

**The Influence of Ambient Weather Parameters on the  
Prediction of an Electrical Power Production of a Combined  
Cycle Power Plant in the UAE**

تأثير معاملات الطقس المحيط على التنبؤ بإنتاج الطاقة الكهربائية لمحطة توليد  
طاقة ذات دورة مركبة في دولة الإمارات العربية المتحدة

by

**HAJER ALI AHMED SAEED ALABDOULI**

**Dissertation submitted in partial fulfilment**

**of the requirements for the degree of**

**MSc ENGINEERING MANAGEMENT**

at

**The British University in Dubai**

**October 2022**

## DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.



---

Signature of the student

## **COPYRIGHT AND INFORMATION TO USERS**

The author whose copyright is declared on the title page of the work has granted to the British University in Dubai the right to lend his/her research work to users of its library and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar or the Dean of Education only.

Copying for financial gain shall only be allowed with the author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

## **ABSTRACT**

To improve the utilisation of power plants and enhance production, this study is devoted to predicting the baseload electrical power production of a combined cycle power plant in the UAE. The data for this study was taken from plant sensors over a period of one month (September 2021) from specific sensors installed in the power plant, and provided the data for input features that correspond to affect and change the electrical power production. In the UAE, the hot summer climate and ambient weather conditions adversely affect the performance of gas turbines (GT) and have an influence on steam turbines too. Accordingly, this paper studies four input variables: ambient temperature (ranges from 25.29°C to 36.5°C), relative humidity (ranges from 35.47% to 90.28%), atmospheric pressure (ranges from 0.99 bar to 1.01 bar) and exhaust steam vacuum (ranges from 0.057 bar to 0.126 bar). All influence the target variable (power production), which ranges from 506.32MW to 864.44MW. The change in the exhaust vacuum pressure in the steam turbine is affected by the change in ambient temperature, relative humidity, and atmospheric pressure in the gas turbine.

The analysis includes applying machine learning methods such as linear regression and artificial neural networks (ANNs) to develop a predictive power production model using different interactive computer programs such as Minitab, RStudio and Microsoft Excel. The linear regression model R-sq value was found to be 53.49%. Consequently, Minitab software is found to be a slightly more accurate statistical package compared to RStudio. In addition, the best data subset is found to be for week 1, with R-sq value of 82.16%. Moreover, the power linear regression model is ascertained to be more accurate than the ANN power predictive model, with a mean

absolute deviation of 46.385, symmetric mean absolute per cent error of 6.719 and residual standard error of 57.392 (Minitab outputs).

## ملخص

لتحسين استخدام محطات الطاقة ولتعزيز الإنتاج، تم تخصيص هذه الدراسة للتنبؤ بإنتاج الطاقة الكهربائية بالحمل الأساسي لمحطة توليد طاقة ذات دورة مركبة في دولة الإمارات العربية المتحدة. تم أخذ بيانات هذه الدراسة من أجهزة تسجيل بيانات في المحطة على مدى شهر كامل (سبتمبر 2021) بواسطة مستشعرات محددة ومثبتة في محطة الطاقة وقد قدمت هذه البيانات قيم متغيرات تؤثر على إنتاج الطاقة الكهربائية.

في الإمارات العربية المتحدة، يؤدي مناخ الصيف الحار والظروف الجوية المحيطة إلى تدهور أداء التوربينات الغازية والتوربينات البخارية أيضاً. وفقاً لذلك، يدرس هذا البحث أربع متغيرات إدخال وهي درجة الحرارة المحيطة (تتراوح من 25.29 درجة سيليزية إلى 36.5 درجة سيليزية)، والرطوبة النسبية (تتراوح من 35.47% إلى 90.28%)، والضغط الجوي (يتراوح من 0.99 بار إلى 1.01 بار)، وضغط العادم (يتراوح من 0.057 بار إلى 0.126 بار) والتي تؤثر جميعها على المتغير المستهدف (إنتاج الطاقة) والذي يتراوح من 506.32 ميغاواط إلى 864.44 ميغاواط. كما ويتأثر ضغط تفرغ العادم في التوربينات البخارية بالتغير في درجة الحرارة المحيطة والرطوبة النسبية والضغط الجوي في التوربينات الغازية. يتضمن تحليل هذه الدراسة تطبيق بعض الأساليب باستخدام التكنولوجيا مثل طريقة الانحدار الخطي وطريقة الشبكات العصبية الاصطناعية لتطوير نموذج إنتاج الطاقة التنبؤي عن طريق برامج حاسوب تفاعلية مختلفة مثل Minitab و RStudio و Microsoft Excel.

تم إيجاد قيمة R-sq لنموذج الانحدار الخطي وهي 53.49%. وبالتالي، وجد أن برنامج Minitab عبارة عن حزمة إحصائية أكثر دقة قليلاً مقارنةً ببرنامج RStudio وبالإضافة إلى ذلك، تم العثور على أفضل مجموعة فرعية للبيانات وهي للأسبوع الأول بقيمة R-sq مقدارها 82.16%. علاوة على ذلك، تم التأكد من أن نموذج الانحدار الخطي للطاقة المنتجة هو أكثر دقة من نموذج الشبكات العصبية الاصطناعية بمتوسط انحراف مطلق قدره 46.385، ومتوسط نسبة الخطأ المطلق المتمثل بمقدار 6.719 والخطأ المعياري المتبقي بمقدار 57.392.

## **ACKNOWLEDGEMENTS**

I want to express my deepest gratitude to Prof. Alaa A. A-Ameer, Head of the MSc in the Engineering Management Programme, who supervised this work and shared his advice and guidance throughout this research journey.

I would also like to recognise the invaluable support of Dr Sa'Ed M. Salhieh, Associate Professor in BUiD, who is much appreciated for his assistance with the statistics used in this research.

**Table of Contents**

**List of Tables..... iv**

**List of Figures .....vii**

**Table of Contents ..... i**

**List of Abbreviations..... x**

**CHAPTER I..... 1**

**INTRODUCTION..... 1**

**1.1 Background..... 1**

    1.3.1 Electrical Energy Production and Consumption ..... 1

**1.2 Research Problem Statement ..... 3**

**1.3 Research Conceptual Framework ..... 4**

**1.4 Dissertation Research Questions ..... 6**

**1.5 Dissertation Aims and Objectives ..... 6**

    1.5.1 Research Aims .....6

    1.5.2 Research Objectives .....6

**1.6 Research Limitations ..... 7**

**1.7 The Organisation of the Dissertation ..... 7**

**CHAPTER II ..... 10**

<b>LITERATURE REVIEW .....</b>	<b>10</b>
<b>2.1 Machine Learning Methods and Predicting Quality .....</b>	<b>10</b>
<b>2.2 Statistical Modelling and Prediction Equations .....</b>	<b>11</b>
<b>2.3 Regression Methods Used for Modelling Dynamic Processes and Systems .....</b>	<b>12</b>
<b>2.4 Checking the Prediction Model’s Accuracy .....</b>	<b>26</b>
<b>2.5 Prediction of the Power Production Using Regression Analysis.....</b>	<b>29</b>
<b>2.6 Power Plants’ Production Concepts.....</b>	<b>31</b>
<b>2.7 Types of Power Plants.....</b>	<b>32</b>
<b>2.8 An Overview of a Combined Cycle Power Plant.....</b>	<b>35</b>
<b>2.9 Power Generation in UAE .....</b>	<b>38</b>
<b>CHAPTER III.....</b>	<b>42</b>
<b>RESEARCH METHODOLOGY .....</b>	<b>42</b>
<b>3.1 Linear Regression Modelling .....</b>	<b>42</b>
<b>3.2 Artificial Neural Networks.....</b>	<b>43</b>
<b>3.3 Checking the Accuracy of Models.....</b>	<b>48</b>
<b>CHAPTER IV .....</b>	<b>50</b>
<b>RESEARCH RESULTS AND DISCUSSION .....</b>	<b>50</b>
<b>4.1 Data Summary .....</b>	<b>50</b>

<b>4.2</b>	<b>Linear Regression Analysis</b> .....	<b>55</b>
<b>4.3</b>	<b>Linear Regression Using Program Coding</b> .....	<b>71</b>
<b>4.4</b>	<b>Artificial Neural Networks</b> .....	<b>79</b>
<b>4.5</b>	<b>Checking the Model’s Accuracy</b> .....	<b>82</b>
	<b>CHAPTER V</b> .....	<b>87</b>
	<b>CONCLUSION AND RECOMMENDATIONS</b> .....	<b>87</b>
<b>5.1</b>	<b>Conclusion</b> .....	<b>87</b>
<b>5.2</b>	<b>Recommendations</b> .....	<b>89</b>
	<b>References</b> .....	<b>90</b>
	<b>Appendices</b> .....	<b>103</b>
	Appendix A: Data Analysis Details .....	103

## LIST OF TABLES

Table 1: Linear regression functions and their transformations (Hadi & Chatterjee 2012).....	21
Table 2: Descriptive Statistics for electrical power production .....	50
Table 2: Descriptive Statistics for electrical power production .....	51
Table 3: Descriptive Statistics for ambient temperature .....	51
Table 4: Descriptive Statistics for relative humidity. ....	52
Table 5: Descriptive Statistics for atmospheric pressure .....	52
Table 6: Descriptive Statistics for exhaust steam pressure.....	53
Table 7: Covariance table for the dataset .....	53
Table 8: Correlation table for the four input variables .....	54
Table 9: R-squared and adjusted R-squared values for the whole dataset model and each week's predictive model .....	57
Table 10: Dataset model summary using Minitab.....	60
Table 11: Analysis of variance for the whole dataset prediction model .....	60
Table 12: Coefficients table for the whole dataset prediction model .....	60
Table 13: Data summary using RStudio software for the actual value of power production.....	71
Table 14: Data summary using RStudio software for the 70% learning data .....	72
Table 15: Data summary using RStudio software for the 30% testing data.....	77
Table 16: Summary table for predicted and actual power production values with the input variables .....	78
Table 17: The first ten actual and predicted values and the accuracy calculations performed on the Minitab linear regression prediction model.....	83

Table 18: The first ten actual and predicted values and the accuracy calculations performed on RStudio linear regression prediction model .....	83
Table 19: The first ten actual values and the accuracy calculations performed on neural network predicted values .....	84
Table 20: Model accuracy checking for linear regression models and neural network predicted values .....	84
Table A. 1: Week 1 power prediction model equation.....	105
Table A. 2: Week 1 model summary using Minitab .....	105
Table A. 3: Coefficients table for week 1 prediction model .....	105
Table A. 4: Analysis of variance for the week 1 prediction model .....	105
Table A. 5: Week 2 power prediction model equation.....	107
Table A. 6: Week 2 model summary using Minitab .....	107
Table A. 7: Coefficients table for week 2 prediction model .....	107
Table A. 8: Analysis of variance for week 2 data subset .....	108
Table A. 9: Week 3 power prediction model equation.....	110
Table A. 10: Week 3 model summary using Minitab.....	110
Table A. 11: Coefficients table for week 3 prediction model .....	110
Table A. 12: Analysis of variance for week 3 data subset .....	110
Table A. 13: Week 4 power prediction model equation.....	112
Table A. 14: Week 4 model summary using Minitab.....	112
Table A. 15: Coefficients table for week 4 prediction model .....	113
Table A. 16: Analysis of variance for week 4 data subset .....	113
Table A. 17: Week 5 power prediction model equation.....	115

Table A. 18: Week 5 model summary using Minitab.....	115
Table A. 19: Coefficients table for week 5 prediction model .....	115
Table A. 20: Analysis of variance for week 5 data subset .....	115
Table A. 21: A sample of the actual data after normalisation .....	117

## LIST OF FIGURES

Figure 1: Domestic consumption of electricity and the amount of used water in the UAE from 2007 to 2016 (Banhidarrah et al. 2020) .....	3
Figure 2: Interactions between inputs, system and outputs .....	5
Figure 3: Plant overview with inputs and output sensor positions .....	6
Figure 4: Graphical representation of Neural Network Method (Cao, Sai, Fu, and Duan 2020)..	18
Figure 5: Graphs of the linear regression function $Y = \alpha X \beta$ (Hadi & Chatterjee 2012) .....	22
Figure 6: Graphs of the linear regression function $Y = \alpha e^{\beta X}$ .....	23
Figure 7: Graphs of the linear regression function $Y = \alpha + \beta \log X$ .....	23
Figure 8: Graphs of the linear regression function (a) $Y = X \propto X - \beta$ and (b) $Y = e\alpha + \beta X$ .....	24
Figure 9: Linear model and Poisson model comparison .....	25
Figure 10: Conventional and non-conventional power plants .....	32
Figure 11: Total power generation classified by energy source in US in 2007 .....	34
Figure 12: A simple flow diagram of a combined cycle .....	36
Figure 13: The nuclear power plant's net capacity under construction .....	40
Figure 14: ANN method's flow chart .....	47
Figure 15: Power in MW versus time in weeks .....	55
Figure 16: Normality plot for the dataset prediction model .....	61
Figure 17: Residuals plot versus fitted values for the dataset prediction model .....	61
Figure 18: Histogram of residuals for the dataset prediction model .....	62
Figure 19: Scatter plot of power production versus ambient temperature .....	64

Figure 20: Scatter plot of power production (P) versus relative humidity (RH).....	66
Figure 21: Scatter plot of power production (P) versus atmospheric pressure (AP).....	68
Figure 22: Scatter plot of power production (P) versus exhaust vacuum (V).....	69
Figure 23: Residuals of the actual power production model using RStudio .....	73
Figure 24: Coefficients of the actual power production model using RStudio .....	74
Figure 25: Other values for the actual power production model's hypothesis testing .....	74
Figure 26: Residuals versus leverage plot for the actual power production.....	76
Figure 27: Actual versus predicted values of power production using linear regression by RStudio .....	78
Figure 28: The correlation between the actual and predicted power values .....	79
Figure 29: ANN model for the CCPP data.....	80
Figure 30: Neural network plot for the dataset after normalisation .....	81
Figure 31: Actual versus normalised predicted values of power production using linear regression .....	82
Figure 32: The correlation between the actual and predicted power values after normalisation ..	82
Figure A. 1: Power in (MW) versus time in (seconds) .....	103
Figure A. 2: Power in (MW) versus time in (min).....	104
Figure A. 3: Power in (MW) versus time in (hours) .....	104
Figure A. 4: Power in (MW) versus time in (days).....	104
Figure A. 5: Normality plot for the week 1 prediction model.....	106
Figure A. 6: Residuals plot versus fitted values for the week 1 prediction model.....	106
Figure A. 7: Histogram of residuals for week 1 prediction model.....	107
Figure A. 8: Normality plot for week 2 prediction model.....	108

Figure A. 9: Residuals plot versus fitted values for week 2 prediction model.....	109
Figure A. 10: Histogram of residuals for week 2 data subset prediction model .....	109
Figure A. 11: Normality plot for week 3 prediction model.....	111
Figure A. 12: Residuals plot versus fitted values for week 3 prediction model.....	112
Figure A. 13: Histogram of residuals for week 3 data subset prediction model .....	112
Figure A. 14: Normality plot for week 4 prediction model.....	114
Figure A. 15: Residuals plot versus fitted values for week 4 prediction model.....	114
Figure A. 16: Histogram of residuals for week 4 data subset prediction model .....	115
Figure A. 17: Normality plot for week 5 prediction model.....	116
Figure A. 18: Residuals plot versus fitted values for week 5 prediction model.....	116
Figure A. 19: Histogram of residuals for week 5 data subset prediction model .....	117
Figure A. 20: Residuals versus fitted values plot for the actual power production.....	118
Figure A. 21: Normal probability plot using RStudio for the actual power production.....	119
Figure A. 22: Scale location plot for the actual power production.....	119

## LIST OF ABBREVIATIONS

<b>HRSG</b>	Heat Recovery Steam Generator
<b>UAE</b>	United Arab Emirates
<b>P</b>	Electrical Power Production
<b>AT</b>	Ambient Temperature
<b>RH</b>	Relative Humidity
<b>AP</b>	Atmospheric Pressure
<b>V</b>	Exhaust Vacuum
<b>ANN</b>	Artificial Neural Networks
<b>MLM</b>	Machine Learning Methods
<b>AI</b>	Artificial Intelligence
<b>Y</b>	Output Variable
<b>X<sub>n</sub></b>	Independent Variable
<b>β<sub>0</sub></b>	Intercept of the Linear Equation
<b>β<sub>1</sub></b>	Slope of the Linear Equation
<b>ε</b>	Error Value
<b>GLM</b>	Generalised Linear Model
<b>CCPP</b>	Combined Cycle Power Plant
<b>H<sub>0</sub></b>	Null Hypothesis
<b>H<sub>1</sub></b>	Alternative Hypothesis
<b>MLP</b>	Multi-Layer Perceptron
<b>z<sub>i</sub></b>	Normalised Value
<b>R-sq</b>	R-squared
<b>R-sq (adj)</b>	Adjusted R-squared
<b>DF</b>	Degree of Freedom
<b>SOFC</b>	Solid Oxide Fuel Cell
<b>w<sub>i</sub></b>	Weights
<b>b</b>	Bias Weights
<b>MSE</b>	Mean Squared Error
<b>RMSE</b>	Root Mean Square Error
<b>MAE</b>	Mean Absolute Error

<b>MAPE</b>	Mean Absolute Percentage Error
<b>RSE</b>	Relative Standard Error

# CHAPTER I

## INTRODUCTION

### 1.1 Background

Developing prediction models is a technique used to solve real-life problems. It shows a clear relationship between variables, and hence supports decision-making regarding the interactions between these variables, and comes up with conclusions (Tüfekci 2014). From an engineering perspective variables interact together and affect the final output of a system; what happens in power plants is an example (Kunming & Zhou 2012). A power plant consists of a group of systems and subsystems that work together to produce electrical power sent subsequently to be used in facilities (Raja, Srivastava & Dwivedi 2006). Pertaining to this engineering facility, it is essential to predict the net energy yield or the electrical power production (P) of the plant to maximise the profit, support development of the performance, help managers in the decision-making process and for general economic purposes (Tüfekci 2014).

