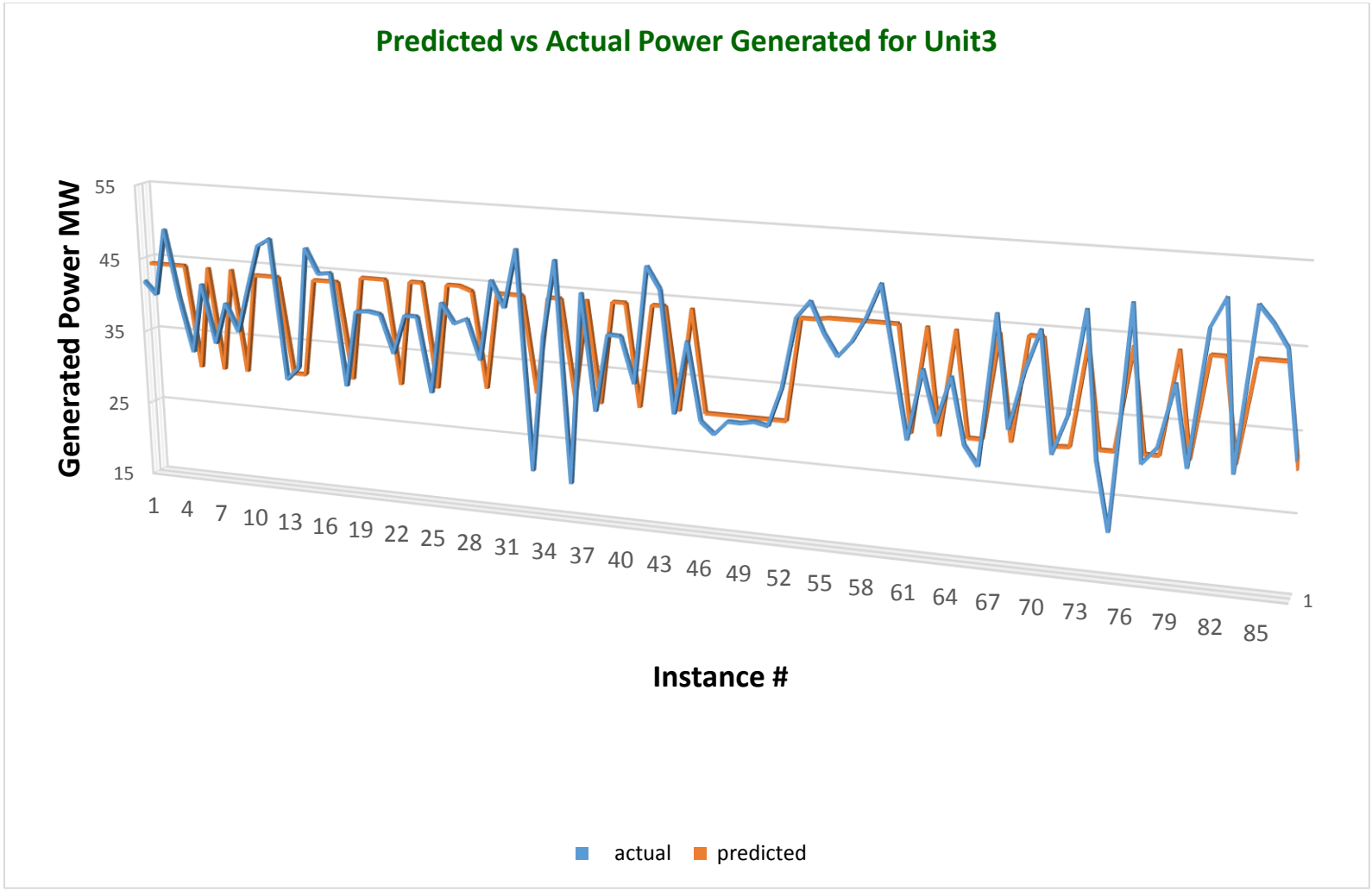


Figure 5.5 Graph for comparison between Actual and Predicted values Of the Generated Power, using One-level Decision Tree , for Test Dataset of Unit 3

5.5.2.1 Isotonic Regression Model

Isotonic regression is the technique of fitting a free-form line to a sequence of observations under the following constraints: the fitted free-form line has to be non-decreasing everywhere, and it has to lie as close to the observations as possible. Isotonic regression model picks the attribute that results in the lowest squared error (Jan, Kurt and Patrick, 2009). Isotonic regression is selected to develop the prediction model for Unit 4 because it showed the highest correlation coefficient, and minimum errors. Below is the details about the model and its evaluation.

i. **Power Prediction Model**: Isotonic Regression implements the method for learning an isotonic regression function based on the pair-adjacent violators approach. *PLSClassifier* learns a partial least squares regression model. It uses the *PLSFilter* to transform the training data into partial least-squares space and then learns a linear regression from the transformed data (Witten et al, 2011). Figure 5.6 (a) shows the Isotonic Regression model for Unit 4 dataset. As seen from the figure the model depends only on the steam flow to predict the amount of the generated power. Figure 5.6 (b) gives some basic information about the model and the used dataset. The model used 10-fold cross validation for evaluation, because the number is only 87 which is relatively low. The time required to build the model is less than one second.

   The efficiency engineers, beside the accurate prediction they need to know the behavior of all the three predictors, and how they influence the amount of the generated power. Although Isotonic Regression achieved high accuracy (correlation coefficient = 0.9943), but it can't show the influence of the other parameters, this because it depends only on Steam Flow. However, the correlation coefficient of the Support Vector Machine algorithm is also high (0.9915), and it provide the required understanding about the behavior of the other features. Figure 5.7 shows the Support Vector Machine for Regression model.

Isotonic regression Model
Based on attribute: SteamFlow

cut point:	28.36	prediction:	26.02
cut point:	29.63	prediction:	28.07
cut point:	30.46	prediction:	28.4
cut point:	31.08	prediction:	29.85
cut point:	31.21	prediction:	29.88
cut point:	31.3	prediction:	30.09
cut point:	31.82	prediction:	30.14
cut point:	32.53	prediction:	30.83
cut point:	32.81	prediction:	31.2
cut point:	33.1	prediction:	31.22
cut point:	33.57	prediction:	33.74
cut point:	37.68	prediction:	33.8
cut point:	43.99	prediction:	39.28
cut point:	45.84	prediction:	43.45
cut point:	46.78	prediction:	43.6
cut point:	47.99	prediction:	44.12
cut point:	48.71	prediction:	45
cut point:	48.81	prediction:	45.46
cut point:	50.03	prediction:	45.91
cut point:	50.76	prediction:	46.51
cut point:	51.24	prediction:	49.72
cut point:	51.68	prediction:	49.84
cut point:	54.51	prediction:	49.87
cut point:	56.13	prediction:	50.43
cut point:	57.82	prediction:	55.23
cut point:	58.87	prediction:	55.48
cut point:	59.92	prediction:	55.85
		prediction:	56.93

(a) Isotonic Regression Model for Unit 4

Data set : Unit 4
Total Number of instances: 83
Training set: 87
Evaluation : 10-fold Cross-validation
Algorithm : Isotonic

(b) General Information about the model

Figure 5.6 Isotonic Regression Model for Unit 4

SMOreg
weights (not support vectors):
+ 0.9648 * (normalized) SteamFlow
+ 0.0375 * (normalized) Pressure
+ 0.0825 * (normalized) Temperature
- 0.0409

Figure 5.7 Support Vector Machine for Regression Model for Unit 4

Table 5.13 Sample of comparison between Actual and Predicted values Of the Generated Power, using Isotonic Regression, for Test Dataset of Unit 4

Instance No.	Actual	Predicted	Error
1	30.828	31.199	0.371
2	49.758	49.842	0.084
3	55.424	55.033	-0.391
4	28.073	28.327	0.254
5	50.286	49.842	-0.444
6	51.399	49.915	-1.484
7	28.464	28.327	-0.137
8	49.719	49.842	0.123
9	29.538	29.892	0.354
10	49.895	49.797	-0.098
11	49.973	49.797	-0.176
12	49.719	49.87	0.151
13	42.237	43.683	1.446
14	50.188	49.797	-0.391
15	29.968	29.796	-0.172
16	40.088	39.072	-1.016
17	30.066	29.796	-0.27
18	45.753	46.887	1.133
19	55.482	55.229	-0.254
20	39.951	39.106	-0.845
21	49.973	50.657	0.684
22	49.328	49.865	0.537
23	44.933	42.74	-2.193
24	30.027	29.838	-0.189
25	44.835	42.74	-2.095
26	46.886	44.366	-2.52
27	50.227	49.865	-0.362
28	46.027	46.75	0.722
29	49.426	50.429	1.003
30	29.675	29.857	0.182

Table 5.12 Isotonic Regression Accuracy for Unit 4 Data set

Correlation coefficient	0.9943
Mean absolute error	0.644
Root mean squared error	0.9224
Relative absolute error	8.57%
Root relative squared error	10.61%

ii. **Model Evaluation**: Model evaluation is done using 10-fold cross validation. Each fold is done using the equations in Figure 3.2 (Equations of the Evaluation methods). Table 5.12 give the details about the model accuracy, as shown in the table, the correlation coefficient is high (0.9943) , and error factors are low (Mean absolute error is 0.644, Root mean squared error is 0.9224, Relative absolute error is 8.57%, and the Root relative squared error is 10.61%). The model accuracy is high. Table 5.13 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Isotonic Regression, for 30 instances of Unit 4, the Error column shows how much the predicted values are near to the actual ones. The graph in figure 5.8 gives more clear vision about the model accuracy, the graph compares between the actual and predicted values of the amount of the generated power.

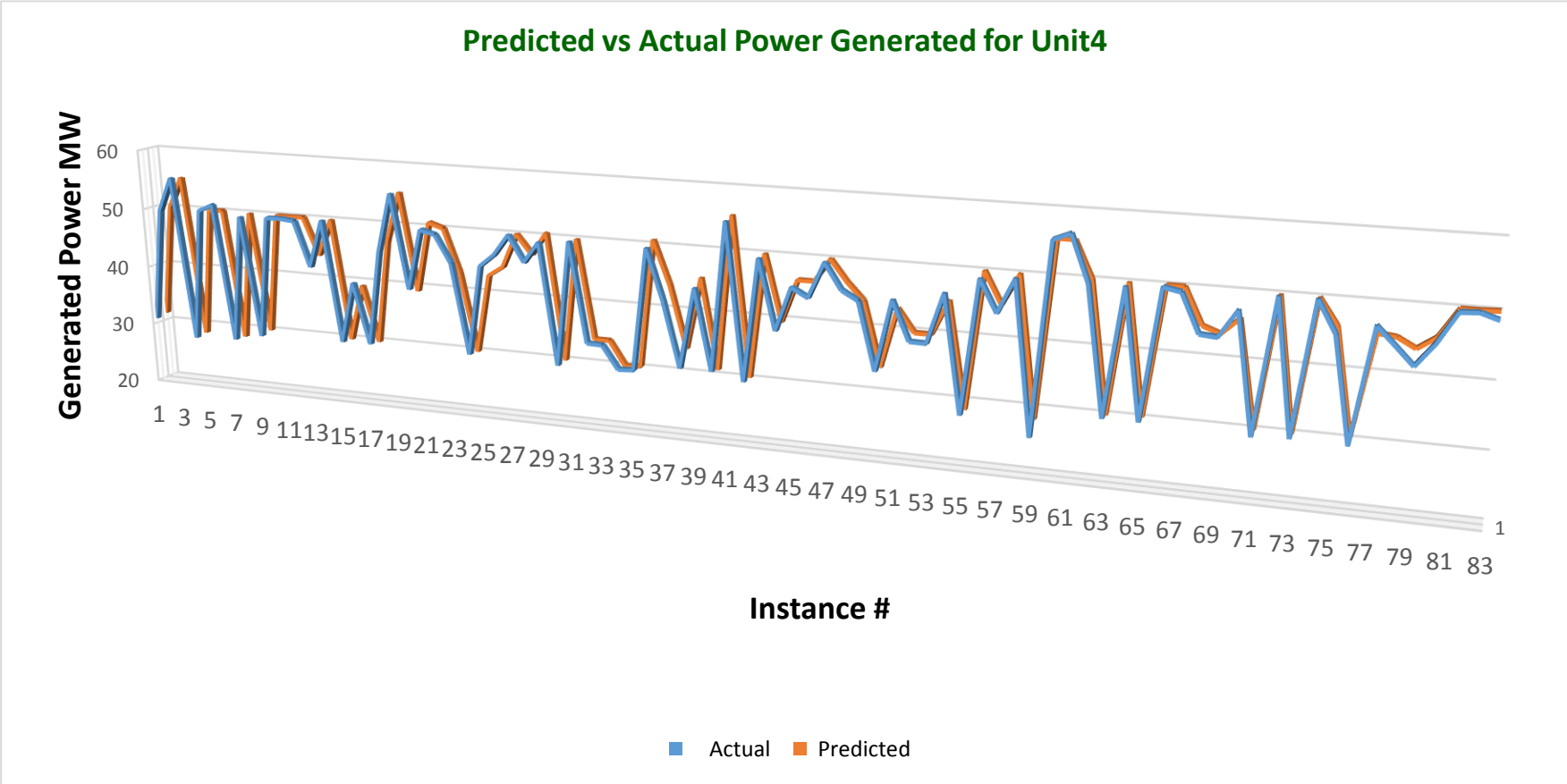


Figure 5.8 Graph for comparison between Actual and Predicted values Of the Generated Power, using Isotonic Regression, for Test Dataset of Unit 4

5.6.2.2 Conjunctive Rule Model

Another model was built using Conjunctive Rule algorithm. This algorithm implements a single conjunctive rule learner that can predict for numeric and nominal class labels. A rule consists of antecedents "AND"ed together and the consequent (class value) for the classification/regression. In this case, the consequent is the mean for a numeric value in the dataset. This learner selects an antecedent by computing the Information Gain of each antecedent and prunes the generated rule using Reduced Error Pruning (Witten, Frank and Hall, 2011). The model showed very poor accuracy, this model was built only to compare its results with the first one which is the most accurate model in Unit 4. Below is the details about the Conjunctive Rule model and it evaluation.

i. **Power Prediction Model**: Figure 5.9 (a) shows the Conjunctive Rule model for Unit 4 dataset. As seen from the figure, like Isotonic Regression model, the Conjunctive Rule algorithm used only the *Steam Flow* as a predictor. Because the model depends on a single feature, it can't assist the efficiency engineers to understand the behavior of other features, and how they influence the amount of the generated power. The model predicts the power by a very primitive equation which is shown in figure 5.9(a).

ii. **Model Evaluation**: The model evaluation is done using the equations in Figure 3.2 (Equations of the Evaluation methods). Table 5.14 gives the details about the model accuracy, as shown in the table, the correlation coefficient is low (0.8902), compared to (0.9943) for Isotonic Regression model for this dataset. Error factors are higher than errors in Isotonic Regression model:
   - Mean absolute error is 3.1135 compared to 0.644 in Isotonic Regression,
   - Root mean squared error is 3.9343 compared to 0.9224 in Isotonic Regression,
   - Relative absolute error is 41.43 % compared to 8.57% in Isotonic Regression,
   - Root relative squared error is 45.24% compared to 10.61% in Isotonic Regression.

So, it is clear that the model accuracy is very low. Table 5.15 shows a sample of a comparison between the Actual and Predicted values of the Generated Power, using Conjunctive Rule, for 30 instances of the Dataset of Unit 4. The Error column shows how much the predicted values are very far from the actual ones. Also the graph at Figure 5.10 gives more clear vision about the model accuracy, the graph compares

between the actual and predicted values of the amount of the generated power. The big difference between the actual and predicted values is very clear from the graph.

ConjunctiveRule Model
Single conjunctive rule learner:

(SteamFlow > 43.9885) =>
Power = 48.327471

(a) Conjunctive Rule Model for Unit 4

Data set : Unit4
Total Number of instances: 83
Training set: 87
Evaluation : 10-fold Cross-validation
Algorithm: ConjunctiveRule
Time (s) : 0

(b) General Information about the model

Figure 5.9 Conjunctive Rule Model for Unit 4

Table 5.15 Sample of comparison between Actual and Predicted values Of the Generated Power, using Conjunctive Rule , for Test Dataset of Unit

Instance No.	Actual	Predicted	Error
1	30.828	32.112	1.284
2	49.758	48.441	-1.317
3	55.424	48.441	-6.983
4	28.073	32.112	4.039
5	50.286	48.441	-1.845
6	51.399	48.441	-2.958
7	28.464	32.112	3.648
8	49.719	48.441	-1.278
9	29.538	32.112	2.574
10	49.895	48.736	-1.159
11	49.973	48.736	-1.237
12	49.719	48.736	-0.983
13	42.237	48.736	6.499
14	50.188	48.736	-1.452
15	29.968	32.425	2.457
16	40.088	32.425	-7.663
17	30.066	32.425	2.359
18	45.753	48.736	2.983
19	55.482	48.884	-6.598
20	39.951	30.758	-9.193
21	49.973	48.884	-1.089
22	49.328	48.884	-0.444
23	44.933	48.884	3.951
24	30.027	30.758	0.731
25	44.835	48.884	4.049
26	46.886	48.884	1.998
27	50.227	48.884	-1.343
28	46.027	48.39	2.363
29	49.426	48.39	-1.036
30	29.675	32.468	2.793

Table 5.14 Conjunctive Rule Accuracy for Unit 4 Data set

Correlation coefficient	0.8902
Mean absolute error	3.1135
Root mean squared error	3.9343
Relative absolute error	41.43%
Root relative squared error	45.24%