#### 1.3.1 Electrical Energy Production and Consumption

As per Aranda et al. (2012), the consumption of power based on a worldwide range was 22 trillion KW/hr in 2011, which is more than double consumption in 1975. In addition, some predictions state that the growth rate of power consumption will be 2.2% in 2040. According to Aranda et al. (2012), this growing power consumption will require the infrastructure of the whole world to increase its budget by over \$12 trillion over 30 years, starting from 2011. Moreover, Omer (2012) states that the worldwide manufacture of hydrocarbons is harming the environment in a continuous manner. As a result, environmental regulations are requested to reduce greenhouse gases emissions. This is done by lowering the cost of emissions mitigation and by effective

planning by power producers. In addition, effective planning is a crucial management tool that shows its importance in dealing with worldwide transmission expansion (Eccleston & March 2014).

As per Granit & Löfgren (2010), the United Arab Emirates (UAE) is an arid country where economic growth means a high demand for electricity and water, especially with the primary considerations of climate change and climate instability. In the UAE, Banhidarah et al. (2020) showed that electrical power consumption has increased gradually from 2007 to 2016, which drives the concentration on better planning and more efficient power production. Figure 1 shows the domestic consumption of electricity in units of TWh in the UAE during the same period and the quantity of consumed water in a million cubic meters as illustrated by Banhidarah et al. (2020).

Raja, Srivastva and Dwivedi (2006) stated that electricity has a significant impact on the development sectors, and it is the most multipurpose type of energy. Consequently, its growth rate is running faster than any other type of energy. In fact, the power industry has shown a remarkable growth rate in its usage in technological and economic development over the last few decades. Moreover, electricity consumption in any country indicates its high productivity and growth.

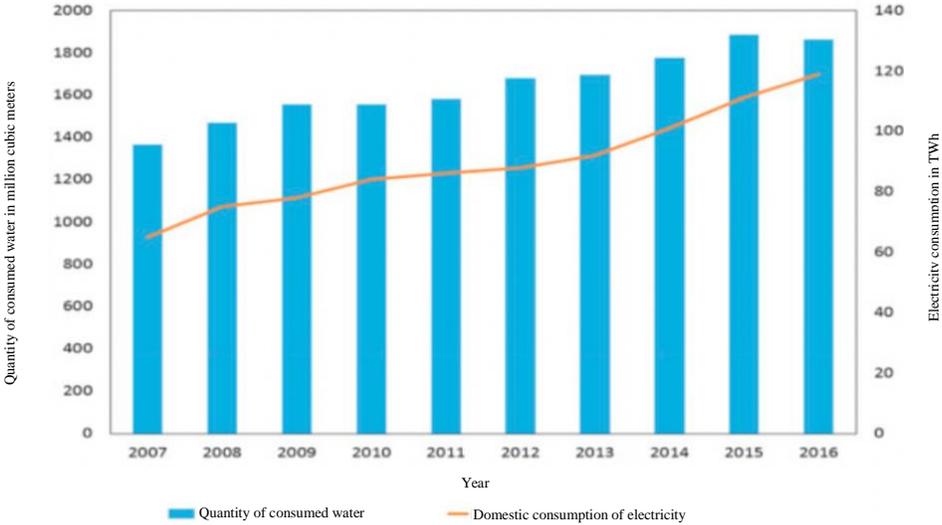


Figure 1: Domestic consumption of electricity and the amount of used water in the UAE from 2007 to 2016 (Banhidarrah et al. 2020)

A combined cycle power plant (CCPP) consists of gas and steam turbines along with heat recovery steam generators (HRSG), all grouped to produce electrical power (Tüfekci 2014). Few studies have discussed the effect of ambient parameter conditions on the target power production in the UAE and Arab countries. This paper will give recommendations to power plants based on relative weather conditions. Thatcher (2007) mentions that planning strategies for electricity generation are essential to provide feedback on economics and costs, to reduce losses and give recommendations for efficiency considerations. In this research, the power generation and some factors that affect its production will be revealed. Moreover, the deviations and uncertainties in the prediction model will be shown, to develop a plan that will reduce the load, balance it and reduce costs. In general, robust planning and the usage of strong models mitigate the uncertainties in the production charts. This research will focus mainly on the effect of ambient parameters on power production using linear regression analysis, artificial neural networks (ANNs), graphs and trends to help in decision-making, give recommendations, optimise the utilisation of power production in the UAE and highlight the efficiency concerns over baseload power plant production with respect to ambient variables.

## **1.2 Research Problem Statement**

A CCPP in the UAE intended to optimise the utilisation of power production and increase its efficiency was affected by many factors. These factors may reduce the performance of GTs, steam turbines and HRSGs and lead to lower power production. Some factors such as ambient weather parameters were neglected when focusing on optimising the performance of the power

plant. There have been few studies of the effect of ambient weather parameters, which leads to their effects being ignored and left uncontrolled.

In addition, power prediction models certainly facilitate forecasting future production. Accordingly, some statistical packages are used to predict power with lower accuracy compared to others and, hence, less accurate prediction models are presented. This study will compare some statistical software and suggest the advantages of using each piece of software to achieve better calculations.

### **1.3 Research Conceptual Framework**

Input variables were chosen based on studies that showed how these ambient variables might affect steam and gas turbine functionality. For instance, the load of the gas turbine is impacted by some ambient weather conditions, such as ambient temperature (AT), atmospheric pressure (AP) and relative humidity (RH). On the other hand, a steam turbine's performance is influenced by the exhaust steam pressure (Tüfekci 2014). All the input and output variables correspond to measured data that is collected from specific sensors. The ranges are illustrated below:

- ambient temperature: input to the system measured in Celsius and ranges between 25.29°C and 36.5°C.
- atmospheric pressure: input to the system measured in units of hectopascal and ranges from 998.75 hPa to 1009.82 hPa.
- relative humidity: input to the system that is measured as a percentage and ranges from 35.46% to 90.27%.

- exhaust steam pressure (V): input to the system measured in bars and ranges from 0.057 to 0.126 bars.
- electrical power production: this is the production variable measured in megawatts and ranges from 506.32MW to 864.44MW.

The P is affected by all these inputs, as Figure 2 shows:

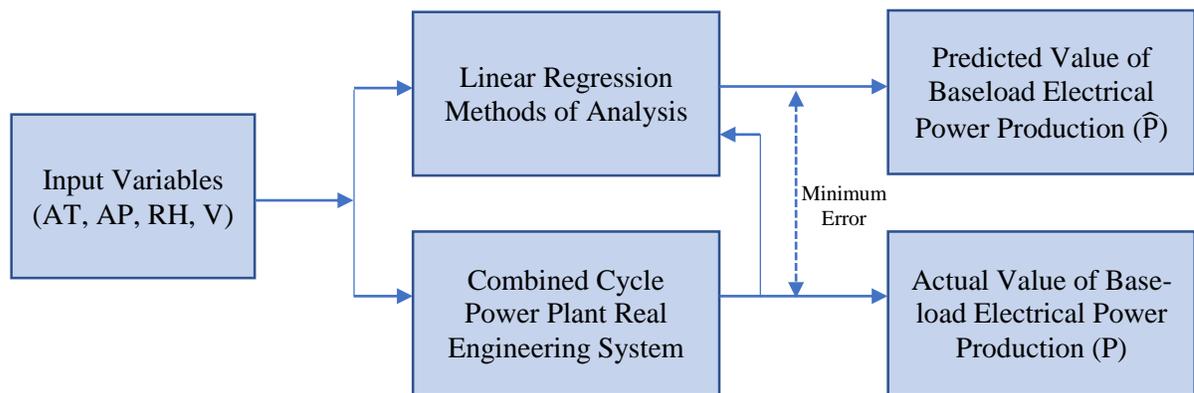


Figure 2: Interactions between inputs, system and outputs

Figure 3 shows the CCPP overview and the sensors where the input and output measurements were recorded.

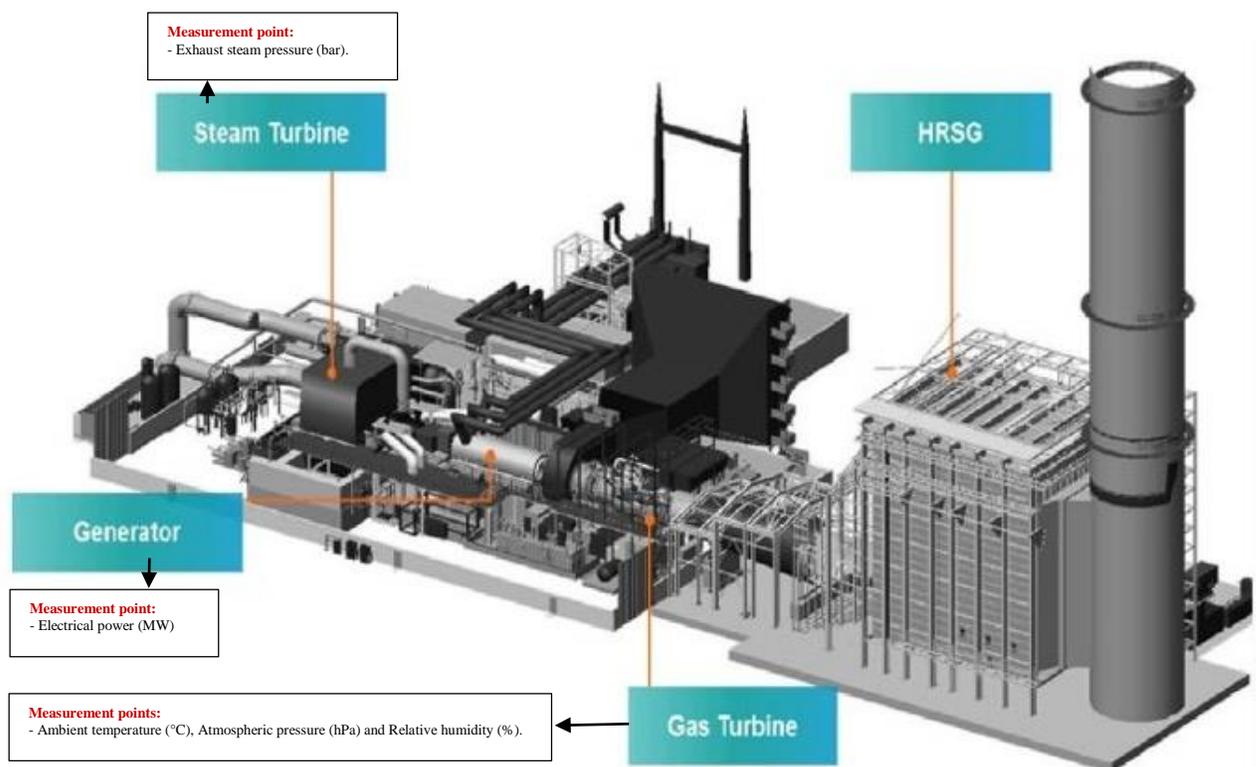


Figure 3: Plant overview with inputs and output sensor positions

(Verdict 2020)

## **1.4 Dissertation Research Questions**

What are the main ambient parameters that might affect the performance of gas and steam turbines in a CCPP? Are there any suggestions to improve the efficiency of CCPP production based on UAE weather conditions? How does linear regression differ from ANNs when performing statistical analysis and checking the model's accuracy? Is there any preferred statistical package when performing such analysis?

## **1.5 Dissertation Aims and Objectives**

This section illustrates the aims and objectives of the research. Generally, research aims demonstrate the main objectives of the research while the research objectives define the specific aims of the research.

### **1.5.1 Research Aims**

The main aim of this research is to improve the CCPP utilisation in the UAE by performing two methods of analysis. This involves using different software on real data to find the P based on ambient weather parameters, to enable better managerial decision-making.

### **1.5.2 Research Objectives**

- To predict electrical power as a product of a CCPP in the UAE based on the influence of input variables such as AT, AP, RH and V pressure using linear regression.

- To find the correlation between the input variables of the system.
- To use a machine learning method such as ANNs to analyse the data and compare the results with the linear regression model.
- To compare the accuracy of two statistical packages.

## **1.6 Research Limitations**

This section illustrates some research limitations and alternatives that could be used, such as:

- Lack of previous research on the topic in the UAE: there is a lack of research discussing the aims and objectives of this study in the UAE and the Arab region, so an intense search was done to find the pertinent facts and compare them with other regions, to help with the interpretation of the model and support the calculations.
- Data collection process: data was collected over a one-month period – September 2021. The weather conditions during the month were studied to show their effect on power production. The limitation was generalising the conclusions of this one-month data to the 12 months of the year.
- Method of analysis: this research focuses on two analysis methods, the linear regression method, as studied in the Engineering Statistics Module (ENGM501), and ANNs. However, there are many methods of analysis that are considered to be a measure for determining the relationship between input and output variables or variable interaction. This study uses three software packages: Minitab, RStudio and Microsoft Excel. There are many other software programs developed for statistical analysis.

## **1.7 The Organisation of the Dissertation**

This section explains the organisation of the dissertation, which is illustrated as follows.

**Chapter I** is the introduction of the research, starting with a background section about the techniques used to develop prediction models and the concept of CCPPs and P. Following that, the electrical energy production and consumption section demonstrates the development of P in the UAE, growth rates and the domestic consumption of electricity. The second part of the introduction is the research problem statement, which explains the problems of the research that need to be solved. The research conceptual framework section in the introduction shows the list of variables that will be studied in the research and how these variables interact with each other. The dissertation research question section presents some questions that will be answered after performing the analysis and finalising the research discussions. Accordingly, the objective of the dissertation section is divided into the main and specific objectives described specifically in this chapter. Finally, the dissertation limitation section illustrates the limitations of the research, which are the characteristics of implementing the analysis that can influence the interpretation of the research findings.

**Chapter II** is the literature review chapter, which represents the techniques used in developing power models and in predicting quality. Furthermore, it shows the advantages and disadvantages of using some statistical techniques used in the research to answer the dissertation question, such as regression analysis and ANNs. It discusses the accuracy and deficiencies of some types of prediction analysis and the concept of transformation of data. This section also presents a detailed description of different types of power plants, electrical energy production processes and why the power production is shifted towards nuclear power production. Additionally, it describes P in the UAE, presents previous studies in the field of predicting power production, and compares the outcomes of each study.

**Chapter III** describes the research methodology. In this chapter, the methods of analysis used to analyse the dataset are described in detail. They include linear regression and ANNs. Linear regression is a method by which prediction models are developed to predict a phenomenon that happens for a variable based on some input parameters. ANN is a method based on artificial intelligence (AI) whereby specific statistical packages are used to generate network connections between variables. The steps involved in both methods are illustrated in this chapter. Additionally, checking the accuracy of a model by using equations is also specified. Generally, this section is related to identifying the processes of data collection, data analysis and finding alternatives to the problems specified.