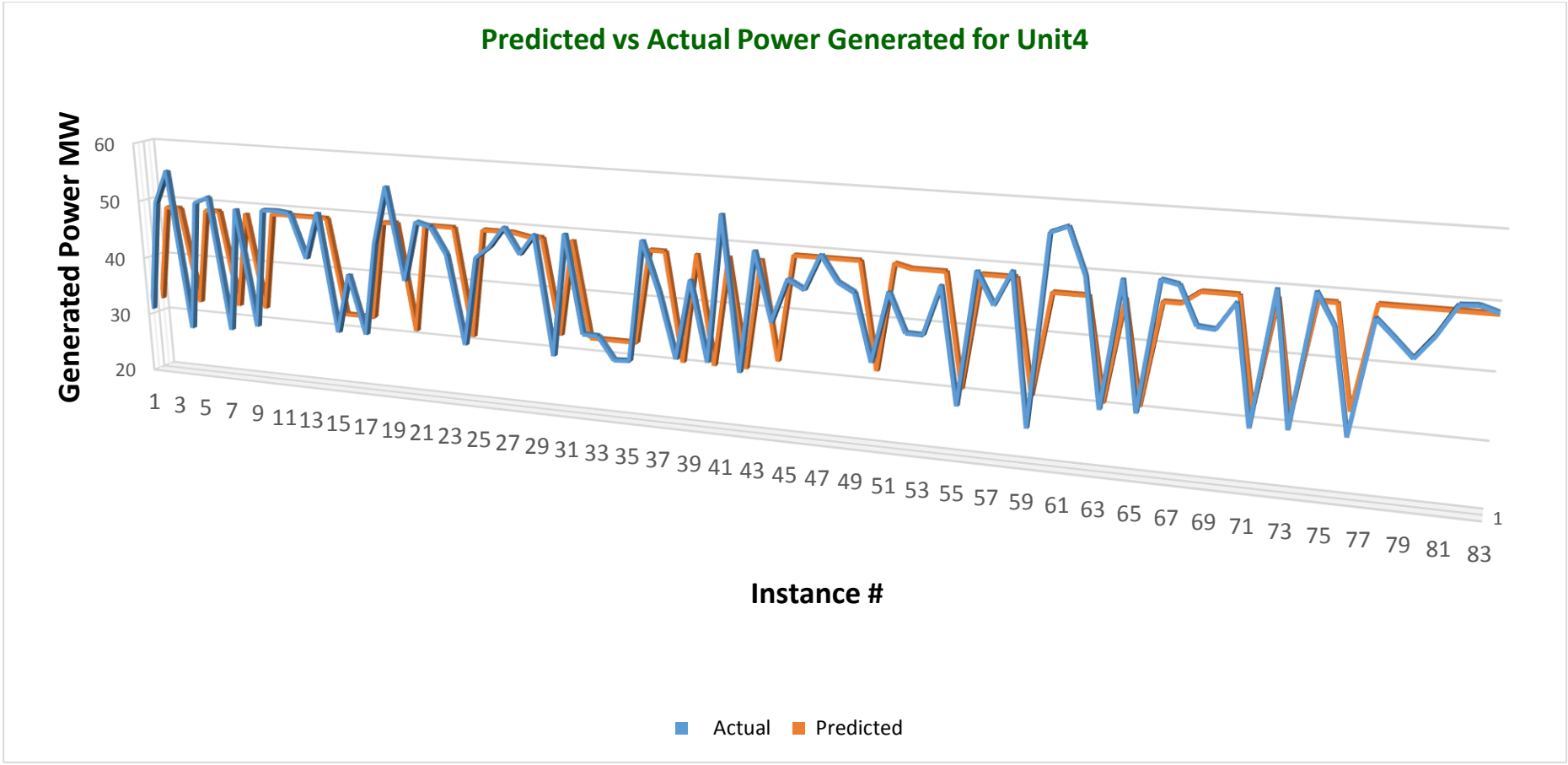


Figure 5.10 Graph for comparison between Actual and Predicted values Of the Generated Power, using Conjunctive Rule , for Test Dataset of Unit 4

5.7 Results Discussion and Models Comparison

In unit 3, a big difference in standard deviation of the Steam Pressure at Turbine Inlet is observed (28.875), this is due to the high difference between the observed values of the Steam Pressure (128 and 0.4 bar). If these values are correct, they will cause damage in tubes and power plant components. Because of this it is very clear that Unit 3 data is noisy, and we can not depend on it to get accurate prediction. That is very clear from the results obtained from Unit 3. The Pressure at Turbine Inlet is one of the most important features in calculating the amount of power, consequently, the correlation coefficient of all prediction models in Unit 3 becomes very low, and the Mean Absolute Error is high. After initial comparison between seventeen algorithms the Pace Regression was selected to build the prediction models for Unit 3. The value of the generated power which is predicted by Pace Regression model, is much accurate than those which are calculated using either the thermodynamic laws, or the Steam Consumption Graph. Table 5.16 shows a sample of the four value of the amount of the generated power from Unit 3: Actual generated power, Calculated power using Steam Consumption graph, the calculated power using Thermodynamic Laws, and the Predicted power. It is very clear from the graph of figure 5.11 that the Predicted Amount using the Pace Regression is much accurate than those calculated using the thermodynamic laws, or the Steam Consumption Graph.

Unit 4 dataset is much better than unit 3, that is obvious from the table 5.4 which shows basic statistical analysis about unit 4 dataset. The mean values of pressure and temperature are very near to the optimum values assigned by the power plant's manufacturer. So, the amount of the actual generated power is very near to the amount which is calculated by the Steam Consumption Graph. Table 5.17 shows a sample of the four value of the amount of the generated power from Unit 4: the Actual generated power, the Calculated power using Steam Consumption graph, the calculated power using Thermodynamic Laws, and the Predicted power. It is very clear from the graph of figure 5.11 that the Predicted Amount using the Isotonic Regression is much accurate than those calculated using the thermodynamic laws, or the Steam Consumption Graph.

From the above discussion of model evaluation results, it is clear that the values predicted by the new prediction models (Pace Regression for Unit 3, and Isotonic Regression in Unit 4) are more accurate than those calculated by Thermodynamic Laws, or Steam Consumption Graph. Moreover, the new prediction models depends only on the controllable parameters. Hence, they could be used as efficient tool to predict the amount of the generated power from a thermal power plant accurately. An additional contribution of this research is that: by data exploration and analysis and prediction models; we are able to highlight some of Unit 3 problems, like sensors, problem of Steam Pressure at Turbine Inlet. By all these the developed prediction models succeeded to answer all the research questions.

5.8 Summary

The calculation of the amount of the generated power from a thermal power plant is done using five features which are (*1. Main steam flow, 3. Pressure at Turbine Inlet, 4. Temperature at Turbine Inlet, 61. Pressure at Turbine Outlet, 62. Temperature at Turbine Outlet*), only the first three features are controllable. Any prediction model to be effective and usable, it should depend only on these three features. Hence, the objective of this chapter is "*To design a prediction technique that can accurately predicts the amount of the generated power from a thermal power plant, using only the controllable parameters*". To achieve this objective, 2 datasets were used with 17 prediction algorithms. The models that showed higher correlation coefficient and minimum errors were selected to design the prediction models. The predicted values were compared with the actual amount of the generated power, and the calculated power (using both Thermodynamic Laws, and Steam Consumption Graph). The predicted values are more accurate than the values calculated using the traditional methods.

Pace Regression attained the highest Correlation Coefficient (0.9383) and lowest MAE (2.1045) with Unit 3 dataset, and Isotonic Regression model attained the highest Correlation Coefficient (0.9943) and lowest MAE (0.644) with unit 4. Although Isotonic Regression attained the highest results, but it depends only on one feature. Hence, Support Vector Machine model was also developed because it achieved high correlation coefficient (0.9915), and provided the required explanation about the behavior of the other features.

Table 5.16 Sample of Comparison between Actual, Predicted, and Calculated (Using Steam Consumption graph and Thermodynamic Laws) in Unit 3

Actual	Calculated by Manufacturer Graph	Calculated by Thermodynamic Laws	Predicted
19.849	22.569	19.306	19.163
21.021	24.092	21.299	21.754
21.099	24.142	21.384	23.051
27.604	31.223	26.741	24.612
28.327	32.522	26.503	29.759
29.773	32.102	29.137	33.435
29.812	36.761	30.307	31.819
29.851	33.332	27.997	29.546
29.890	34.277	28.213	29.819
29.929	31.919	31.063	30.768
29.929	32.697	26.227	27.500
29.949	32.029	30.585	25.798
30.027	33.022	26.715	29.814
30.046	31.870	28.332	30.989
30.066	31.991	30.857	26.192
30.085	32.068	28.031	30.178
30.105	32.435	27.517	28.267

Table 5.17 Sample of Comparison between Actual, Predicted, and Calculated (Using Steam Consumption graph and Thermodynamic Laws) in Unit 4

Actual	Calculated by Manufacturer Graph	Calculated by Thermodynamic Laws	Predicted
26.022	25.266	22.473	28.073
28.073	27.456	24.322	28.327
28.327	28.215	25.018	28.464
28.464	27.794	24.520	28.327
29.538	29.009	25.682	29.892
29.597	29.055	25.642	29.883
29.675	28.940	25.513	29.857
29.734	29.171	25.713	30.027
29.910	28.994	25.722	29.883
29.968	28.791	25.563	29.796
29.988	28.701	25.381	29.857
30.027	29.101	25.406	29.838
30.066	28.888	25.446	29.796
30.066	28.997	25.709	29.820
30.085	29.249	25.627	30.085
30.085	29.270	25.641	30.085
30.144	29.336	25.761	30.085

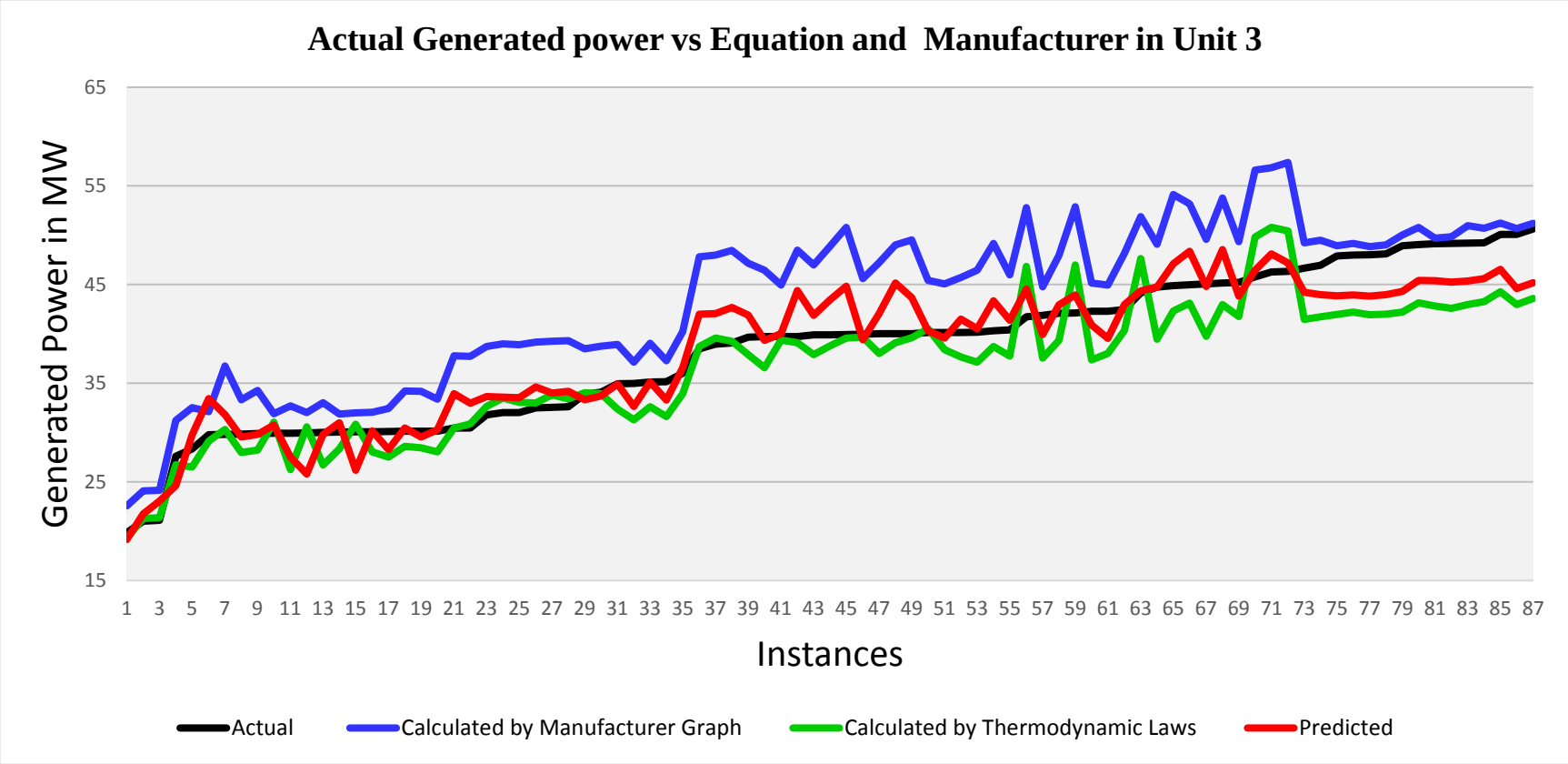


Figure 5.11 Comparison between Actual, Predicted, and Calculated (Using Steam Consumption graph and Thermodynamic Laws) in Unit 3

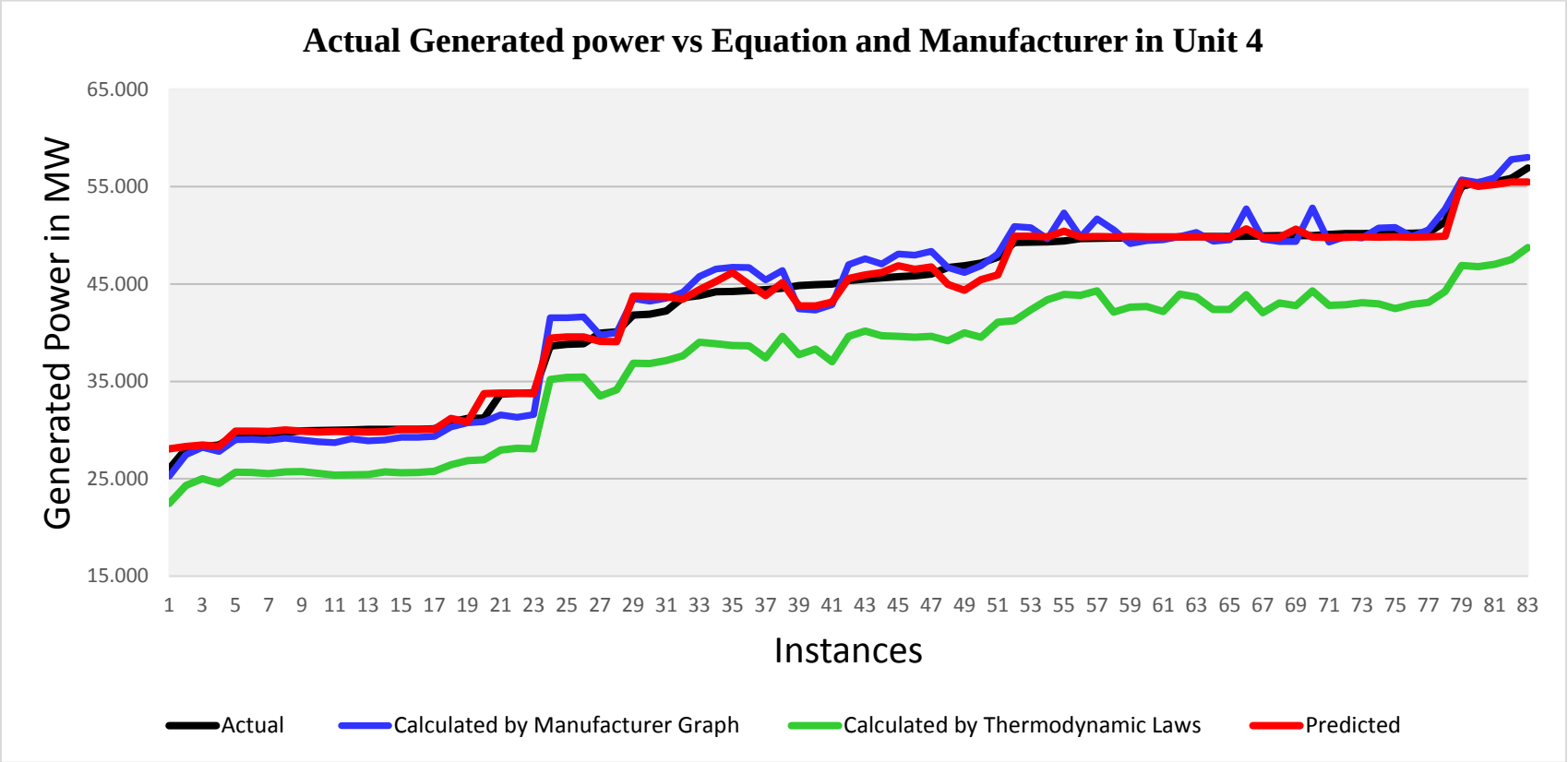


Figure 5.12 Comparison between Actual, Predicted, and Calculated (Using Steam Consumption graph and Thermodynamic Laws) in Unit 4

CHAPTER SIX

CONCLUSION AND FUTURE WORK

6.1 Introduction

The aim of this chapter is to summarize the research objectives and techniques that have been carried out in this thesis. The research concerns the development of techniques that can accurately predict the Amount of the Generated Power from a Thermal Power Plant. CRISP-DM model (Shearer, 2000) is used to is organize the work into six phases, namely: Business Understanding; at this phase the concepts of power generation is reviewed and the research objectives are clearly defined. Then Data Preparation Data Understanding and; where data is prepared to the machine learning tools, and basic statistical analysis was done to give better understanding about the datasets. The forth phase is the modeling; which is composed of two main parts; Feature Selection and Prediction Model development. After that comes the Evaluation phase, where each model is evaluated using the correlation coefficient and the error rates to reflect the model accuracy.

This chapter is organized as follows: Section 7.2 summarizes the proposed methods presented in this research. Section 7.3 presents the research contributions in three main points. The future work are discussed in Section 7.4. Finally the whole chapter is concluded with a summary in Section 7.5.

6.2 The Proposed Method

This research has two main objectives, to achieve these objectives, data mining techniques were used. The first one is "to design a feature selection technique, that can determine the best set of features to predict the amount of the generated power from a thermal power plant". The second is "to design a prediction technique that can accurately predicts the amount of generated power, using only the controllable parameters". To achieve these objectives datasets were collected from two identical Units (Unit 3 and 4) in Khartoum North Power Plant. Then 17 algorithms were applied for these datasets, the ones that attained the highest correlation coefficient and minimum errors were selected to build the suitable models for each objective.