**Chapter IV** contains the most important aspects of this study and is about presenting the results obtained and discussions after performing statistical analysis for the dataset. In this section all tables and figures are presented for each method of analysis. Moreover, this chapter contains data summary tables, correlation and covariance tables for the dataset, and prediction model equations developed after performing linear regression and ANN. Following that, a comparison between the prediction models is presented based on accuracy and the most accurate statistical package is illustrated.

**Chapter V** presents the summary of this research, the conclusions and recommendations. The main topic discussed in the literature review chapter is presented here, which is the statistical methods of analysis used to develop prediction models for power plants, especially CCPPs. For instance, linear regression and ANNs are presented in this section as a summary of the methodology used in this research. Additionally, the main findings of this study are illustrated, and the most accurate statistical model presented here along with the preferred statistical package for statistical analysis. Finally, recommendations for future research are presented.

## **CHAPTER II**

### **LITERATURE REVIEW**

#### **2.1 Machine Learning Methods and Predicting Quality**

According to Sankhye and Hu (2020), machine learning methods (MLMs) are ways of prediction that analyse the values of a product and its quality by analysing data of variables that may affect its quality. MLMs are valuable tools for assessing the significance of the supply chain parameters. In addition, Najah Ahmed et al. (2019) state that MLMs are part of AI, which can be used to analyse data and come up with conclusions regarding the interactions between data variables by using computer science and specific algorithms. In fact, as per Sankhye and Hu (2020), industry 4.0 and the concept of smart factories made the process of extensive data collection during production stages easier by using computer systems and specific sensors. While there has been substantial research into forecasting the quality of some production processes, there has been little research into the use of classification algorithms to predict the overall quality of production.

MLMs, according to De et al. (2010), are a major element of the fast-expanding field of statistics. In data mining fields, algorithms are employed to make categorisations or forecasts using statistical methodologies and to provide crucial insights. Accordingly, these insights are used to improve decision-making within corporations and organisations, influencing the critical development of key growth metrics. The demand for data scientists will increase as the database for study expands and evolves, with their help needed to identify business problems and, as a result, detect the data required to answer the questions.

## **2.2 Statistical Modelling and Prediction Equations**

Walter (2013) explained statistical modelling and linear regression, and it is summarised as follows.

Models are considered a mathematical explanation of an observation of a processed dataset. The statistical models include equations illustrating the relationship or the impact of variables on each other. These models explain the probability distribution of these variables, and, usually, the assumptions of these models are mainly the random distribution or random variation. Moreover, statistical models show two parts in one equation: the systematic part and the random part.

Accordingly, Nick (2007) stated that descriptive statistics is the science that gathers, summarises and analyses data based on random variation. Similarly, Holcomb (2016) mentioned more specific terminology descriptions for descriptive statistics, which are the measures in graphs, averages, percentages and equations used to analyse whether this data comes from samples or populations. Likewise, another science called inferential statistics is based on the process of generalisation from sample outcomes to populations. As per Pérez-Vicente and Expósito Ruiz (2009), the importance of using such analysis is to describe the trends and most important features found in a dataset that refers to amounts and information related to an area or a topic of interest.

Furthermore, Nick (2007) explained that graphics and descriptive statistics such as mean, mode, median, standard deviation, maximum value, minimum values and skewness describe quantitative and qualitative data variables. Similarly, Holcomb (2016) defined using the term parameter when the analysed variable comes from the population. On the other hand, it is called a statistic if it comes from a sample.

### **2.3 Regression Methods Used for Modelling Dynamic Processes and Systems**

Regression methods of analysis are management tools and a prediction technique enabling processes and systems to predict a target variable based on various inputs (Tüfekci 2014). For thermodynamic systems such as a CCPP, the P can be predicted by constructing a linear regression (Aranda et al. 2012). These power plants consist of a system of one steam turbine, two gas turbines and one HRSG. For the gas turbine to become more reliable and sustainable, it is essential to predict its power generation, especially when working for a high-profit load and high liability (Tüfekci 2014).

There are various types of regression methods, as discussed by Tso and Yau (2007), such as modelling stationary gas turbines using ANNs technique to estimate its behaviour while operating for wide ranges of input data starting from full speed and zero load to full-load conditions. In research by Amozegar and Khorasani (2016), other methods such as radial basis function and multi-layer perceptron (MLP) are used to identify stationary gas turbines in the startup stage. Moreover, according to Han and Kamber (2011), feed-forward neural networks techniques and dynamic linear models were analysed and assessed for their capacity to predict gas turbine behaviour. It was found that neural network methods of analysis are used to identify the behaviour of gas turbines better than dynamic linear models. As per Chabot and D'Arras (2010), the neural network technique has been used to classify models or factors based on real data by giving a set of variables, inputs or some features assuming they interact to study phenomena where other features are discarded if found to be irrelevant to the phenomena's classification.

Several studies have shown that using MLM to predict electrical energy consumption (Azadeh, Saberi & Seraj 2010; Ekonomou 2010; Che, Wang & Wang 2012; Kavaklioglu 2011;

Leung & Lee 2013). Similar studies related to this paper are found, such as Tüfekci (2014), in which machine learning regression methods were used to predict the P of a CCPP, consisting of one unit of steam turbine and three units of gas turbines and one system of HRSG.

Tüfekci (2014) showed the definition of regression analysis and its types, such as:

**1- Simple Linear Regression:** this method is performed to obtain a model with the smallest value of squared error. It is used to fit each input variable ( $a_1$ ) with the output ( $x$ ) in an equation as stated below, where  $w_0$  is the intercept and  $w_1$  is the slope of the linear regression equation estimated by the method of least squares, which is known as the difference between the predicted values and the actual values.

$$x = w_0 + w_1 a_1 \quad (1)$$

**2- Multiple Linear Regression:** the link between one or more independent variables and a dependent variable is represented by this method, which is based on mathematical modelling. Linear regression is used to predict the output ( $x$ ) for its relationship with the independent input variables such as  $a_1, a_2$ , etc. The least-squares method is also performed here to illustrate the linear relations in the observed data. The equation below specifies the linear dependency of the inputs and the target output variable where  $w_0$  is the intercept of the equation and  $w_1$  is the weight of the first input variable (Tüfekci 2014).

$$x = w_0 + w_1 a_1 + \dots + w_k a_k \quad (2)$$

Here,  $w_k$  is the weight of input in k order and  $a_k$  is the input variable in k order.

Multiple linear regression analysis involves the ANOVA table. In the ANOVA table, the degree of freedom (DF) is the summation of the individual degrees of freedom for the samples.

Moreover, it implies the number of rows in the data used to fit the regression model (Kruggel, Pélégriani-Issac & Benali 2002).

Previous research was done to examine the effect of ambient parameters on power generation. One study was performed using power plant data in Turkey. The analysis was done to show the best machine learning regression method used, with the most accuracy, to predict the full-load electric power yield, which was the bagging algorithm using REPTree analysis. The study showed general trends and effects of some ambient parameters on power and mainly focused on the multilinear regression methods as a method of analysis. The data was collected over six years using a computer system to record the ambient parameters of gas and steam turbines each second, along with power production (Tüfekci 2014).

Aranda et al. (2012) applied multilinear regression analysis to predict energy consumption annually in the Spanish banking sector. This paper mainly focused on data taken from 55 banks in Spain, where the data was validated, analysed and generated models to predict energy consumption. Three models were produced: the first was to estimate electricity consumption in bank branches. The second was intended for areas with low winter climate intensity, and the third was for branches with high winter climate severity. The main results were that the first model had the lowest determination coefficient, which means it is suitable for forecasting the energy utilisation of banks and detecting inefficiency in banks with poor energy consumption performance. In addition, the first model had the lowest uncertainty.

Searle and Gruber (2016) stated that there might be some difficulties and problems with the definition, the occupation and the status of the variable (a) as there are different types and meanings of the representations that can be misunderstood from the model. For example, the correspondent number of the variable as a multiplication (1,2,3, etc.) may not correspond directly and accurately

to the dependent variable  $x$ . It is not accurate to say that a professional worker has a four times higher status than a regular labourer. So, whatever the representation is, it is essential to describe it as arbitrary because it might cause a problem for the proposed model.

Another study described the demand for electricity based on 30-minute intervals in a power plant in Australia. The intervals were consistent with changes in the climate and predicted the power based on demand using duration curves and multilinear regression. Power plants in some Australian states were selected in that study to reach the highest accuracy in the calculations. These datasets were very valuable in predicting the demand for electricity and were essential for economic purposes (Thatcher 2007).

Li (2015) has presented some advantages of using linear regression analysis while showing a relationship between variables, as summarised as follows:

- It is a widely used method whereby the model's calculations allow direct interpretation of relationships.
- If the assumptions of the linear regression method are satisfied, such as the normal distribution of the residuals, the resulting parameters will be efficient and neutral.

On the other hand, there are some disadvantages, as stated by Li (2015), such as:

- In typical situations and in most management researches, the data distribution is heterogeneous and tails are not exponentially bounded (heavy-tailed). As a result, the normality and its assumption will be questioned.
- Moreover, the outliers, which are the values that are substantially different from other variable data, affect the distribution of the data and might intensely affect the result of the calculations and the relationships that would be revealed. In addition, many journals related

to management decisions using the linear regression method do not reveal how the outliers were addressed in the study.

**3- Least Median Square (LMS):** Tüfekci (2014) states that it is a linear regression analysis completed by reducing the squared error of the median. Using the regression equation, where  $i$  is the data variable,  $k$  is the total number of data points, and  $j$  is the corresponding data point, the slope of each input variable is reallocated to reduce the median of the squares of the difference between observed and predicted value as follows:

$$LMS = Median_i \left( x^{(i)} - \sum_{j=0}^k w_j a_j^{(i)} \right) \quad (3)$$

Here, LMS is the least median square,  $Median_i$  is the median in order  $i$  for the iteration equation,  $x^{(i)}$  is the output variable in order  $i$ ,  $w_j$  is the slope of the linear equation,  $a_j^{(i)}$  is the input variable. All the iteration steps starts from order  $j=0$  to  $k$  for the dataset.

This method has been used by Morano and Tajani (2014) to identify and remove the outliers to develop a predictive regression model for real estate analysis. Based on that, the data were classified into normal observations, perpendicular outliers, significant and insignificant leverage points, points to remove and points to keep.

**4- Artificial neural networks:** ANNs and the human brain are the widely known comparison and expressions used to explain the concepts behind its terminology. When it comes to learning, the human brain quickly recognises a picture or some process due to interactions that happen inside the human brain networks between one nerve cell and another. The same happens in computers that require specific coding and continuous improvement to operate some networks that analyse data and understand its phenomena (Krogh 2008).

An ANN is a computer system composed of many neurons that interact with each other and present a supportive model for making decisions and conclusions regarding these interactions. The usage of these models, which came from the analysis of a vast amount of data, will expand and develop as the technology and software are used to build these interactions and show the complexity of the relationships between data variables (Whittle 2010). According to Reese and Bhatia (2018), the model presents the relationship between an input layer with nodes, a hidden layer in the middle, an output layer node and an output target variable. Figure 4 shows the graphical representation of the MLM, which is the neural network technique (Najah Ahmed et al. 2019). In fact, Krose and Smagt (2011) stated that an ANN involves a pool of units that represents a simple configuration in which the units connect by signals sent from one unit to another over a wide range of weighted networks.

Adding to that, Tü (1996) mentioned some advantages of using the neural network method, which include less formal statistical training, discovering complex nonlinear interactions between dependent and independent variables, detecting all possible interactions between response variables, and the fact that there are several training algorithms to perform it. On the other hand, there are some disadvantages of using this method which can be: its black box nature, which is related to the way of interpreting the model as an input-output system while ignoring the internal interactions, a higher software cost, a proclivity for generalisation and the practical implementation of the model. Nevertheless, Krogh (2008) generally mentioned the applications of this method, which can be used to predict the protein's structure, classify cancers and predict genes. For instance, Maier and Dandy (1996) used ANNs to forecast the parameters of the quality and salinity of river water in South Australia. The main results were that high levels of salinity can cause great losses to users in Australia and cost about \$US22 million per year.

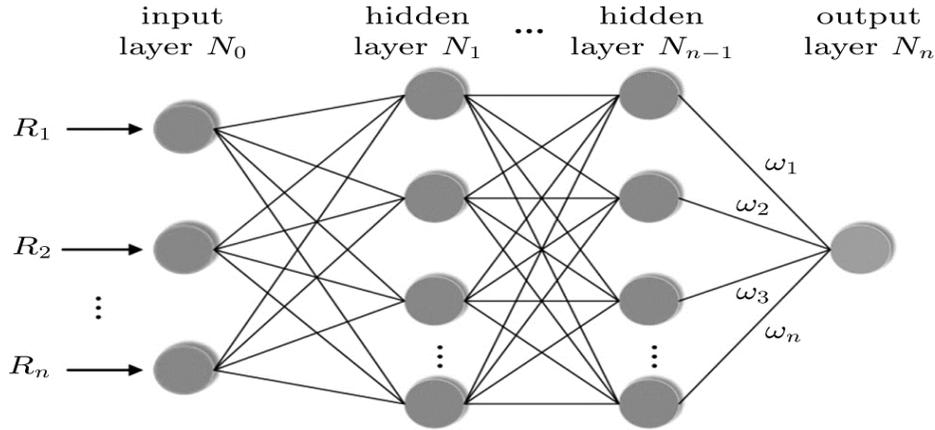


Figure 4: Graphical representation of Neural Network Method (Cao, Sai, Fu, and Duan 2020)

Here,  $R_n$  is the input variable in order n,  $N_0$  is the first input layer,  $N_1$  is the first hidden layer,  $N_{n-1}$  is the following hidden layer of order n-1,  $N_n$  is the output layer and  $w_n$  is the weight of the equation in order n.

Hocking and Hocking (2013) further explained the methodology of implementing neural networks. Once the data is split and trained, the ANN provides the target values in relation to any input pattern observed. Hocking and Hocking (2013) insisted that ANN data validation be considered before questioning the results obtained. In addition, the learning process is described as the process that identifies the weight values of the neurons. The learning process is called ‘supervised’ if the input and output data exist. However, it is called ‘unsupervised’ if only the input data is available. Further, the back-propagation algorithm is used for multi-layer neural networks. This process consists of a forward pass, which means computing the output of each neuron on the training set, and a backward pass, which are the weights assigned to each neuron depending on the

error value obtained by subtracting the actual and predicted output values. This process is stopped when the value of the error is lower than the threshold value.

**5- Structural equation modelling:** Pugesek, Tomer and Eye (2003) mentioned that this technique is used to analyse and evaluate models that exhibit relationships between variables. It is a statistical methodology that represents several cause-effect relationships that have been set as a hypothesis between variables. These relationships are proposed based on some phenomenon to pre-describe the direct effect and the indirect effect of the (observed or not) independent variables have on the (observed or not) dependent variables. In addition, Byrne (2011) showed that the structural equation modelling method is based on two steps representing the causal process by structural equations using regression and later modelling these equations pictorially to relate to the concept of the theory under study. On the other hand, Jitesh (2021) showed that the structural equation modelling approach helps analytics test the theoretical models and validates them. It is used to understand the nature of the relationship between observed values and ensures a stepwise understanding of applications with software programs such as SPSS and R software. Motawa and Oladokun (2015) used this approach to analyse different variable interactions to demonstrate the trends in energy consumption and to study carbon emissions. It was concluded that, using structural equation modelling, there are many factors that affect how much energy a home uses and how much carbon it emits, which includes the floor area, energy efficiency, the number of people in the house, the income of the household, the age of the house, consumption patterns and the age of the homeowners.

**6- Nonlinear regression modelling:** as per Rhinehart (2016), valuable models are usually nonlinear in their influence on the inputs. Moreover, variables must be transformed before performing the analysis. As per Hadi and Chatterjee (2012), transformation is done to ensure the

linearity, ensure the normality in the dataset and stabilise the variance. If linear regression analysis is performed, it is often performed on the transformed variables to get a better regression fit. However, Hocking and Hocking (2013) stated that, for multiple regression, some input variables may require transformation. In contrast, it is not necessary to be performed for all variables. In addition, this step needs careful effort and clear application. In fact, Hadi and Chatterjee (2012) mentioned that it is essential to carry on with this step if there is a clear indication that the regression model violates the regression assumptions. An example of these assumptions is the normality plot and the residuals variance plot. For instance, an application of this method was performed and explained by Mohseni, Stefan and Erickson (1998) to develop nonlinear regression models that aim to measure the weekly stream temperatures required for fish habitat assessment in the USA throughout an annual cycle. Accordingly, the regression models accurately predicted the weekly stream temperatures for 573 stream gauging stations. Two models were developed, one for the warming season and the other one for the cooling season. As a result, the models were considered to be effectively suitable (99% confidence) for approximately 89% of the stream gauging stations.

The following models are considered linear models as the parameters were presented linearly, such as  $\beta_0$  and  $\beta_1$  (even if the input variables are nonlinear).

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (4)$$

$$Y = \beta_0 + \beta_1 X^2 + \varepsilon \quad (5)$$

$$Y = \beta_0 + \beta_1 \log X + \varepsilon \quad (6)$$

$$Y = \beta_0 + \beta_1 \sqrt{X} + \varepsilon \quad (7)$$

However, the model

$$Y = \beta_0 + e^{\beta_1 X} + \varepsilon \quad (8)$$

is considered a nonlinear model because  $\beta_1$  is not entered linearly in the model.

Note that  $Y$  is the output variable,  $\beta_0$  is the intercept of the equation,  $\beta_1$  is the slope of the equation,  $X$  is the input variable and  $\varepsilon$  is the random error value.

As a result, the data of this model must be transformed to satisfy the assumptions and get a better fit from the results. Another reason mentioned by Hadi and Chatterjee (2012) for the transformation of variables is the probability distribution of the mean. If the value of the input  $X$  changes with  $Y$  and the mean of  $X$  is related, the variance of  $Y$  concerning  $X$  will be affected. Usually, the distribution of  $Y$  is not normal in such cases and violates the significance of the model. Sometimes some large samples can be assumed to be normal as they show a normal bell-shaped distribution. Finally, there is no more apparent reason to perform transformation than to perform a linear regression on the data as a first step and test the normality and residuals of the model. Table 1 shows the linear regression functions and their transformations.

Table 1: Linear regression functions and their transformations (Hadi & Chatterjee 2012)

Function	Transformation	Linear Form	Graph
$Y = \alpha X^\beta$ (9)	$\hat{Y} = \log Y$ (10) $\hat{X} = \log X$ (11)	$\hat{Y} = \log \alpha + \beta \hat{X}$ (12)	Figure 5
$Y = \alpha e^{\beta X}$ (13)	$\hat{Y} = \ln Y$ (14)	$\hat{Y} = \ln \alpha + \beta X$ (15)	Figure 6
$Y = \alpha + \beta \log X$ (16)	$\hat{X} = \log X$ (17)	$Y = \alpha + \beta \hat{X}$ (18)	Figure 7

$Y = \frac{X}{\alpha X - \beta}$ (19)	$\hat{Y} = \frac{1}{Y}$ (20) $\hat{X} = \frac{1}{X}$ (21)	$\hat{Y} = \alpha - \beta \hat{X}$ (22)	Figure 8 (a)
$Y = \frac{e^{\alpha+\beta X}}{1 + e^{\alpha+\beta X}}$ (23)	$\hat{Y} = \ln \frac{Y}{1-Y}$ (24)	$\hat{Y} = \alpha + \beta X$ (25)	Figure 8 (b)

The following figures 5 to 8 show the graph of the linear functions, respectively.

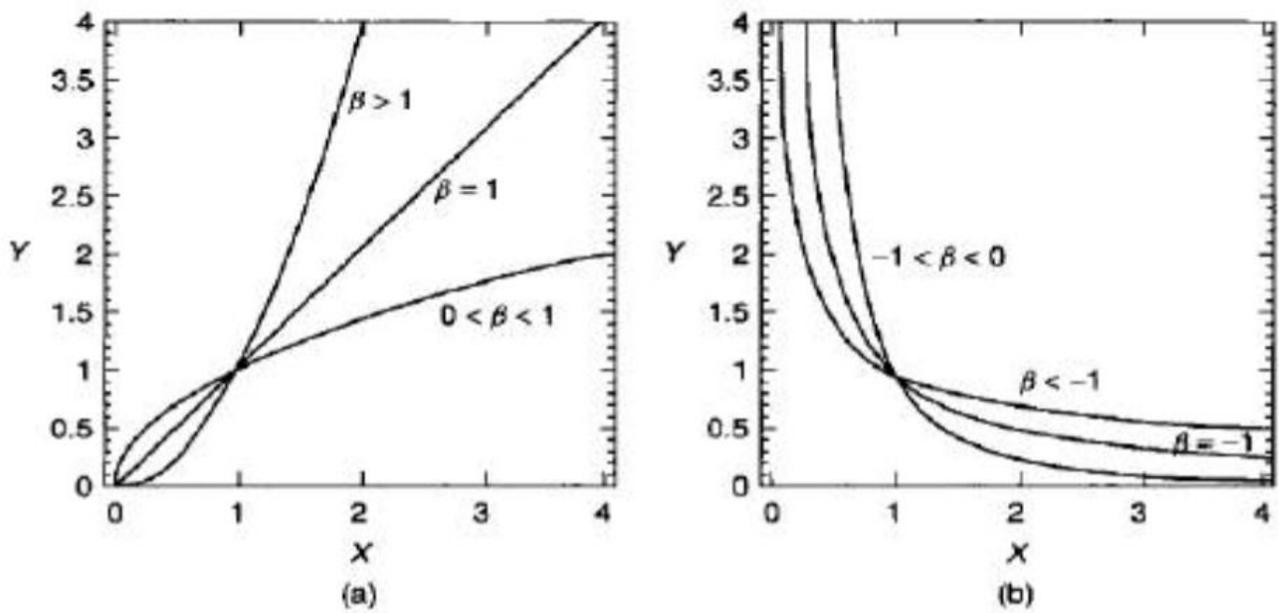


Figure 5: Graphs of the linear regression function  $Y = \alpha X^\beta$

(Hadi & Chatterjee 2012)

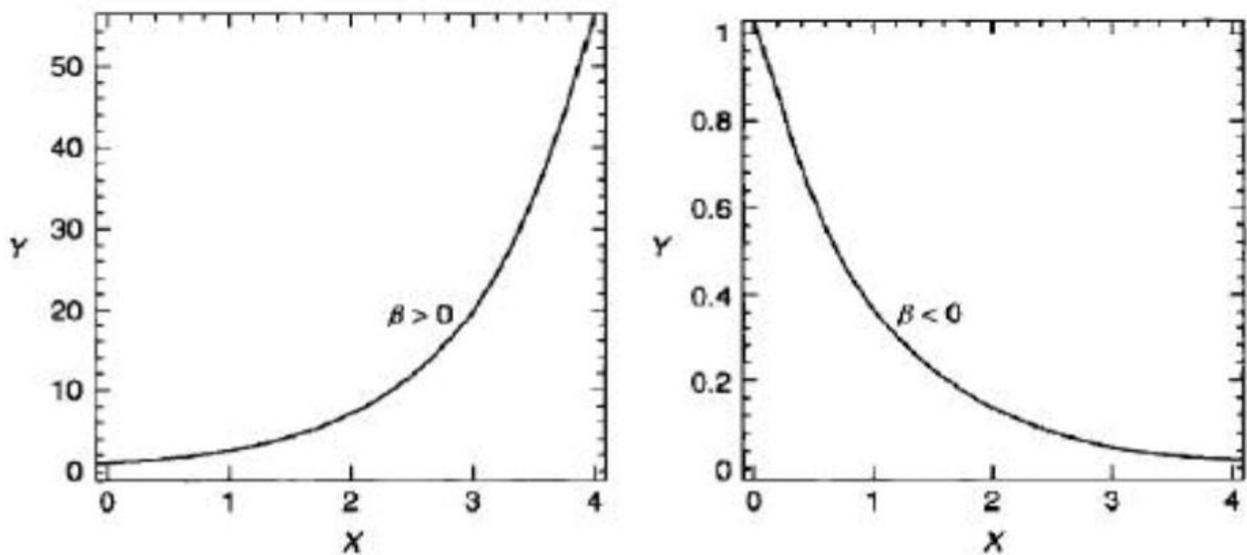


Figure 6: Graphs of the linear regression function  $Y = \alpha e^{\beta X}$

(Hadi & Chatterjee 2012)

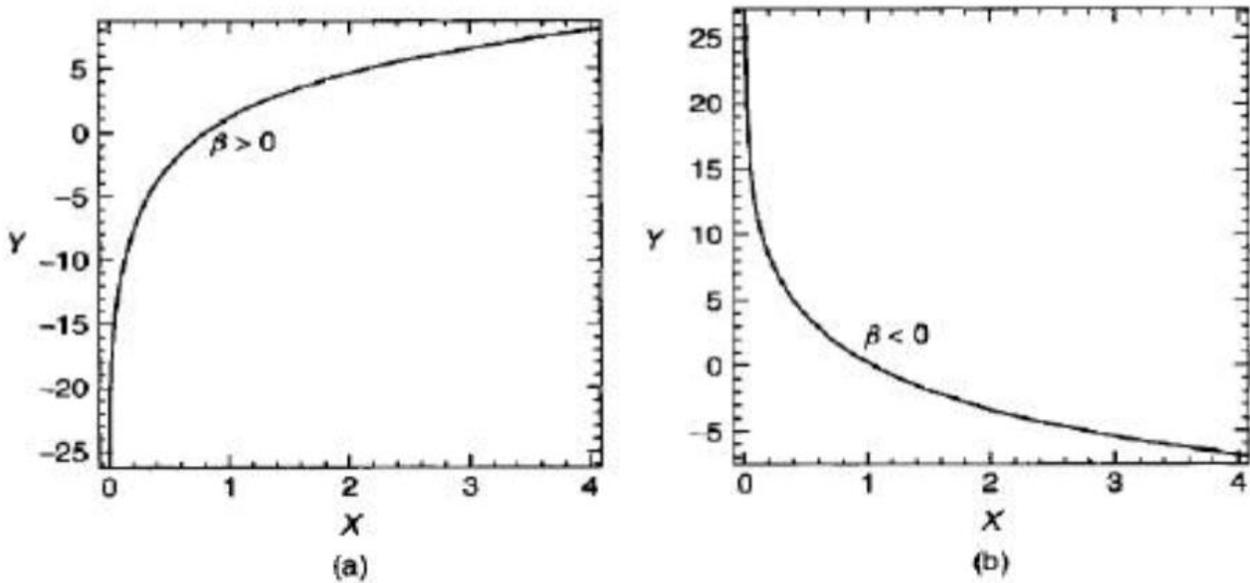


Figure 7: Graphs of the linear regression function  $Y = \alpha + \beta \log X$

(Hadi & Chatterjee 2012)

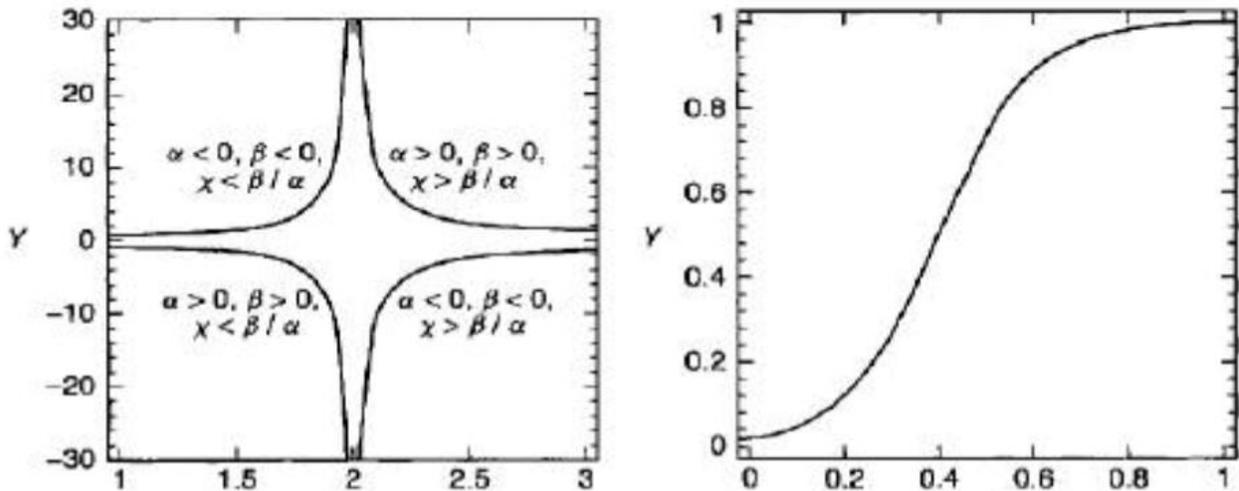


Figure 8: Graphs of the linear regression function (a)  $Y = \frac{X}{\alpha X - \beta}$  and (b)  $Y = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}}$

(Hadi & Chatterjee 2012)

Rönkkö et al. (2022) stated, after viewing 20 articles that used regression as a method of analysis, that around 66% of these studies have applied transformation at least once. The most common approaches used were Poisson and negative binomial models (see Figure 9 for the comparison of a linear regression model and a Poisson model). In fact, these analyses were followed by probit and logit transformations, which are usually used in logistic regression and some categorical data models. In addition, for U shape variable effects, the power transformations were used. The most common way to apply a transformation of data is to use GLM, which stands for a generalised linear model and which is based on a function that connects a curve by the output variable's mean as a function of the independent input variable. It's important to note that the distribution is considered a conditional distribution for a group of independent variables and an unconditional distribution for the dependent variables.

The GLM technique can be used to generalise the variance analysis using log-likelihoods (Pugesek, Tomer & Eye 2003). As per McCullagh and Nelder (2019), GLM is related to four types of distribution, which are: the normal distribution of data, binomial distribution such as probit analysis, Poisson distribution (contingency tables) and gamma distribution (variance components).

As shown in Figure 8 (Rönkkö et al. 2022), the first plot shows how the data points around the regression line are in a normal distribution. This is called conditional distribution for all the points that have the same value of X. On the other hand, the distribution of Y is unconditional and might not be considered normally distributed. The second part of Figure 9 is the Poisson model, which shows a conditional distribution of Y for any value of X. However, the total distribution of Y doesn't need to be Poisson. This example illustrates that choosing a specific GLM distribution cannot be specified by looking at unconditional distributions.

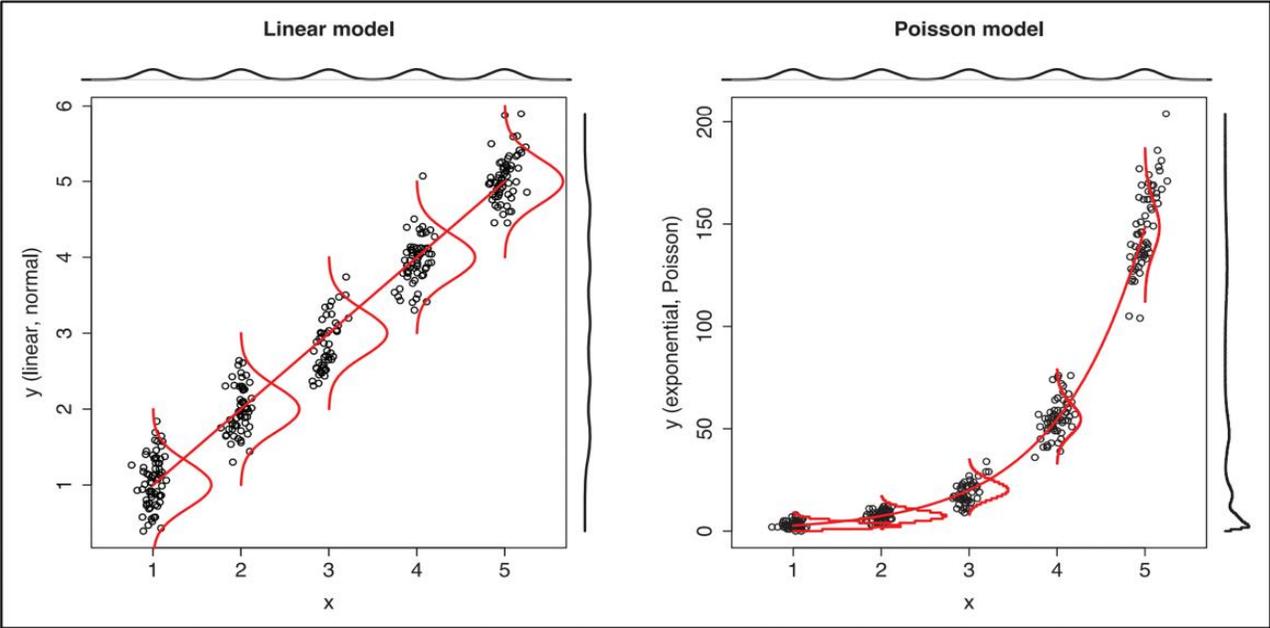


Figure 9: Linear model and Poisson model comparison

(Rönkkö et al. 2022)

Pugesek, Tomer and Eye (2003) studied the GLM, especially the generalisation of the correlation coefficients of a general linear regression model. This study showed the comparison of four types of measurements, which are bias, mean squared error and the presence of

overparameterisation. One of the main results was that the cross-validation estimator had a high negative bias and significant mean squared error.