For the first objective, the Wrapper method of feature selection was used to select the best set of features, that can accurately predict the amount of the generated power. Linear Regression model attained the highest Correlation Coefficient (0.9998) in Unit 4, 52 attribute were selected by the model to predict the power, and the model succeeded in determining the other features with their influence (weight) to the amount of the generated power. In Unit 3 Pace Regression also achieved high Correlation Coefficient (0.9997) using only 28 features, but the Main steam header pressure, which is one of the features needed to calculate the power, was not selected. Some features like T/A axial displacement, which was selected by the two models, draw the attention of the domain experts because they appeared with high factors.

For the second objective, Pace Regression attained the highest Correlation Coefficient (0.9383) with Unit 3 dataset, and Isotonic Regression model achieved (0.9943) with unit 4. Because Isotonic Regression depends only on one feature, Support Vector Machine model was also developed to provided the required explanation about the behavior of the other features. Beside the correlation coefficient and the error rates, graphs were used to compare the predicted vs the actual values, to give better vision about models' accuracy. Graphs were also used to compare the predicted power vs the actual and calculated values (using thermodynamic laws, and manufacturers' guides). These comparisons shows the superiority of data mining techniques to the thermodynamic laws, and manufacturers' guides.

6.3 Contribution of the Study

As mentioned earlier in the previous section. The main goal of this thesis is to use data mining to identify the parameters that influence the amount of generated power and to accurately predict this amount. The contribution of this thesis is to highlight the superiority of data mining to traditional methods in calculating the amount of generated power from a thermal power plant. This contribution could be summarized as follows:

i. Identification of important features that influence the amount of generated power in a thermal power plant.

The starting point is to eliminate the unimportant attributes in the dataset, to get better results of prediction. This is done in two steps: the first one is an expert guide to exclude attributes that are not related to the problem, this step reduced the dataset from 83 to 62 attributes. Then wrapper method was applied as a feature selection techniques to this new dataset which is composed of 62 attributes. The result of this step differ from dataset to another, 28 out of 61 attributes were selected for unit 3, 52 for unit 4, and 18 for unit 3&4 dataset. Moreover all attribute selection results were collected and summarized in one sheet, the attributes that were selected by many algorithms are given higher ranks, which gives an indication about the relevancy of this attribute to class. The result of this part is the attribute selection result summary sheet which had been submitted to the domain expert, who found some useful observations. For example (3. Main steam header pressure) which is one of the top five features had never been selected by Unit 3 dataset, and (41. T/A axial displacement B (mm)) which hade never been considered as an important factor was selected by 7 models. So, this summary sheet could be used as a basic analysis tool to the power plant status.

ii. Better understanding of the effects of selected parameters in thermal power plant using feature selection and prediction algorithms.

This point is directly related to the first one because to exclude the unimportant parameters, we used feature selection. Then the selected set of parameters were used to predict the amount of the generated power. Although this is not a practical way to predict the power because of two reasons; the first is that the models use both controllable and non-controllable parameters as predictors, the second reason is the high number of features, which will lead to overfitting. However the factor related to each predictor in the prediction model (like the factors of regression equation) gives indication of the parameter's effect in assigning the amount of the predicted power. This is very clear from the Linear Regression model, which achieved the highest accuracy in Unit 4 (0.9998). Also the results of this part had been submitted to the domain expert, they also get some observations like those related to the T/A axial displacement B (mm).

iii. Develop an accurate method to predict the amount of the generated power based on real data collected from the power plant, by using data mining techniques.

The goal of this part is to use only the set of controllable parameters to predict the amount of generated power. All units showed different results for different algorithms, so the conclusion is that; according to the current situation of each unit in the power plant, the prediction results may differ. The predicted values were compared with the actual amount of the generated power, and the calculated power (using both Thermodynamic Laws, and Steam Consumption Graph). The predicted values are more accurate than the calculated values. This accuracy proves the contribution of this research. For unit 4 Isotonic Regression was used, which attained high (0.9943) correlation coefficient. While in unit 3 Pace Regression was used, which also attained high (0.9383) correlation coefficient.

6.4 Future Work

This research achieved all the objectives of the study, the current work has focused only on the amount of the generated power from a thermal power plant. However, a number of research opportunities still exist and further researches can be conducted in the area of thermal power plants. Specifically, further studies can be conducted in the following areas:

- Identify the parameters that influence the steam pressure at turbine outlet, because this is the main non controllable factor that affects the amount of the generated power.
- Prediction of power plant component failure, this task needs more collaboration from power plant engineers to identify the problems and failure types of these components.
- Build a data warehouse for power plant data, to increase the analysis efficiency, by designing the data warehouse and adding more space to database servers.
- Use Big Data technology and Hadoop to gain more efficient analysis capabilities, and leverage the replication techniques of Hadoop.
- Investigate the use of real time data mining, to get instant results and direction in addition to the analyzing the historical data.
- Investigate new data mining algorithm like deep learning to be used in power plant efficiency prediction, and failure detection.
- Develop a complete application that could be integrated with existing databases and the proposed data warehouse. The target of the new application is to ease data management and prediction of power plant efficiency and failure of main components.

The main components of all thermal power plants are identical. Hence, the model which is designed in this research could be generalized and implemented as is, at any thermal power plant for the same objectives (to determine the features that influence the amount of power, and to predict the power using only the controllable parameters). This could be done by the following steps:

- Data collection and preparation for the 63 features.
- Run Wrapper feature Selection method using Linear Regression, to select the best set of features and build prediction models using the selected set of features.
- Use Isotonic Regression with the controllable parameters to accurately predict the amount of the generated power.

Although all thermal power plants follow Rankine Cycle (Learn Engineering, 2013), but there are a lot of differences between them, which may lead to differences in instances and features of the datasets. These differences like: the size, status, manufacturers, fuel type, and even the location and weather conditions. Because of this it is better to precede the second step by an initial comparison between algorithms' performance for each datasets, to use the most suitable algorithm for that dataset, rather than depending on fixed algorithms.

The preferred way to do this is by developing a flexible software package that can take as in put the features, and generate all prediction reports and comparison graphs as its output.

6.5 Summary

The goal of this research is to use data mining techniques to solve two problems: the first one is to identify the parameters that influence the amount of generated power from a thermal power plant. The second is to predict the amount of generated power accurately, using the existing data of the controllable parameters.

This chapter presents the summaries of the techniques used in this research to achieve these goals. This chapter proves that depending on thermodynamic laws or manufacturers expectations will not give accurate results for the amount of generated power. Moreover each unit shows different results for different algorithms, so according to the situation of the power plant and specifically for each unit the prediction results may differ. The proposed solution is composed of two parts: the first one identifies the parameters that influence the amount of generated power using Wrapper Feature Selection method, to select the best set of features, and predict the power using this selected set of features. The selected algorithms are Paces Regression for unit 3, Linear Regression for unit 4, and Neural Network for unit 3&4 datasets. The second part is the power prediction using controllable parameters, the selected algorithms for this part are are Paces Regression for Unit 3, Isotonic Regression for Unit 4.

The results achieved by power prediction models of this research outperform the traditional approaches of power calculation. This confirms the effectiveness of the proposed methods. Furthermore, this research discusses the plan for future work to improve the current work and how it will be applied for all thermal power plants.

REFERENCES

Abolhosseini, A., Heshmati, A. and Altmann, J. (2014). A Review of Renewable Energy Supply and Energy Efficiency Technologies. [pdf] Bonn: Institute for the Study of Labor. Available at ftp.iza.org/dp8145.pdf [Accessed 12 Aug 2017]

Aha, D. and Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*. 6, pp. 37-66.

Ali, S. and Shahzad, W. (2012). A Feature Subset Selection Method based on Conditional Mutual Information and Ant Colony Optimization. *International Journal of Computer Applications*, 60(11), pp. 5–10.

Amooee, G., M-Bidgoli, B. and B-Dehnavi, M. (2011). A Comparison Between Data Mining Prediction Algorithms for Fault Detection (Case study: Ahanpishegan co.). *IJCSI International Journal of Computer Science*, 8(6), p 3.

Bach and Francis R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In: *Proceedings of the 25th international conference on Machine learning*, pp. 33–40.

Bermingham, L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A., Wilson, F., Agakov, F., Navarro, P. and Haley, C. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man.

Booth, R. and Roland, W. (1998). Neural network-based combustion optimization reduces NOx emissions while improving performance. *In: Proc. IEEE Industry Applications Dynamic Modeling Control Applications Industry Workshop*, pp. 1–6.

Bradley P. and Mangasarian, O. (1998). Feature selection via concave minimization and support vector machines," In: *Proc.15th International Conference on Machine Learning* , Madison, Morgan Kaufmann, pp. 82–90.

Büche, D., Stoll, P., Dornberger, R. and Koumoutsakos, P. (2002). Multiobjective evolutionary algorithm for the optimization of noisy combustion processes, *IEEE Trans. Syst., Man, Cybern. C*, 32(4), pp. 460–473.

Burns, A., Kusiak, A. and Letsche, T. (2004). Mining transformed data sets. In: Khosla, R., Howlett, R. and Jain L., ed., Knowledge-Based Intelligent Information and Engineering Systems, . Heidelberg: Springer, pp. 148–154.

Cadenas, J., Garrido, M. and Martínez, R. (2013). Feature subset selection Filter–Wrapper based on low quality data. *Expert Systems with Applications*, 40, pp. 6241–6252.

Cass, R. and Radl, B. (1997). Adaptive process optimization using functional link networks and evolutionary optimization, *Control Eng. Practice*, 4(11), pp. 1579–1584.