## **2.4 Checking the Prediction Model's Accuracy**

This section demonstrates the various tests and terms that will be used in this study to check the accuracy of the prediction models. The following are the terminologies and specifications:

**1- Mean square error (MSE):** according to Campora, Cravero and Zaccone (2018), the ANN performance is often calculated through the MSE, which is the difference between the actual and predicted values. Moreover, Thompson (1990) insisted that MSE is a valuable statistical tool used for comparing and forecasting accuracy. However, Thompson (1990) proposed a ratio called log mean squared error ratio, which is mainly designed to cover the shortages of MSE in evaluating the overall accuracy of a model through many series. Marmolin (1986) discussed MSE as a criterion for checking quality and illustrated that MSE is not a perfect measure for checking the accuracy of visual systems. For human observation, the picture quality of an optical system can be checked by reviewing the error concerning the assumed parameters. Moreover, none of the tests or measures (other than MSE) was a valuable tool for checking the accuracy of all pictures of the visual systems. Thus, the properties of pictures were subjected to error measures instead.

**2- The root mean square error (RMSE):** this is the residual's standard deviation or what is called prediction error. Accordingly, the residuals are defined as the extent to which the data points are far from the regression line. In addition, it shows how the data are concentrated around the line for better fitting (Chai & Draxler 2014). RMSE is considered a standard test used to assess models, especially models of meteorology, quality of air and in climate research and studies. In view of this, Chai and Draxler (2014) discussed the advantages of using RMSE, compared it with the MSE

test and emphasised that both tests are usually used in model accuracy tests. It was concluded that RMSE is not a decent accuracy tool for checking the overall performance of a model. Therefore, MSE is a better measure for such a purpose.

Conversely, Willmott and Matsuura (2005) confirmed that RMSE is a widely used measure for climate and environmental predictions. However, Willmott and Matsuura (2005) insisted that RMSE can't be used for testing the accuracy of an average performance error as it might be misinterpreted. Nevertheless, Chai and Draxler (2014) suggested that avoiding RMSE and favouring MSE is not a key solution. Additionally, many research papers favour MAD (mean absolute deviation) over RMSE when evaluating the statistics of a model. However, Chai and Draxler (2014) mentioned that there are some cases that must be well classified to know when to use each test. For instance, RMSE is more applicable to be used when the data distribution is expected to follow Gaussian distribution.

Nonetheless, a group of tests is required to evaluate the validity of model performance. Additionally, the Chai and Draxler (2014) study doesn't show that RMSE is a better measurement compared to MAD. On the other hand, Willmott and Matsuura (2005) explained that RMSE is a function of three types of error sets instead of one only (as an average error function). Moreover, RMSE depends on the variation distribution of error magnitudes and the square root of errors, which is like MAD test.

**3- Mean absolute deviation (MAD):** According to Konno and Koshizuka (2005), the accuracy test (MAD) was firstly introduced in 1990 to solve problems related to large-scale data optimisation. In addition, Konno and Koshizuka (2005) concluded that using MAD calculations provides many advantages to the calculations as it is very compatible with the rules of real decision-making. Similarly, Khair et al. (2017) discussed optimisation and prediction models in relation to

the low percentage error. It was shown that using MAD and MAPE for error and accuracy calculations in the least square method resulted in a lower percentage error of 9.77% as the technique was applied for time series data. Further, Willmott and Matsuura (2005) concluded that MAD is a more accurate average error test, which is in contrast to with RMSE. Generally, comparisons of the average performance of a model's error must be using MAD. Berend and Kontorovich (2013) mentioned that MAD generally behaves as the standard deviation; its weaknesses are confined to the endpoints as the test sometimes fails to converge. As a solution, Berend and Kontorovich (2013) provided an optimal estimation for the tail points regarding the total variation or the distance between the actual and predicted values over a realistic range.

**4- Symmetric mean absolute percent error (MAPE):** as per Tayman and Swanson (1999), MAPE is a statistical measure that is widely used to calculate the accuracy of a prediction model. It is extensively used for evaluations of population forecasts. In contrast, Goodwin and Lawton (1999) explained that MAPE shouldn't be used to forecast the accuracy of a model, as it considers the errors above the actual values different from the values below. In addition, a suggestion for this problem was to use the symmetric MAPE. This test usually expresses the forecast error as a ratio of actual and predicted values.

Tayman and Swanson (1999) illustrated that MAPE sometimes overstates the forecast error of a population, which affects the validity of the test. The alternative tests used in the last-mentioned study are the symmetrical MAPE and the M-estimators measure. Some experimental evaluations imply that M-estimators are more accurate measures, as these simply do not overemphasise the forecast error as the MAPE test does. Another suggestion was to focus on the nonlinear transformations before studying the distribution, to minimise the error.

**5- Residual standard error (RSE):** this test was illustrated by Chinn (2000) and confirmed that RSE is used in the regression models as a measure for the residuals' standard deviation. It is also used to see how well the expected and actual data points match up. As per Zhang and Chow (2010), this test can provide entrepreneurs with information and facts about the difference between the projected costs and enable them to compare them with the actual costs. In addition, it can give insights into the extent of variation between the projected costs and the historical costs.

## **2.5 Prediction of the Power Production Using Regression Analysis**

Tüfekci (2014) stated that there are many other regression methods used to build models that vary with accuracy and error, such as radial basis function neural network, MLP, pace regression, support vector poly kernel regression and additive regression. In addition, Thatcher (2007) discussed the effect of climate change in Australia on electricity consumption. It was demonstrated that energy consumption differs among regions and states in Australia due to the change in the climate and other essential factors, such as social and economic patterns. These patterns vary in a seasonal manner, which highlights how important ambient weather conditions are in the electricity industry (either after or before production). In addition, it was found that, during summer, electricity production units increase their production by 0.14% and, hence, models of change in electricity demand were considered valuable tools for planning in Australia.

The Spanish banking sector's energy consumption was discussed by Tso and Yau (2007). It was shown that climate change plays a special role in electricity demand where a regression model was developed for the data from 55 banks. Two models were tested, one for high winter climate severity and the other for low winter climate severity. Those models were efficient for predicting energy consumption for the bank branches.

As per Tü and Gürgen (2014), the increasing demand for electricity is cohesively related to environmental and economic concerns. In addition, power plant systems are established worldwide, consisting of gas turbines that are tested for their reliability and sustainability, especially when they are subjected to full loads and liabilities. Tü and Gürgen (2014) studied the effect of ambient parameters such as AT, ambient pressure, RH and V. Their study showed a clear effect of AT on the performance of gas turbines, and of V on the performance of the steam turbine. Another study, by Kesgin and Heperkan (2005) was a relative study that utilised the ambient parameters, used MLMs such as ANNs to predict the behaviour of gas turbines and studied the target power production variable. This study differs from Tüfekci (2014) by the nature of the dataset, the type of MLMs used and the collection of datasets over a longer period for middle-load gas and steam turbines. Moreover, the scale of power production was different. Therefore, this paper aims to not only analyse the individual parameters but also to find the most appropriate method for utilising the ambient parameters and predicting power production in the UAE. This research is different in the way it focuses on the linear regression method to analyse the dataset based on UAE weather and the ambient variables dataset recorded from a CCPP. In addition, an ANN analysis will be carried out to study the dataset and the results compared to previous research.

Further, Dutta and Ghosh (2021) analysed the P of a CCPP based on ambient parameters. The main findings were the factors such as AT, RH, V and ambient pressure could clearly affect the performance of power plants, and the recognition of this helps to improve the yield per hour. In addition, it can help optimise the utilisation of fuel used for production. Thus, better cost management for better efficiency. Moreover, the prediction model was based on a dataset gathered over six years, from 2006 to 2011, while the plant was working based on a full-load performance. The correlation of each variable with the power production was studied and the main findings were

that the AT is inversely proportional to power production, with a strong correlation of -0.948. In addition, RH is directly proportional to power production with a medium correlation of 0.39; ambient pressure is directly proportional to power with a correlation of 0.518; and there is an inversely proportional relationship between V and power production, with a correlation of -0.87. Finally, a Dutta and Ghosh (2021) study proposed further research to predict the power production of different types of plants.

## 2.6 Power Plants' Production Concepts

Raja, Srivastava and Dwivedi (2006) explained the concepts of power plants and illustrated that these plants consist of systems or subsystems that produce the electricity or power required for public or economic demands. The author emphasised that these power plants must be environmentally friendly and anticipated a shift in the direction of energy production from conventional to non-conventional power production by the year 2050. The non-conventional power production is more efficient, environmentally friendly and favourable to society. Figure 10 shows the classification of power plants as conventional and non-conventional.

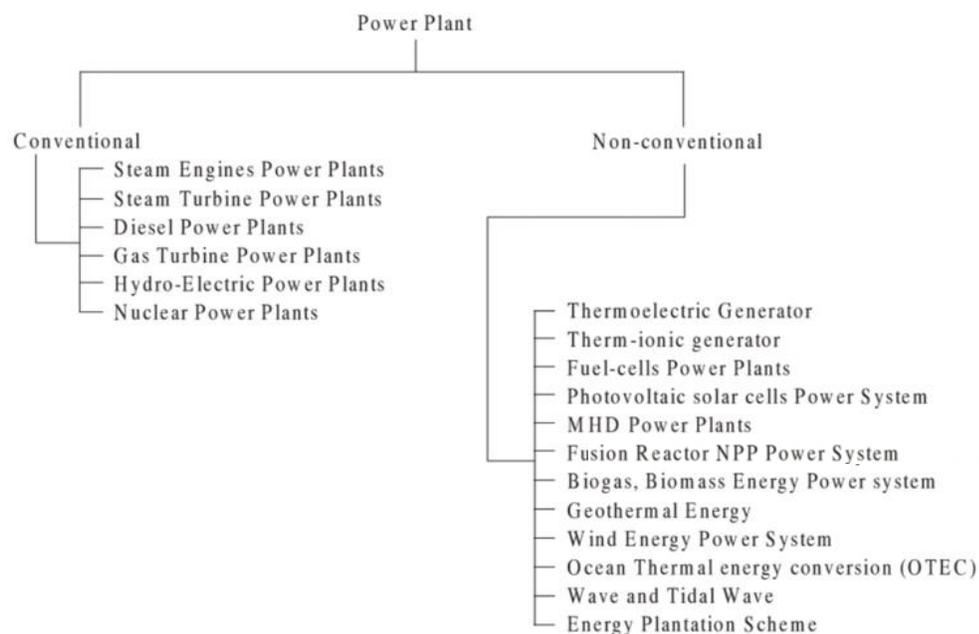


Figure 10: Conventional and non-conventional power plants

(Raja, Srivastava & Dwivedi 2006)

A study by Kotowicz and Brzęczek (2018) represents a comprehensive analysis of attempts to increase the electrical efficiency and performance of gas turbines installed in a CCPP and enable them to produce almost 200MW per day. The main parameters were analysed, based on open and closed air-cooling conditions, consequent combustion and steam cooling. It was concluded that, in gas turbines, the high metal blade temperatures influence the electrical efficiency. Another study (Ibrahim et al. 2017) mentioned a key parameter for increased efficiency and optimum electrical production: the gas turbine inlet temperature. Moreover, as per Kotowicz and Brzęczek (2018), using consequent combustion and steam cooling can increase the efficiency of the production cycle to between 0.63 and 0.65.

A power plant concept is based on delivering a flow of energies such as mechanical and electrical energy. The main machine in the power plant group of systems is the generator, which is coupled to a primary mover to run it and generate the electricity (Raja, Srivastava & Dwivedi 2006).

## **2.7 Types of Power Plants**

This section shows some other types of power plants used to produce power and divided based on economic demand. As per Kaplan (2009), the demand for electricity and load determines which type of power plant to build. These types include:

- nuclear power plants, coal power plants and geothermal baseload units, which can run throughout the year, except if there is forced maintenance, with low fuel costs and, hence, fewer variable costs. However, they are expensive to build. In fact, Petridis and Nicolau (2011) stated that nuclear power production systems are directed towards not only producing electricity but also applications such as producing hydrogen and coal gasification. Moreover, in the last few decades the efforts of nuclear production have mainly focused on obtaining high-system efficiency, usage of nuclear process heat for some applications such as hydrogen production, adherence to safety, reaching better burn-up rates of fuel, length of lifetime and non-proliferation resistance.
- CCPPs, which are very efficient in power production. However, they require expensive fuel such as natural gas and need maintenance many times a year. The power production of these plants meets an intermediary load.
- peaking plants, which use combustion turbines to operate, are inefficient and use highly expensive natural gas. These plants are utilised only to fulfil high loads if needed.
- other types of power plants, such as renewable power plants, which are based on wind energy and solar power. These power plants fall outside of this economic category. They are used to shift production from high-variable-cost gas systems and peaking units to lower-variable-cost wind and solar units. However, this renewable power production can displace the coal generation only if the demand is low, such as during evenings, strong winds and during weekends. In fact, a disadvantage of these systems is the difficulty of meeting loads regularly; they are mainly dependent on weather conditions. In addition, if wind and solar

plants are generating electricity that needs to be stored, the options of storage, from the current technological perspective, are few and expensive (Kaplan 2009).

Figure 11 shows the total power generation classified by energy source in the US in 2007. As shown, the highest percentage of electrical energy production came from coal combustion – 49%.

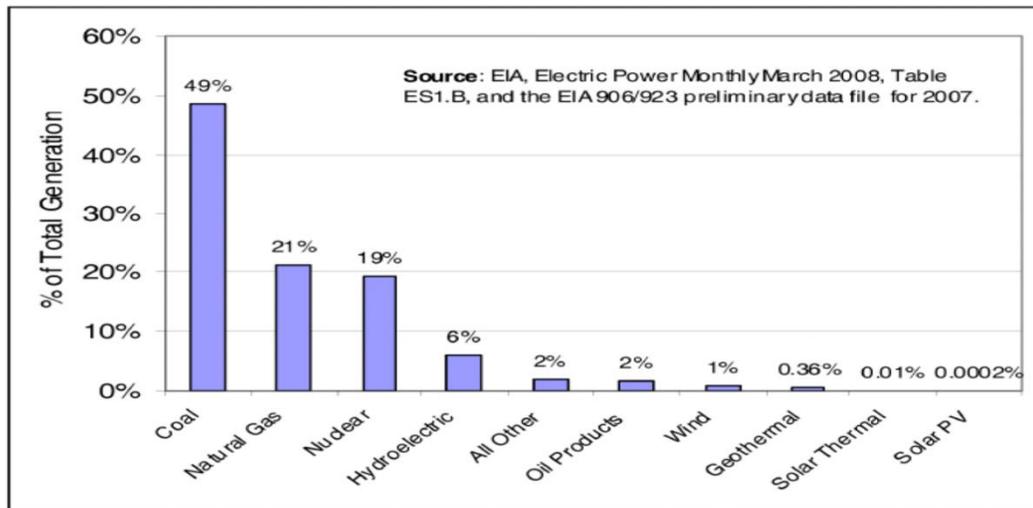


Figure 11: Total power generation classified by energy source in US in 2007

(Kaplan 2009)

There are various types of energies that contribute to producing mechanical energy. As stated by Raja, Srivastava and Dwivedi (2006), there are different types of energy including nuclear, thermal, electrical, radiant and chemical energy. Nuclear energy is based on nuclear fission, where heat is produced to evaporate steam that moves the generator to produce electricity. It is the world's best emission-free source of energy. In addition, this type of energy can be performed using fusion. However, all existing power plants use fission as fusion can't be controlled. Other types of energy such as thermal energy are considered to be a combination of kinetic and potential energy, where the atoms are moving randomly due to heat. Each day, a

significant amount of heat is stored in the oceans, equivalent to 250 billion barrels of oil; this system is called the ocean thermal energy conversion system.