Cawley, G., Talbot, N. and Girolami, M. (2007). Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. In: *B. Schölkopf, J. C. Platt, and T. Hoffmann (eds.), Advances in Neural Information Processing Systems*, MIT Press, pp. 209–216.

Chen, K., Chen, L., Chen, M. and Lee, C. (2011). Using SVM based method for equipment fault detection in a thermal power plant, *Elsevier Computers in Industry*, 62, pp. 42–50.

Chiu, S. and Chen, C. and Lin, T. (2008). Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer. *Artif Intell*, 44, pp. 221–231

Chong, A., Wilcox, S. and Ward, J. (2002). Neural network models of the combustion derivatives emanating from a chain grate stoker fired boiler plant. In: *Proc. Inst. Elect. Eng. Seminar Advanced Sensors Instrumentation Systems Combustion Processes*, pp. 1–4.

Chu, J., Shieh, S., Jang, S., Chien, C., Wan, H. and Ko, H. (2003). Constrained optimization of combustion in a simulated coal-fired boiler using artificial neural network model and information analysis. *Fuel*, 82(6), pp. 693–703.

Cleary, J. and Trigg, L. (1995). An Instance-based Learner Using an Entropic Distance Measure. In: *12th International Conference on Machine Learning*, pp. 108-114.

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. In: *Proc. 18th International Conference on Machine Learning*, San Francisco, Morgan Kaufmann, pp. 74–81.

Dash, M. and Ong, Y. (2011). RELIEF-C: Efficient Feature Selection for Clustering over Noisy Data. In: *Proc. 23rd IEEE International Conference on Tools with Artificial Intelligence*, Roca Raton, Florida, pp. 869–872.

Duda, R., Hart, P. and Stork, D. (2012), *Pattern classification*, 2 nd ed. Wiley-interscience.

Ethem Alpaydın (2010). *Introduction to Machine Learning*. 2 nd ed. London: The MIT Press.

Fazullula, M., Praveen, M. and Reddy, S. (2014). Visual Data Mining: A case study in Thermal Power Plant. *IJISET - International Journal of Innovative Science, Engineering & Technology*, 1(6), pp. 110 – 115.

Figueiredo, V., Rodrigues, F. and Gouveia, J. (2005). An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques. *IEEE Transactions on Power Systems*, 20, pp. 596-602.

Flokal, (2017). Turbine Mass Flowmeters [online] Available at: http://flokal.com/products/turbine-mass-flowmeters [Accessed 13 Oct 2017].

Frank, E., Hall, M. and Pfahringer, B. (2003) Locally Weighted Naive Bayes. In: *19th Conference in Uncertainty in Artificial Intelligence*, pp. 249-256.

Guyon, I. and Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, pp. 1157-1182

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2 nd ed. New York: Springer.

Holmes, G., Hall, M. and Frak, E. (1999). Generating Rule Sets from Model Trees. In: *Twelfth Australian Joint Conference on Artificial Intelligence*, pp. 1-12.

Hoque, N., Bhattacharyya, D. and Kalita, J. (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications*, 41(14), pp. 6371–6385.

Hou, Z., Lian, Z., Yao, Y. and Yuan, X. (2006). Data mining based sensor fault diagnosis and validation for building air conditioning system, *Energy Conversion and Management*, 47, pp. 2479–2490.

Huang, X., Qi, H. and Liu, X. (2006). Implementation of fault detection and diagnosis system for control systems in thermal power plants. In: *Proceedings of the 6th World Congress on Intelligent Control and Automation*, Dalian, pp. 21–23.

Ilamathi, P., Selladurai, V. and Balamurugan, K. (2012). Predictive modelling and optimization of Nox emission from power plant, *IAES International Journal of Artificial Intelligence (IJ-AI),1(1)*, pp. 2252-8938.

Iung, B. and Marquez, A. (2006). Editorial: special issue on e-maintenance, *Computers in Industry*, 57, pp. 473–475.

Iowa Energy Center, (2017).Data-driven Performance Optimization of Wind Farms. [online] Available at: http://www.iowaenergycenter.org/data-driven-performance-optimization-of-wind-farms [13 Sep 2016].

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer

Jan, L., Kurt, H. and Patrick, M. (2009). Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. *Journal of Statistical Software*, 32 (5), pp. 1–24.

Jović, A., Brkić, K. and Bogunović, N. (2015). A review of feature selection methods with applications, *DeMSI,UIP*, (11), pp. 15-44

Kapooria, R. and Kumar, S. and Kasana, S. (2008). An analysis of a thermal power plant working on a Rankine cycle: *a Theoretical Investigation. Journal of Energy in Southern Africa*, (19), pp. 77-83.

Kenney, F. and Keeping, S. (1962). *Linear Regression and Correlation*."" Ch. 15 . In Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 252–285"

Kim, Y., Street, W. and Menczer, F. (2002). Evolutionary model selection in unsupervised learning, *Intelligent Data Analysis*, 6(6), pp. 531–556.

Kohavi, R. (1995). The Power of Decision Tables. In: *8th European Conference on Machine Learning*, 174-189.

Korn, F., Pagel, B. and C. Faloutsos, (2001). On the dimensionality curse and the self-similarity blessing. IEEE Trans. *Knowl.Data Eng*, 13(1), pp. 96–111.

Küçüksille, E., Selbas, R. and Sencan, A. (2011). Prediction of thermodynamic properties of refrigerants using data mining. *Elsevier Energy Conversion and Management* , 52, pp. 836–848.

Kusiak, A. and Li, W. The Prediction and Diagnosis of Wind Turbine Faults, *Renewable Energy*, (36)1, pp. 16-23.

Kusiak, A., Zhang, Z. and Li, M. (2011). Optimization of Wind Turbine Performance With Data-Driven Models. *IEEE Transactions On Sustainable Energy*, 1, pp. 66-76.

Kusiak, A., Burns, A. and Milster, F. (2005). Optimizing combustion efficiency of a circulating fluidized boiler: A data mining approach, *International Journal of Knowledge Based Intelligent Engineering Systems*, 9, pp. 263-274.

Kusiak, A., Li, M. and Tang, F. (2010). Modeling and optimization of HVAC energy consumption, *Applied Energy*, 87, pp. 3092–3102.

Kusiak, A., Zhang, Z. and Li, M. (2011). Optimization of Wind Turbine Performance With Data-Driven Models. *IEEE Transactions On Sustainable Energy*, 1, pp. 66-76.

Kusiak, A., Zheng, H. and Song, Z. (2009). On-line Monitoring of Power Curves, *Renewable Energy*, 34(6), pp. 1487-1493.

Kusiak, A., Zheng, H. and Song, Z. (2009). Short-Term Prediction of Wind Farm Power: A Data-Mining Approach, *IEEE Transactions on Energy Conversion*, 24(1), pp. 125-136.

Learn Engineering, (2013). *How does a Thermal Power Plant Work ?*. [online] Available at: www.learnengineering.org/2013/01/thermal-power-plant-working.html [Accessed 4 Sep 2016]

learnengineering.org, (2013). How does a Thermal Power Plant Work ?. [online] Available at www.learnengineering.org/2013/01/thermal-power-plant-working.html [Accessed 4 Sep. 2016].

Li, Y., Dong, M. and Hua, J. (2008). Localized feature selection for clustering. *Pattern Recognition Letters*, 29(1), pp. 10–18.

Li, Y., Wang, Z. and Yuan, J. (2006). On-line fault detection using SVM-based dynamic MPLS for batch processes, *Chinese Journal of Chemical Engineering*, 14 (6), pp. 754–758.

Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*, London: Kluwer Academic Publishers.

Liu, H. and Setiono, R. (1996). A Probabilistic Approach to Feature Selection-A Filter Solution. In: *Proc. 13th International Conference on Machine Learning*. Bary, Morgan Kaufmann, pp. 319–327.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl.Data Eng*., 17(4), pp. 491–502.

Lu, X., Dong, Z. and Li, X. (2005). Electricity market price spike forecast with data mining techniques, *Electric Power Systems Research*, 73 , pp. 19–29.

Mackay, D. (1998). Introduction to Gaussian Processes

Mahadevan, S. and Shah, S. (2009). Fault detection and diagnosis in process data using one class support vector machines, *Journal of Process Control* ,19 (10), p. 1627.

Maldonado, S., Weber, R. and Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines, *Information Sciences*, 286, pp. 228–246.

Martínez, I. and Roldán, É. and Dinis, L. and Petrov, D. and Parrondo, J. and Rica, A. (2016). Brownian Carnot engine. *Nature Physics*, pp. 67-70.

Miyayama, T., Tanaka, S., Miyatake, T., Umeki, T., Miyamoto, Y., Nishino, K. and Harada, H. (1991). A combustion control support expert system for a coal-fired boiler. In: *Proc. IEEE Industrial Electronics, Control, Instrumentation*, Kobe, pp. 1513–1516.

Modha, D. and Spangler, W. (2003). Feature weighting in k-means clustering. *Mach. Learn.*, 52(3), pp. 217–237.

Morais, J. Pires, Y. Cardoso, C. and Klautau, A. (2009). An Overview of Data Mining Techniques Applied to Power Systems. *Data Mining and Knowledge Discovery in Real Life Applications*, ISBN 978-3-902613-53-0, pp. 438.

Morais, J. Pires, Y. Cardoso, C. and Klautau, A. (2009). An Overview of Data Mining Techniques Applied to Power Systems. *Data Mining and Knowledge Discovery in Real Life Applications*, pp. 438.