The power unit is the joule, which is work per second or energy per time. This means that energy is required to generate power, and, hence, energy is used to operate power plants and produce electricity. The units of power are watts, horsepower and joules per second. The conversion of these units is as follows:

$$1 \text{ watt} = 1 \text{ joule per second} \quad (26)$$

$$1000 \text{ kilowatts} = 1 \text{ horsepower} \quad (27)$$

## **2.8 An Overview of a Combined Cycle Power Plant**

According to Kehlhofer et al. (2009), a CCPP, as a definition, is a plant composed of two combined thermal cycles that produces higher efficiency. The so-called ‘topping cycle’ is the higher-temperature cycle. The heat produced as waste from that cycle is used for the second bottoming cycle of lower temperature, where those two cycles are coupled in a heat exchanger. According to Eckardt (2014), a CCPP consists of a group of high-temperature engines that convert heat to mechanical energy. The term ‘combined cycle’ refers to collecting and utilising the gas turbine’s waste heat to produce steam, which increases the efficiency of production and produces more electricity. The compressed air is injected into the gas turbine and burns with a high-temperature fuel. This mixture of hot air and fuel flows in the blades of the gas turbine, making it spin to run the electrical generator and produce electricity. The heat is transformed into mechanical work, which is used to generate electricity. The waste heat from the gas turbine exhaust enters the HRSG. It is then used to generate steam, which is used to operate the steam turbine and move the output shaft, resulting in mechanical work that is later converted to electricity. Figure 12 shows a

simplified flow diagram of a combined cycle. In addition, as per Franco and Casarosa (2002), the overall efficiency of the CCPP range is 50% to 60%.

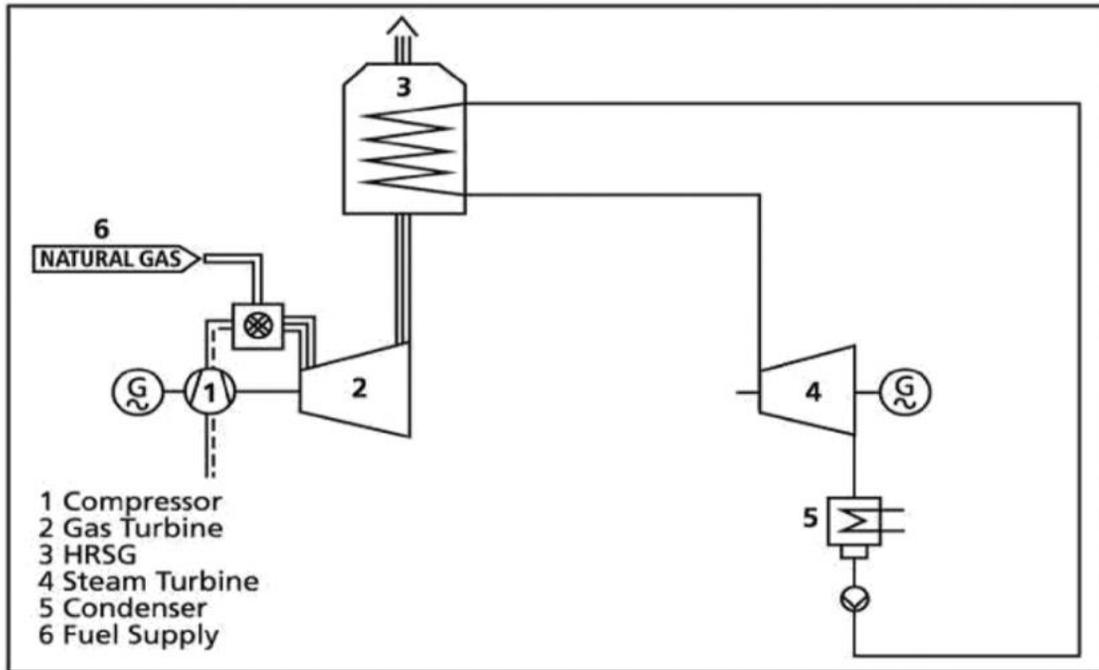


Figure 12: A simple flow diagram of a combined cycle

(Kehlhofer et al. 2009)

There are several strategies for monitoring and increasing the efficiency of the CCPP, other than studying the ambient weather conditions, such as the following:

- Kaviri et al. (2013) explained the advantages of using CCPP for power production for its high efficiency and low emissions. In addition, this study focused on CCPP with supplementary firing. It concluded that the temperature of the the HRSG inlet gas influences the efficiency of the steam cycle. The research showed that increasing this temperature to 650°C; increases the cycle's thermal efficiency and energy efficiency (which is the efficiency of the amount of energy produced by the system when it is in thermodynamic

equilibrium). However, after some time, the two types of efficiencies start decreasing due to other factors.

- Franco and Casarosa (2002) showed that the efficiency of a CCPP might increase to more than 60% without the attainment of new technology for the gas turbine. This can be done by optimising the HRSG, by using parallel sections and limiting the conditions to up to 220 bars. In addition, optimisation of HRSG by reheating gas turbines (the process of post-combustion) and recovering gas might lead to an efficiency of up to 65%.
- another efficiency optimisation method of HRSG was proposed by Kehlhofer et al. (2009), which explores the applications of the influence coefficients utilising the Newton-Raphson approach. It concentrated on the cycle design factors to achieve improved efficiency. The major aim of this study was to demonstrate the importance of design factors and their impact on cycle performance. An optimisation of the allocation of the boiler's area and its various components was presented as an example.
- Kotowicz, Job and Brzeczek (2015) mentioned methods used to increase the efficiency of the CCPPs and concluded that the efficiency for the gas turbine, as an example, would be higher when focusing on the aspects of the gas turbine and improving it by efficient heat utilisation of the turbine's cooling air. The methodology and calculations presented an extreme range of temperatures and compression ratios, assuming a constant value of the turbine's outlet gas temperature. Moreover, it was possible to improve the gas turbine's efficiency while keeping the heat value constant, by focusing on the aspects of the gas turbine and improving it. In addition, this study proposed an additional steam cycle that allowed for an increase in efficiency by 2 to 3%. However, this higher efficiency led to an

increase in the compression ratio. In addition, an economic study was proposed, which showed that the gas turbine's utilisation development could be made easier by keeping a constant and reasonable low cost of the gas turbine's financial investments.

To finalise, Ibrahim et al. (2017) illustrated that to identify the characteristics of a CCPP and locate all the optimum conditions for optimum operation, the simulation model unique tool was irreplaceable. The importance of using such models lies in the difficulty of implementing ideas, due to the size of the system and the difficulty of experimental examinations.

## **2.9 Power Generation in UAE**

An independent country such as the UAE focuses on its development plan mainly on the power sector and production efficiency. Sushil Jha and Tandon (2019) mentioned that the country's wealth depends primarily on income from petrol, and ever since its formation, the country has focused on developing the infrastructure of power generation. As per Ho et al. (2019), the UAE is one of a few countries focusing on nuclear energy as a source of energy. Ibrahim et al. (2017) stated that valuable improvements for plant efficiency and lower pollutants are crucial concerns for any design or any type of power generation. Accordingly, Sushil Jha and Tandon (2019) explained the reason for using nuclear power plants for power generation: the country's economic performance varies according to the volatility of oil prices, which is the primary input to the systems of CCPPs. Thus, the dependence on oil must be reduced by the robust development of UAE infrastructure. Additionally, another view was presented by Saghafifar and Gadalla (2016), which mainly focused on the replacement of natural gas and fossil fuel in power plants. Solar energy was promoted as a potential solution that takes advantage of the sunny climate in the UAE to activate solar-based power plants. The study explained the benefits of using this kind of power production in the UAE

as it is environmentally friendly and a technology-based solution. In particular, a case study of a new hybrid solar energy-based CCPP with a capacity of 50MW was presented by Saghafifar and Gadalla (2016), who concluded that hybridisation of an existing power plant enhances the utilisation and cost of electrical power in the UAE, and increases the savings, the net present value and the payback period for the plant. The study showed that a hybrid CCPP in the UAE could present a net present value of \$34.918 per MWh and a cost of electricity of \$77.7 per MWh. Moreover, the solar share is 8.87% and the specific CO<sub>2</sub> emissions are 371.9 kg per MWh.

Many countries rely on the CCPP system to generate electricity, such as those in the UAE. However, Talukder and Soori (2015) discussed the disadvantages of having this type of power plant in the UAE's arid climate and hot summer temperatures. The study concluded that the main losses could be due to hot weather, as the amount of gas turbine power production might decline with increasing ATs. To overcome this, Ho et al. (2019) outlined the advantages of constructing nuclear power stations in the UAE, such as the Barakah in Abu Dhabi, which involves supplementing baseload clean-power production. Early growth of nuclear power production started around the world in the 1950s, and many designs were presented.

Consequently, the dominant design was the pressurised water reactor, as it was very compact and economic. Around the next ten years, other designs might be functioning, such as the small modular reactor, which can produce up to 300MW. However, Al Rashdi et al. (2020) illustrated that the contamination of rare earth elements is harmful to the environment. Moreover, the concentrations of these elements were measured around the area of the nuclear power plant in Barakah, Abu Dhabi. Generally, the findings revealed that the area is normally free of any hazardous amounts. Figure 13 shows the nuclear power plant's net capacity under construction (Ho et al. 2019). There are 454 nuclear power reactors (four in the UAE – their construction is led by

China) around the world, corresponding to 10% of electricity production. As shown by the figure, UAE is third ranked compared to other countries, indicating the fast development of nuclear power production.

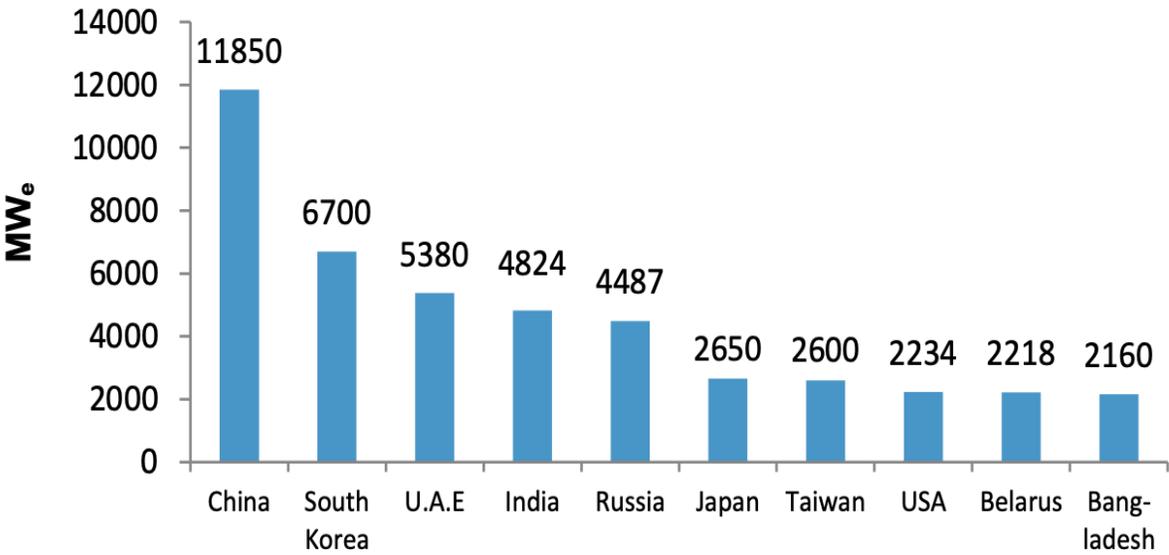


Figure 13: The nuclear power plant’s net capacity under construction

(Ho et al. 2019)

Talukder and Soori (2015) discussed the disadvantages of CCPP with supplementary firing, which can increase the overall power generation: it can result in low thermal efficiency. In contrast, the low thermal efficiency resulted in an industrial direction towards solar hybridisation in the UAE’s CCPP and in other countries. Inefficiency can be reduced by reducing carbon dioxide emissions. In addition, this new direction presented by Talukder and Soori (2015) towards a new plant production configuration is an integration of solar power and CCPPs. This technology optimises the usage of renewable energy. Generally, the study showed that in the UAE, high overall efficiency comes from the integrated solar CCPPs. Generally, (Ho et al. 2019) stated that, over the next few decades, the advanced nuclear power plants in the form of sodium-cooled and molten salt-

cooled nuclear plants are expected to provide high-efficiency P along with the high-temperature gas-cooled plant reactors. Moreover, these types of plant configurations can be used for heavy industry and hydrogen generation for the synthetic variety of fuels.

The technical and economic characteristics of the cooling system's input air in a gas turbine inside a CCPP were examined in a study by Barigozzi et al. (2015) comparing three plants in different countries: Phoenix in the USA, New Orleans in the USA and a power plant in Abu Dhabi (UAE). The system that was analysed used chilled water, which basically cools down the system during very hot or sunny days. The study concluded that modelling the power plant and sizing the amount of inlet air in the cooling system depends on the weather conditions, such as whether there is a hot climate. The plant operation hours and power production were higher in hot weather areas. In addition, wet climate conditions require a vast amount of thermal storage, resulting in higher investment costs. Further, the best performance recorded was for areas with high ATs and low RH (deserts).

## CHAPTER III

### RESEARCH METHODOLOGY

This chapter describes the research data, interactions between variables and how the data is quantitatively tested in the CCP system to produce electrical power. Firstly, the dataset contains 14,400 data points collected and recorded in the datasheet (each parameter has 2,881 points). This dataset would fit into a linear regression model to predict the baseload power production. Another method is used to analyse the dataset, which is ANN. The two methods will be compared and discussed.

#### 3.1 Linear Regression Modelling

The variables were obtained from the power plant after gaining formal consent, so there is no need for a walk-through audit. Independent variables are AP, AT, RH and V. These variables will be analysed firstly using the linear regression method, which is a technique for depicting the relationship between independent variables and the dependent variable (Tso & Yau 2007) as in the following equation below (Aranda et al. 2012):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i + \varepsilon \quad (28)$$

Here  $\varepsilon$  is the random error,  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$  and  $\beta_i$  are the changes in the response variable  $Y$  in terms of a change of one unit in the independent variable  $X_i$ , while all other parameters remain constant. Regression coefficients will be calculated using the least-squares method, a statistical method used to analyse data points and find the best fit by minimising the sum of squares of the residual points in the linear regression model (Kong, Li & Zhang 2019). The predicted value  $\hat{Y}$  linear equation with the predictors is (Aranda et al. 2012):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_i X_i + \varepsilon \quad (29)$$

Variables outside the range of this predicted model might show an extrapolation error if tested using this regression model to obtain the response variable. It should be noted that Microsoft Excel will be used for calculations, along with Minitab and RStudio to show the graphical representations of data.

According to Banhidarah et al. (2020), to find the ambient parameters that affect the power production a null hypothesis is set to see if the independent variables, such as AT, ambient pressure, RH and V ( $X_i$  variables) influence the dependent variable P. So, two types of hypotheses will be tested. The first is the null hypothesis ( $H_0$ ) which states that the independent variables and the P have no relationship. The second is the alternative hypothesis ( $H_1$ ), which claims that there is evidence of a relationship between the independent variables and P. The two hypotheses are as follows (Banhidarah et al. 2020):

$$H_0: \beta_1 X_1, \beta_2 X_2, \dots, \beta_i X_i = 0 \quad (30)$$

$$H_1: \beta_1 X_1, \beta_2 X_2, \dots, \beta_i X_i \neq 0 \quad (31)$$

Sample analysis will be carried out to find the trends using scatter plots, show patterns and by dividing the data into subsamples to explain the power production. Subsequently, regression analysis will be discussed with appropriate calculations to find the best model, perform model validation and to discuss the results and trends to provide recommendations and conclusions.