Ogilvie, T., Swidenbank, E. and Hogg, B. (1998). Use of data mining techniques in the performance monitoring and optimization of a thermal power plant. In: *Proc. Inst. Elect. Eng. Colloq. Knowledge Discovery Data Mining*, pp. 1–4.

Oh, I., Lee, J. and Moon, B. (2004). Hybrid genetic algorithms for feature selection. *IEEE Trans. Pattern Anal. Mach. Intell*., 26(11), pp. 1424–1437.

Patel, M. and Vaghela, D. (2016). A Review on Feature Selection Methods for Data Mining. *IJARIIE*, 2(3), pp. 2395-4396

Penn State Eberly College of Science. Online courses. Science, (2017). *What is Simple Linear Regression?*. [online] Available at: https://onlinecourses.science.psu.edu/stat501/node/251 [Accessed 1 Oct 2017]

Prasad, G., Swidenbank, E. and Hogg, W. (1999). A novel performance monitoring strategy for economical thermal power plant, *IEEE Transactions on Energy Conversion*, 14 (3), pp. 802–809.

Quinlan, R. (1992). Learning with Continuous Classes. In: *5th Australian Joint Conference on Artificial Intelligence*, Singapore, pp. 343-348.

Quinlan, R. (1993). C4.5: Programs for Machine Learning. Machine Learning, Morgan Kaufmann Publishers, 16(3), pp. 235-240.

Refrigeration basics, (2017).The Refrigeration basics [online] Available at: http://www.refrigerationbasics.com/RBIII/definitions2.htm [Accessed 5 Aug 2017]

Rousseeuw, P., and Leroy, A. (1987). Robust regression and outlier detection.

Sandri, M. and Zuccolotto, P. (2006). Variable Selection Using Random Forests. In: S. Zani, A. Cerioli, M. Riani, and M. Vichi (eds.), Data Analysis, Classification and the Forward Search, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, pp. 263–270.

Sandri, M. and Zuccolotto, P. (2006). Variable Selection Using Random Forests. In: S. Zani, A. Cerioli, M. Riani, and M. Vichi (eds.), Data Analysis, Classification and the Forward Search, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, pp. 263–270.

Sayad, S. (2017). An Introduction to Data Mining. [online] Available at: http://www.saedsayad.com/data_mining_map.htm [Accessed 3 Jul 2017].

Şencan, A. (2007). Modeling of thermodynamic properties of refrigerant/absorbent couples using Data Mining Process, *Energy Conversion and Management*, 48, pp. 470– 480.

Sensorland, (2017).The Information Venter for Sensors and Data Systems[online] Available at:http://www.sensorland.com/HowPage004.html[Accessed 13 Oct 2017]

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data warehousing, (16), pp. 419 – 438

Shevade, S., Keerthi, S., Bhattacharyya, C. and Murthy K. (1999). Improvements to the SMO Algorithm for SVM Regression. In: *IEEE Transactions on Neural Networks*.

Shu, Y. (2007). Inference of power plant quake-proof information based on interactive data mining approach, *Advanced Engineering Informatics*, 21 (3), pp. 257–267.

Sikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of Relief and ReliefF. *Mach. Learn*., 53, pp. 23–69.

Song, Z. and Kusiak, A. (2007). Constraint-Based Control of Boiler Efficiency: A Data-Mining Approach, *IEEE Transactions On Industrial Informatics*, 3(1).

Tang, J., Alelyani, S. and Liu, H. (2013). Feature Selection for Clustering: A Review. In: C. Aggarwal and C. Reddy, ed., Data Clustering: Algorithms and Applications, Boca Raton: CRC Press, pp.29-55.

Tang, J., Alelyani, S. and Liu, H. (2014). Feature Selection for Classification: A Review. In: C. Aggarwal, ed., Data Classification: Algorithms and Applications. Boca Raton: CRC Press, pp.38-58.

TLV Global, (2017).The TLV Steam Specialist[online] Available at: https://www.tlv.com/global/TI/steam-theory/how-to-read-a-steam-table.html[Accessed 21 Oct 2017]

Tso, G. and Yau, K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural Networks, *Elsevier Energy*, Volume 32, pp. 1761–1768. Available at: https://www.journals.elsevier.com/energy [Accessed 6 Aug. 2017].

Tyagi, H. and Kumar, R. (2014). Optimization of a Power Plant by Using Data Mining and its Techniques. *International Journal of Advances in Science Engineering and Technology*, 2, pp. 83-87.

Wang, Y. (2000). A new approach to fitting linear models in high dimensional spaces.

Wayne, I. and Pat, L. (1992). Induction of One-Level Decision Trees. In: *ML92 Proceedings of the Ninth International Conference on Machine Learning*, Aberdeen, Morgan Kaufmann, pp. 233–240.

Witten, I. Frank, E. and Hall, M. (2011). *Data Mining : Practical Machine Learning Tools and Techniques*. 3rd ed. New York : Morgan Kaufmann.

Wold, S., Sjöström, M. and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems. 58 (2): 109–130

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Angus, N., Liu, B., Yu, P., Zhou, Z., Steinbach, M., Hand, D., and Steinb, D., (2008). Top 10 algorithms in data mining. *Knowl Inf Syst*, 14, pp. 1-37.

Yang, P. and Liu, S. (2004). Fault Diagnosis for boilers in thermal power plant by data mining, in: *Proceedings of Eighth International Conference on Control, Automation, Robotics and Vision*, Kunming, pp. 6–9.

Yiming, Y. and Jan, P. (1997). A comparative study on feature selection in text categorization. *ICML*.

Yu, L. and Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: *Proc. 20th International Conference on Machine Learning.* Washington, AAAI Press, pp. 856–863.

Yu, L. and Liu, H. (2004) Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res*., 5, pp. 1205–1224.

Yu, Z., Haghighat, F., Fung, B. and Yoshino, H. (2010). A decision tree method for building energy demand modeling. *Elsevier Energy and Buildings*, 42, pp. 1637–1646.

Zemansky and Mark, W. (1968). *Heat and Thermodynamics*, 5th ed. New York: McGraw-Hill. p. 275

Zhang, X. (2010). Fundamentals of Electric Power Systems, Restructured Electric Power Systems: Analysis of Electricity Markets with Equilibrium Models. Hoboken: Wiley,

Zhang, Y. (2009). Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM, *Chemical Engineering Science.* 64 (5) 801–811.

Zhou, H., Zhao, J., Zheng, L., Wanga, C., Fa, k. and Cen F., (2012), Modeling Nox Emissions From Coal- Fired Utility Boilers Using Support Vector Regression With Ant Colony Optimization, *Engineering Applications of Artificial Intelligence*, 25, pp. 147–158.

Zhou, Z. (2003). Three perspectives of data mining, *Artif Intell*, 143, pp.139–46.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), pp. 301–320.

APPENDICES

Appendix A : Steam Tables.

Appendix B: Moiller Chart

Appendix C: Components Diagram of Unit 3

Appendix D: Components Diagram of Unit 4

Appendix A : Steam tables

Press. (Abs.)	Temp.	Specific Volume		Specific Enthalpy		
kPa	ºC	m^3/ kg		kJ / kg		
P	T	Vf	Vg	hf	hg	hfg
1.0	6.970	0.00100014	129.183	29.30	2513.68	2484.38
2.0	17.495	0.00100136	66.9896	73.43	2532.91	2459.48
4.0	28.962	0.00100410	34.7925	121.40	2553.71	2432.31
6.0	36.160	0.00100645	23.7342	151.49	2566.67	2415.17
				173.85	2576.24	2402.39
200	120.21	0.00106052	0.885735	504.68	2706.24	2201.56
300	133.53	0.00107318	0.605785	561.46	2724.89	2163.44
400	143.61	0.00108356	0.462392	604.72	2738.06	2133.33
500	151.84	0.00109256	0.374804	640.19	2748.11	2107.92
600	158.83	0.00110061	0.315575	670.50	2756.14	2085.64
700	164.95	0.00110797	0.272764	697.14	2762.75	2065.61
800	170.41	0.00111479	0.240328	721.02	2768.30	2047.28
900	175.36	0.00112118	0.214874	742.72	2773.04	2030.31
1000	179.89	0.00112723	0.194349	762.68	2777.12	2014.44
1100	184.07	0.00113299	0.177436	781.20	2780.67	1999.47

Example of Saturated Steam Table (TLV Global, 2017).