### **3.2 Artificial Neural Networks**

As per Dehghani Samani (2018), ANN is a method of simulation that mimics the way the human brain thinks and analyses data. It contrasts with mathematical models, which predict models based on mathematics and the laws of physics. According to Said et al. (2020), this method is

commonly used to accurately anticipate the production of electrical power. Meteorological data such as the variables provided by the plant are fluctuating and very unsettled, so this method is found to be more suitable for dealing with such data. Dehghani Samani (2018) explained how this method works; a simple understanding of the human brain is enough, as the human brain learns from the experiences of life and creates interactions between the so-called ‘neurons’. The ANN approach replicates this repeating process by analysing the relationship between many inputs and outputs, regardless of how they interact in the real world. A type of ANN is the MLP network, which is composed of neurons as elements that have transfer functions connected linearly or nonlinearly by a weights matrix. As shown in Figure 4, these neurons are organised in three levels: one input layer, hidden layers and an output layer.

Normalisation and scaling of data to the 0–1 range reduce the error for the data analysis. In addition, the data analysis might be affected by the measurement units used. For example, the change of one unit measurement from inches to metres for height might cause a difference in the results. Generally, expressing the variables using a small unit would lead to a more extensive range and more significant effect for that variable. The data must be normalised to avoid dependence on one choice of units. This includes transforming data to a small or common range such as the (0–1) range. The following equation will be used to normalise dataset values:

$$z_i = \frac{(x_i - \min(x))}{(\max(x) - \min(x))} \text{ (Github 2013) } \quad (32)$$

Here,  $z_i$  is the normalised value of power production,  $x_i$  is the power production value,  $\min(x)$  is the minimum value of power production, and  $\max(x)$  is the maximum value of power production.

As mentioned before, the electrical power production of a CCPP will be studied based on ambient parameters such as: AT, AP, RH and V. So, the equation will show (P) as a function of the rest parameters:

$$P = F(AT, AP, RH, V) \quad (33)$$

As per Said et al. (2020), the function F can be described as a complex function with a nonlinearity trend. In addition, it is hard to express it in analytical form, so ANN is a powerful method used to do so. Moreover, ANN expresses neurons as a single unit ( $x_i$ ) and uses weights ( $w_i$ ) or coefficients to connect neurons with each other; it adds up all of these weights to get a numerical value b (bias), which is then, given the pre-activation equation for function A:

$$A = b + \sum_{i=1}^N w_i x_i \quad (34)$$

Here, (i) is the unit's index and (N) is the number of units connected. After that, the pre-activation function will be converted to a single output (y) of a single neuron by passing through the transfer function (g):

$$y = g(A) \quad (35)$$

There are many types of activation function, including the linear activation function as shown above, and the sigmoid activation function, which is:

$$g(A) = \text{sigm}(A) = \frac{1}{(1 + \exp(-A))} \quad (36)$$

$$g(A) = A \quad (37)$$

During the training process of the ANN method, the weight coefficients ( $w_i$ ) and the bias (b) can be modified for specific goals. For example, subsets can be divided randomly into three independent categories: 70% of the samples are for the package of learning in which the weights

and bias can be adjusted, 15% for the internal validation of the ANN method and the last 15% is for testing the ANN. Note that all these processes and equations are performed using AI and, specifically, a developed software.

After that, the ANN functions by taking the inputs ( $X_{\text{Learning}}$ ) and the randomly selected values of weights and biases from the learning stage to perform the network and to get the output values ( $Y_{\text{Predicted}}$ ) and  $\theta$ , which is the error. Then, these values of ( $Y_{\text{Predicted}}$ ) are compared to the output of the learning stage ( $Y_{\text{Learning}}$ ) by the cost-error equation  $E_{\theta}$  as follows:

$$E_{\theta} = \frac{1}{M} \sum_{i=1}^M (Y_{\text{Predicted},\theta}(i) - Y_{\text{Learning}}(i))^2 \quad (38)$$

Here,  $\theta$  is the number of iterations and  $M$  corresponds to the total number of samples chosen in the learning stage. The calculations of this equation are repeatable to get a minimum error and, hence, allow adjustments on the weights and biases. Following that, the internal validation process comes where it can be considered the first test done by ANN on the learning data. This is performed automatically by some algorithms in the software that will be used to perform ANN. For the data of this research, no data validation is needed because the data are real, have formal consent before use and don't need a walk-through audit. The network generalisation will then be completed, since internal validation will be used to test it, and the training process will be terminated when the generalisation is not continuously improving. In fact, this internal testing method has no consequence for the whole training process; it just serves as a metric of network performance that is independent of the whole training process. Finally, the network will be ready to function. Note that all the above equations are inherently used in RStudio software.

In this research, the sigmoid function will be applied in RStudio software for the nodes of the hidden layer. The linear function will be used for the nodes of the output layer. In addition, empirical trials will be done to find the number of nodes in the hidden layer.

Figure 14 illustrates a flowchart of the ANN method's steps and how to use it, where the training process involves the optimisation of the neurons' weights and biases in the network.

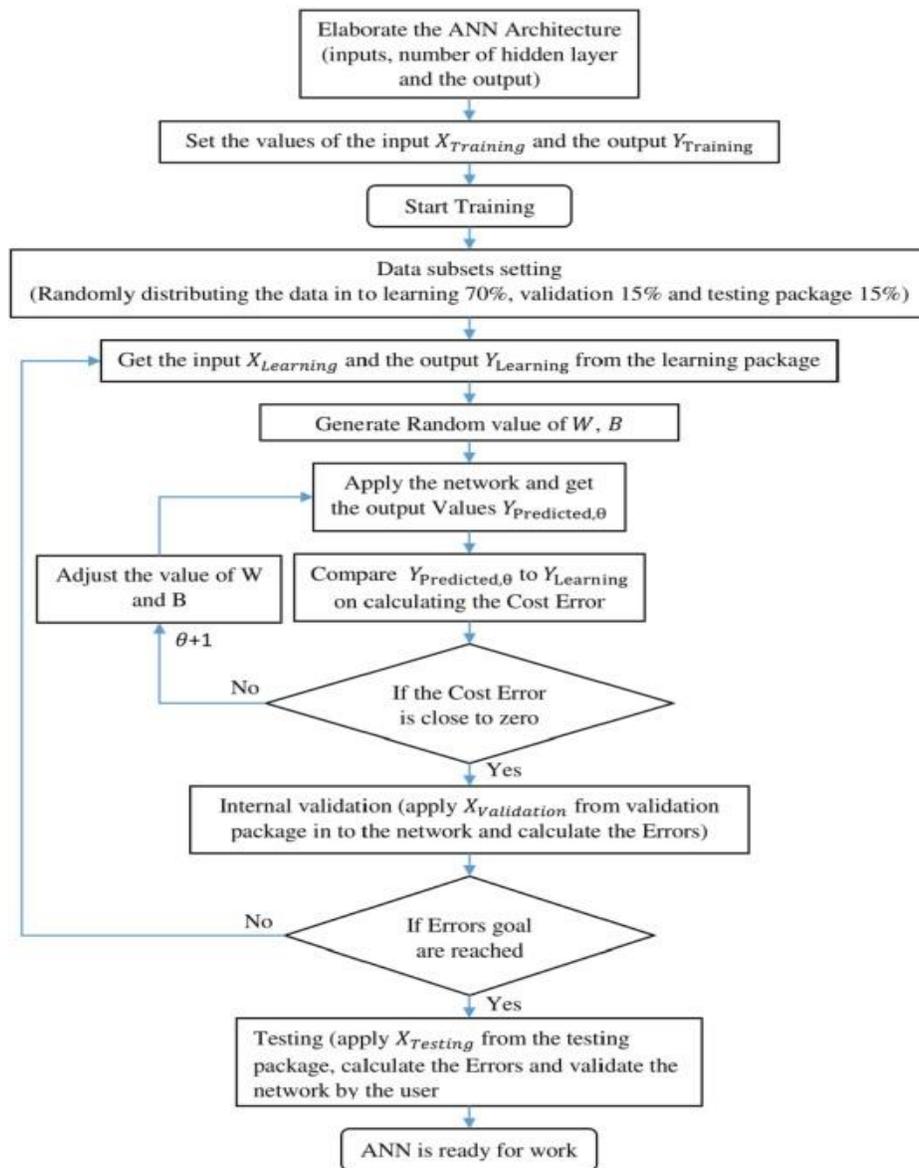


Figure 14: ANN method's flow chart

(Said et al. 2020)

According to Tü and Gürgen (2014), an essential key to performing this analysis is differentiability, which is considered a prerequisite for the method. It is used to update weights in the learning process in relation to the error value. In the forward neural network, when the connection between the nodes is not forming a cycle, the units start taking inputs from the lower layer. Later, processing starts moving to the higher layers. Accordingly, the error function is the output of ANN minus the expected value. Back propagation takes place in the error function using partial derivatives. Finally, the learning stage stops either when the values of the learning stage stabilise or when the error is below the limit. All these processes are performed implicitly using RStudio software.

### 3.3 Checking the Accuracy of Models

In this section, several equations will be illustrated to check the accuracy of linear regression and ANN techniques. These equations are often used to compare the precision of results from one method to another. The following are the equations that will be used in this study:

1 – Mean square error (MSE):

$$MSE = \frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n} \text{ (Wallach \& Goffinet 1989) (39)}$$

Here n is the number of data points, t is the point number,  $\hat{y}_t$  is the predicted value and  $y_t$  is the actual value.

2 – The root mean square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}} \text{ (Chai \& Draxler 2014) (40)}$$

3 – Mean absolute deviation (MAD):

$$MAD = \frac{\sum_{i=1}^n |\hat{y}_i - y_i|}{n} \text{ (Chai \& Draxler 2014) (41)}$$

4 – Symmetric mean absolute percent error (MAPE):

$$MAPE = \frac{\sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{\hat{y}_i} \right|}{n} * 100 \text{ (Goodwin \& Lawton 1999) (42)}$$

5 – Residual standard error (RSE):

$$RSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n-2}} \text{ (Saadeh, Burqan \& El-Ajou 2022) (43)}$$

# CHAPTER IV

## RESEARCH RESULTS AND DISCUSSION

### 4.1 Data Summary

This research examines the effect of four input parameters, where the dataset was gathered from a CCPP in the UAE concerning the baseload power production (P) in MW, by performing statistical analysis. The four variables in the dataset are AT in °C, AP in bar, RH in % and exhaust steam pressure (V) in bar (some units were adjusted to be compatible with the model). The primary target variable is P, and the dataset was recorded using the power plant automation system from the installed transmitters every 15 minutes for one month (September 2021).

The dataset contains 14,400 data points collected and recorded in the datasheet (each parameter has 2,881 data points). This dataset fits into a regression model using Minitab and RStudio. In addition, the dataset is modelled by applying the ANN technique using RStudio to predict the power production and, hence, to predict the input parameters' best performances and study other essential features of these parameters.

The following tables (2–6) contain simple descriptive statistics completed using Microsoft Excel for each parameter in the dataset.

**(a) Variable 1:** Electrical power production

Table 2: Descriptive Statistics for electrical power production

<b>Descriptive Statistics for Electrical Power Production</b>	
Mean	715.079611
Standard Error	1.56763639
Median	717.2125
Mode	805.0924
Standard Deviation	84.1428013
Sample Variance	7080.01102

Kurtosis	-0.8527438
----------	------------

Table 3: Descriptive Statistics for electrical power production

Skewness	-0.1174933
Range	358.1227
Minimum	506.3229
Maximum	864.4456
Sum	2060144.36
Count	2881
Confidence Level (95.0%)	3.07380268

As shown in Table 2, the maximum power production during September 2021 was 864.4456MW, while the minimum power production was 506.3229MW. The power load has sufficiently decreased, utilising a significant shift in power production in the UAE from CCPPs that operate through natural gas and oil to nuclear power plants mainly characterised by zero emissions and high efficiency (Treyer & Bauer 2016).

**(b) Variable 2: Ambient temperature**

Table 4: Descriptive Statistics for ambient temperature

<b>Descriptive Statistics for Ambient Temperature</b>	
Mean	31.8737214
Standard Error	0.03804081
Median	31.91143
Mode	31.14844
Standard Deviation	2.04183848
Sample Variance	4.16910439
Kurtosis	0.11664114
Skewness	-0.430269
Range	11.20415
Minimum	25.29688
Maximum	36.50103
Sum	91828.1912
Count	2881
Confidence Level (95.0%)	0.07458996

**(c) Variable 3: Relative Humidity**

Table 5: Descriptive Statistics for relative humidity.

<b>Descriptive Statistics for Relative Humidity</b>	
Mean	71.3825006
Standard Error	0.1619903
Median	71.57124
Mode	#N/A
Standard Deviation	8.69482083
Sample Variance	75.5999093
Kurtosis	0.34605316
Skewness	-0.4499609
Range	54.81283
Minimum	35.4665
Maximum	90.27933
Sum	205652.984
Count	2881
Confidence Level (95.0%)	0.31762864

**(d) Variable 4: Atmospheric pressure**

The values of the AP were converted from (hPa) to (bar) to match the exhaust vacuum pressure unit. The conversion is shown below:

$$\text{Bars} = \text{hectopascals} \div 1,000 \text{ (Hub 2022) (44)}$$

Table 6: Descriptive Statistics for atmospheric pressure

<b>Descriptive Statistics for Atmospheric Pressure</b>	
Mean	1.00349808
Standard Error	4.7801*10 <sup>-5</sup>
Median	1.003297
Mode	0.999625
Standard Deviation	0.00256573
Sample Variance	6.583*10 <sup>-6</sup>
Kurtosis	-0.8056635
Skewness	0.18953718
Range	0.01107
Minimum	0.99875
Maximum	1.00982
Sum	2891.07797
Count	2881

Confidence Level (95.0%)	9.3728*10 <sup>-5</sup>
--------------------------	-------------------------

(e) **Variable 5:** Exhaust steam pressure

Table 7: Descriptive Statistics for exhaust steam pressure

Descriptive Statistics for Exhaust Steam Pressure	
Mean	0.08614671
Standard Error	0.00026183
Median	0.08453701
Mode	0.10632
Standard Deviation	0.01405394
Sample Variance	0.00019751
Kurtosis	-0.4215716
Skewness	0.50916197
Range	0.06956392
Minimum	0.05722638
Maximum	0.1267903
Sum	248.188659
Count	2881
Confidence Level (95.0%)	0.0005134

- **Covariance Table:**

Table 7 shows the covariance matrix for the five variables.

Table 8: Covariance table for the dataset

	P	AT	RH	AP	V
P	7077.55353				
AT	48.3937111	4.16765729			
RH	-219.9314	-11.893866	75.5736685		
AP	-0.0312872	-0.0013171	-0.0024796	6.5807E-06	
V	0.83628986	0.01253844	-0.0370074	-5.124E-06	0.0001974

The covariance table shows how two variables vary positively or negatively and how strong the relationship is (Kloeckner et al. 2019). For instance, the power production and the AT have medium positive covariance (47.3937111), which means that they also have a directly proportional medium relationship. This contradicts the paper published by Tüfekci (2014), which showed an

inversely proportional relationship between the two variables. Some reasons might include the low scope of data of this research. On the other hand, the power production and RH showed a strong inversely proportional relationship (-219.9314). Other variables appeared to have a weak relationship with power production. Complementing that, some input variables showed an indication of a relationship: RH and AT (-11.893866), which is an inversely proportional relationship.

- **Correlation Table:**

Table 8 shows the correlation matrix for the four input variables.

Table 9: Correlation table for the four input variables

	P	AT	RH	AP	V
P	1				
AT	0.28177428	1			
RH	-0.3007187	-0.6701805	1		
AP	-0.1449736	-0.2514995	-0.1111898	1	
V	0.70744484	0.43709399	-0.3029565	-0.1421456	1

As shown, the highest multicollinearity found between the input variables is for AT and RH, which is a negative correlation (-0.67). It approximately conforms to the result obtained by Tüfekci (2014). Nevertheless, the highest input variable correlation is presented as multicollinearity, and by Tüfekci (2014) was between the AT and the V, which is a positive correlation with a value of (0.84) and similar to the results obtained by Dutta and Ghosh (2021). Moreover, the highest correlation with power production is found with the V, which is a positive correlation (0.707). Other variables such as AP and RH negatively correlate with the power production that is (-0.1449736) and (-0.3007187), respectively. Furthermore, AT positively

correlates with power production (0.28177428). Additionally, Tüfekci (2014) showed that the highest correlation with the power production was with the AT variable which is (-0.95).