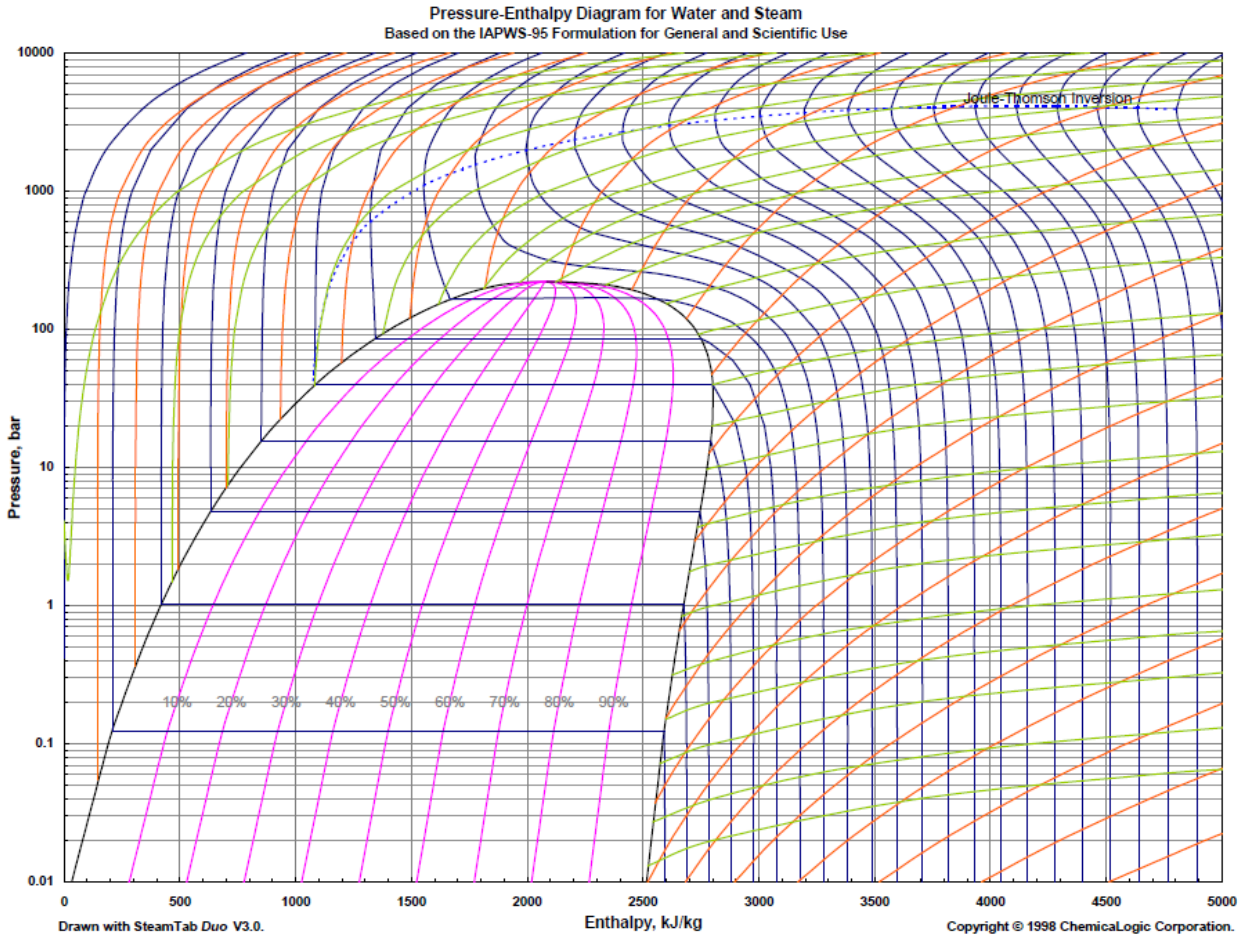
A saturated steam table is an indispensable tool for any engineer working with steam. It's typically used to determine saturated steam temperature from steam pressure, or the opposite: pressure from saturated steam temperature. In addition to pressure and temperature, these tables usually include other related values such as specific enthalpy (h) and specific volume (v).

The data found in a saturated steam table always refers to steam at a particular saturation point, also known as the boiling point. This is the point where water (liquid) and steam (gas) can coexist at the same temperature and pressure. Because H_2O can be either liquid or gas at its saturation point, two sets of data are required: data for saturated water (liquid), which is typically marked with an "f" in subscript, and data for saturated steam (gas), which is typically marked using a "g" in subscript (TLV Global, 2017).

Legend:

- P = Pressure of the steam/water
- T = Saturation point of steam/water (boiling point)
- v_f = Specific volume of saturated water (liquid).
- v_g = Specific volume of saturated steam (gas).
- h_f = Specific enthalpy of saturated water (energy required to heat water from 0°C (32°F) to the boiling point)
- h_{fg} = Latent heat of evaporation (energy required to transform saturated water into dry saturated steam)
- h_g = Specific enthalpy of saturated steam (total energy required to generate steam from water at 0°C (32°F)).

Appendix B: Mollier chart



Moiller Chart (Refrigeration basics, 2017).

Energy

Energy is the capacity of a system to do work where "system" refers to any physical system, not just a refrigeration system.

Enthalpy

Enthalpy is the total amount of heat in one Lb. of a substance. It's units are therefore BTU/Lb. The metric counter part is kJ/kg. (kilo joules/kilogram)

Entropy

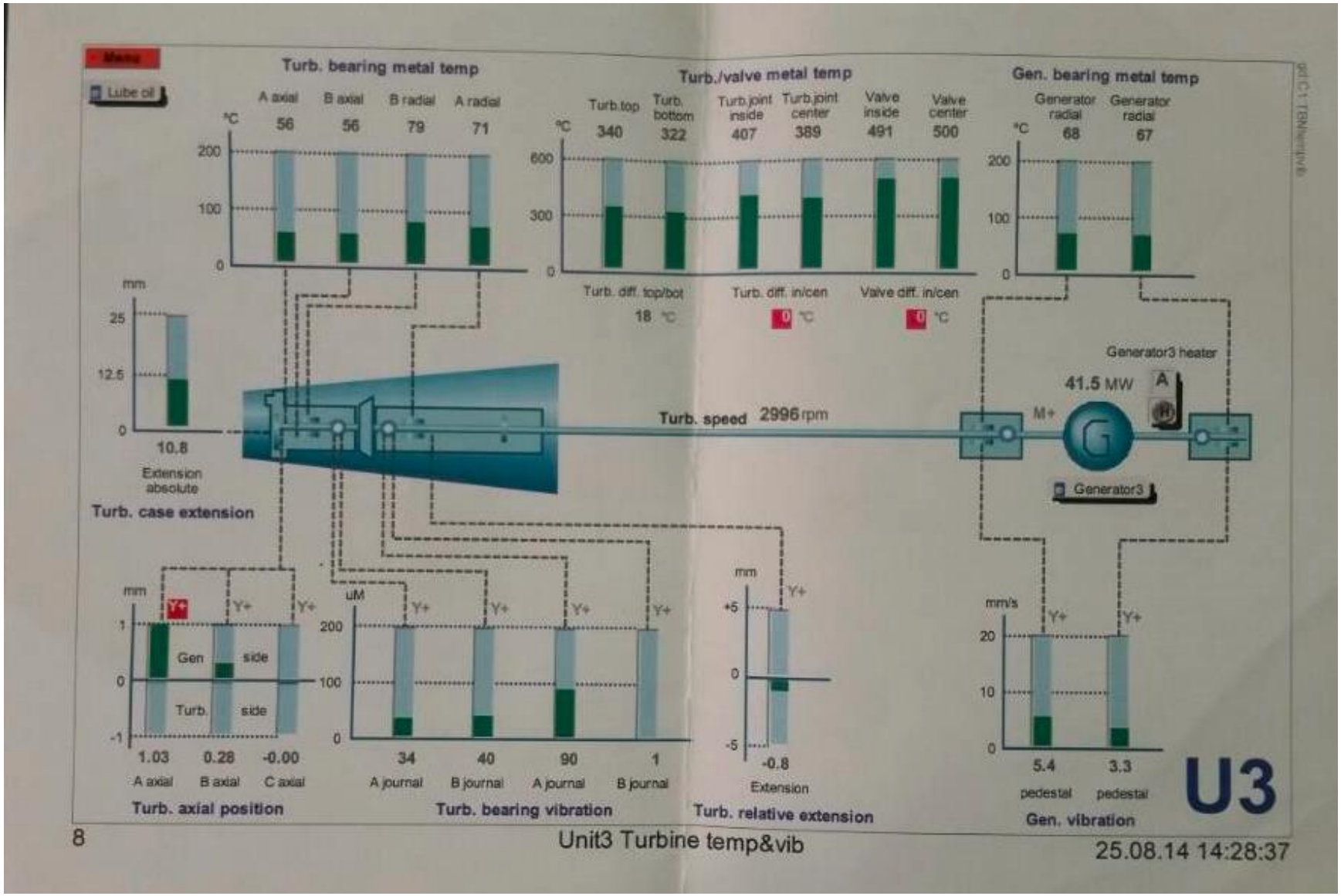
Entropy measures the energy dispersion in a system divided by temperature. This ratio represents the tendency of energy to spread out, to diffuse, to become less concentrated in one physical location or one energetic state. That spreading out is often done by molecules because molecules above absolute zero always have energy inside of them. That's why they are incessantly speeding through space and hitting each other and rotating and vibrating in a gas or liquid. Entropy is measured in BTU per lb. per °F.

Mollier Charts

A Mollier diagram is a graphic representation of the relationship between air temperature, moisture content and enthalpy - and is a basic design tool for building engineers and designers.

Mollier charts are used in designing and analyzing performance of vapour compression refrigeration systems. Each refrigerant has it's own chart which is a graph of the Enthalpy of a refrigerant during various pressures and physical states. Mollier charts are also called Pressure-Enthalpy diagrams. Pressure is shown on the verticle axis, enthalpy is on the horizontal axis. You can compare Imperial versus SI Unit Mollier Charts by clicking on the buttons below the chart. (Refrigeration basics, 2017).

Appendix C: Components Diagram of Unit 3



Appendix D: Components Diagram of Unit 4