### 4.2 Linear Regression Analysis

This section illustrates the analysis of the dataset using regression analysis using Minitab software and some other calculations by Microsoft Excel. Each input variable’s impact on the power production variable is discussed below using hypothesis testing and assuming equal variances, normal distribution of data and a 95% confidence interval for all Minitab analyses.

The first step in performing this analysis is to check whether there is a trend in the data with time and to check the outliers based on minutes, hours, days and week-long periods.

#### (1) Plotting power production with time:

To perform this step, Microsoft Excel was used to plot power in MW with time on weekly basis to check the possible trends. Figure 15 shows power versus time where the data is segmented into four and a half weeks. Accordingly, the data contains 672 data points for week one, 733 data points for week two, 672 data points for week three, 672 data points for week four and 121 data points for the last two days of the month.

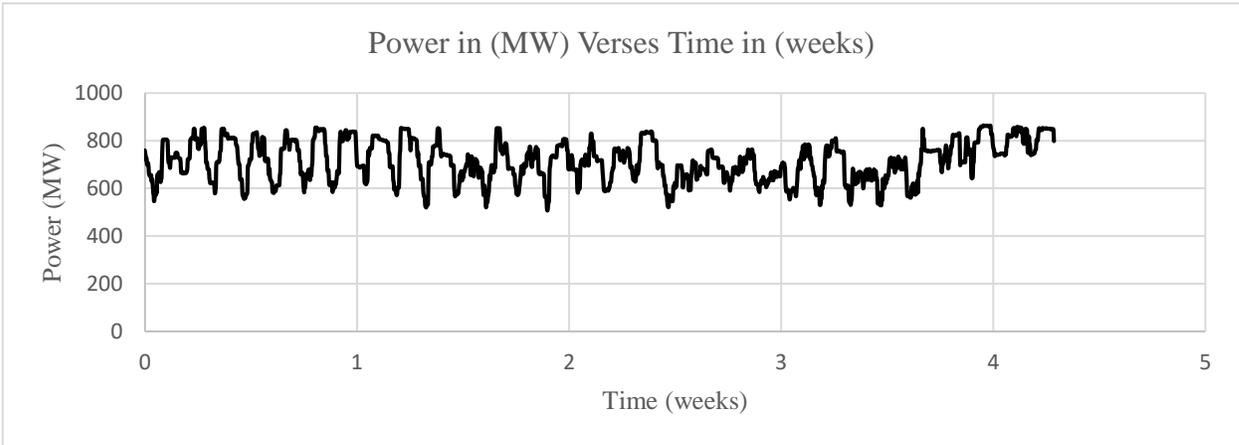


Figure 15: Power in MW versus time in weeks

As shown in Figure 15, there are no main outliers that might affect the analysis or that must be removed before the calculations start, as the data ranges from an interval of 500MW to nearly 900MW. In addition, this figure shows that there are almost seven cycles every week (for each day there is one cycle). Moreover, figures A.1, A.2, A.3, and A.4 in the Appendix show the power data versus time plots in seconds, minutes, hours and days, respectively, to enable an analysis of the data subsets and check whether there are trends or cycles. The figures show that there are three to four cycles in power production every 5,000 minutes (83 hours) and almost one cycle every 2,000 minutes. Moreover, Figure A.3 shows a trend of approximately one process every 25 hours (just over one day), which is illustrated in Figure A.4 and might explain the gradual increase in production. In addition, the weekly trend of data is shown in Figure 15. It demonstrates that the overall weekly trend of power production was steady during the first week, decreasing during the second and third, increasing during the fourth week and almost constant during the final period.

## **(2) Regression analysis and checking linearity**

Checking whether the data is linear or nonlinear is the first step in this analysis. To start, the whole data was tested for linearity using Minitab. Following that, the dataset was segmented using Microsoft Excel into five data subsets representing weeks one, two, three and four and two days of week five (total of 2,881 data points), as the overall trend of data shown in Figure 15, and to find the best data subset that better predicted the power production. Secondly, hypothesis testing was performed for the whole dataset values and for each weekly predictive model to determine whether the variables have a linear relationship with the power production. In addition, the predictive model equation was found for each week, and plots of residuals, normality and the histogram of residuals were obtained using Minitab (see figures A.5 – A.19 and tables A.1 – A.20 in Appendix A). Regression analysis and ANOVA were used to find the relationship between the

dependent variable (power production) and the independent variables (AT, RH, AP, and V), to select the best data subset that best fits the model. The best data subset should have the highest value of R-sq and mostly related residual graphs to the assumptions of the study. Using Minitab, the following values of R-squared (R-sq) and adjusted R-squared (R-sq(adj)) were found for the entire dataset's predictive model and for each week, respectively (see Table 9 and tables A.2, A.6, A.10, A.14, A.18 in Appendix A).

Table 10: R-squared and adjusted R-squared values for the whole dataset model and each week's predictive model

	<b>Whole dataset</b>	<b>Week 1</b>	<b>Week 2</b>	<b>Week 3</b>	<b>Week 4</b>	<b>Week 5 (2 days only)</b>
<b>R-sq</b>	53.49%	82.16%	54.67%	42.97%	62.27%	91.26%
<b>R-sq (adj)</b>	53.43%	82.05%	54.43%	42.63%	62.05%	90.96%

As shown in Table 9, the highest R-squared and adjusted R-squared values were for the week-five prediction model. However, this model data is only for two days. In addition, the normality plot (Figure A.17 in Appendix A) of the predictive model is not compatible with the assumptions of the study, and the residuals scatter plot (Figure A.18 in Appendix A) is not perfectly spread. Moreover, the histogram of residuals (Figure A.19 in Appendix A) is not clearly bell-shaped. As a result, this model is insufficient and needs more data to be adjusted. In addition, it is inaccurate to generalise the results of two days to the whole year of production.

The other predictive models obtained, such as models of weeks two, three and four (refer to Appendix A for figures and tables of each model), have low values of R-squared and adjusted

R-squared compared to the week-one prediction model, which has the highest value of R-squared (82.16%) and adjusted R-squared (82.05%) (see Table A.2 in the Appendix). This result implies that the four input variables cause almost 82% of the variability on the power production variable model equation. Accordingly, this prediction model data analysis will be explained along with the whole dataset prediction model. However, the data is for one week only and cannot be generalised to an entire year. Moreover, other explanations are provided below for testing the assumptions of this study, which confirms the choice of this model and its validity. Note that the following variable symbols illustrate the symbols for the whole dataset.

– Hypothesis test:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad (45)$$

there is no relationship between P and AT, RH, AP, V

$$H_1: \text{at least one } \beta_i \neq 0 \text{ (for } i = 1,2,3,4) \quad (46)$$

there is at least one relationship between P and AT, RH, AP, V

t-Test was performed using Minitab for  $t(1 - \frac{\alpha}{2}, df)$  where  $\alpha$  is the significance level which is 0.05 and  $df$  is the DF (the number of observations minus the estimate parameters number:  $n - 2$ ) (Dutta & Ghosh 2021). The P-values (Pr) in the ANOVA table (Table 11) are (0.0) for all the four input variables, which are all less than  $\alpha = 0.05$ . Accordingly, this rejects the null hypothesis and means that there is 95% confidence that there is a linear relationship between the power production and the AT, RH, AP, and V (significant model).

The following equation was found using Minitab and represents the power production predictive model and all the input variables' coefficients for the whole data:

$$P = 5152 - 9.698 AT - 2.485 RH - 4304 AP + 4274.0 V \quad (47)$$

This model contradicts Tüfekci (2014) and Dutta and Ghosh (2021) in the correlation of some variables. In Tüfekci (2014), AP and RH are positively correlated to the power production. In Dutta and Ghosh (2021), RH is positively correlated to the power production and V is negatively correlated to P. Some reasons might include the limited data in this research and a broader data collection over a more extended period in the Tüfekci (2014) and Dutta and Ghosh (2021) studies, which lasted over six years. In addition, the V has a positive linear relationship with power that opposes Tüfekci (2014), who showed a strong negative correlation of (-0.87).

The regression coefficients were obtained from Table 12 and represent each variable's slope and the intercept. The power changes for each unit change in the input variable (slope value) while keeping other variables constant. Moreover, Table 12 shows the variance inflation factor (VIF) values which represents the multicollinearity in the set of multiple regression variables. It ranges from one to five, demonstrating a strong relationship between power and V, which indicates the validity of the model.

The normal probability plot of the residuals shows almost a straight line (approximately normal distribution of data variables) (see Figure 16). Moreover, the residuals plot (Figure 17) shows that the residuals are randomly scattered, and no clear pattern is observed. However, the model needs more data for future adjustments and there is space for improvements. In addition, the histogram plot of the frequency of residuals shows almost a bell shape, which indicates the normality of the data (see Figure 18). The following are the tables and figures obtained from Minitab for the dataset prediction model:

Table 11: Dataset model summary using Minitab

<b>S</b>	<b>R-sq</b>	<b>R-sq(adj)</b>	<b>R-sq(pred)</b>
57.4209	53.49%	53.43%	53.35%

Table 12: Analysis of variance for the whole dataset prediction model

<b>Source</b>	<b>DF</b>	<b>Adj SS</b>	<b>Adj MS</b>	<b>F-Value</b>	<b>P-Value</b>
<b>Regression</b>	4	10907805	2726951	827.06	0.000
<b>AT</b>	1	460615	460615	139.70	0.000
<b>RH</b>	1	627516	627516	190.32	0.000
<b>AP</b>	1	278467	278467	84.46	0.000
<b>V</b>	1	8385744	8385744	2543.32	0.000
<b>Error</b>	2876	9482627	3297		
<b>Total</b>	2881	20390432			

Table 13: Coefficients table for the whole dataset prediction model

<b>Term</b>	<b>Coef</b>	<b>SE Coef</b>	<b>T-Value</b>	<b>P-Value</b>	<b>VIF</b>
<b>Constant</b>	5152	487	10.58	0.000	
<b>AT</b>	-9.698	0.820	-11.82	0.000	2.45
<b>RH</b>	-2.485	0.180	-13.80	0.000	2.14
<b>AP</b>	-4304	468	-9.19	0.000	1.26
<b>V</b>	4274.0	84.7	50.43	0.000	1.24

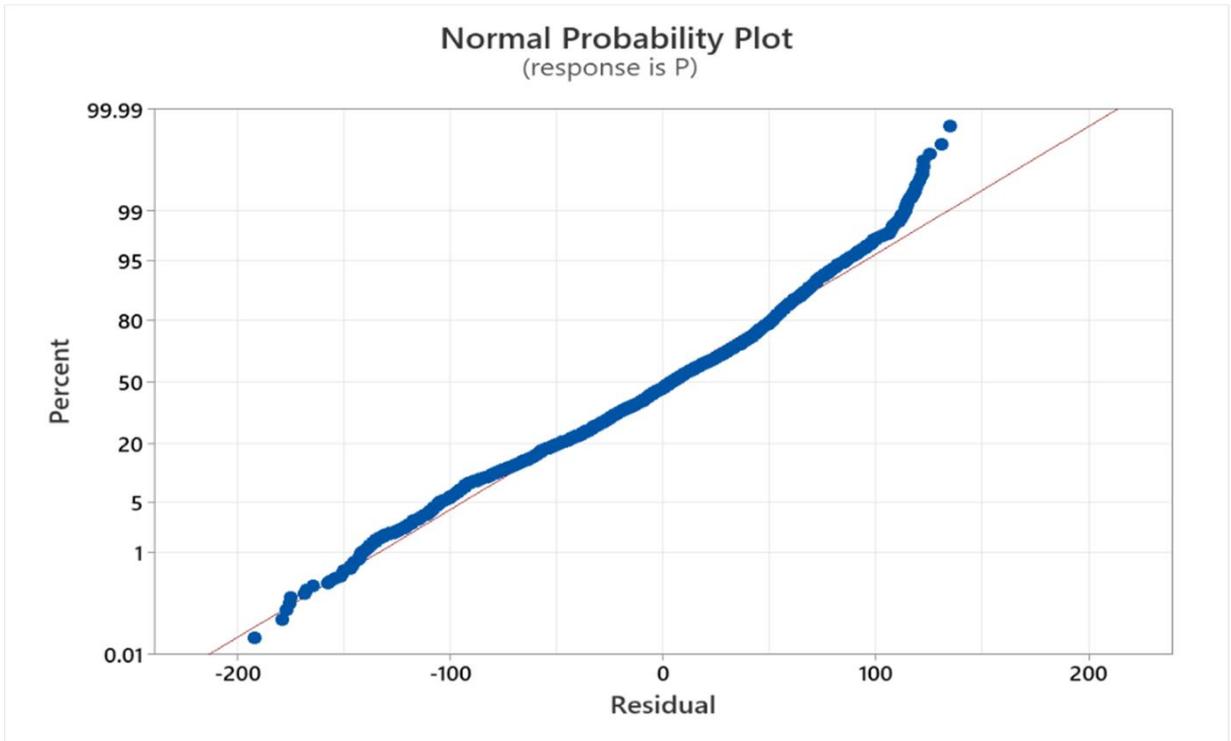


Figure 16: Normality plot for the dataset prediction model

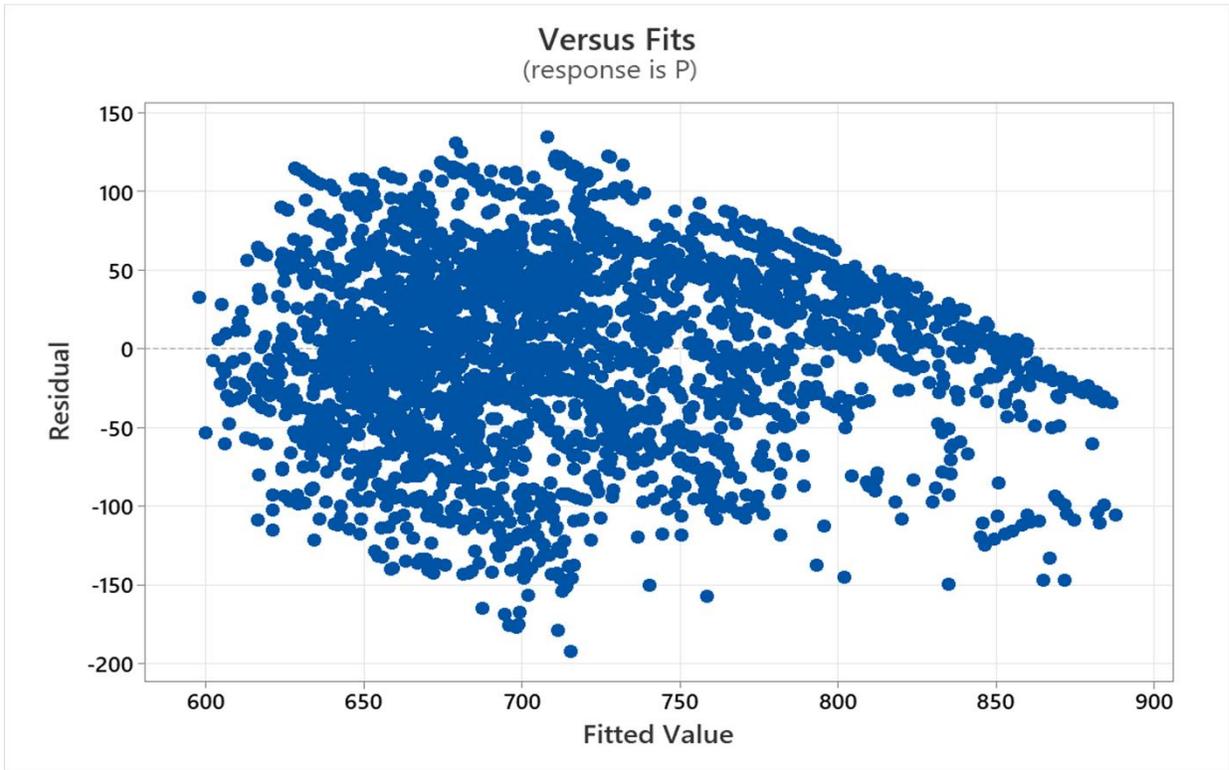


Figure 17: Residuals plot versus fitted values for the dataset prediction model

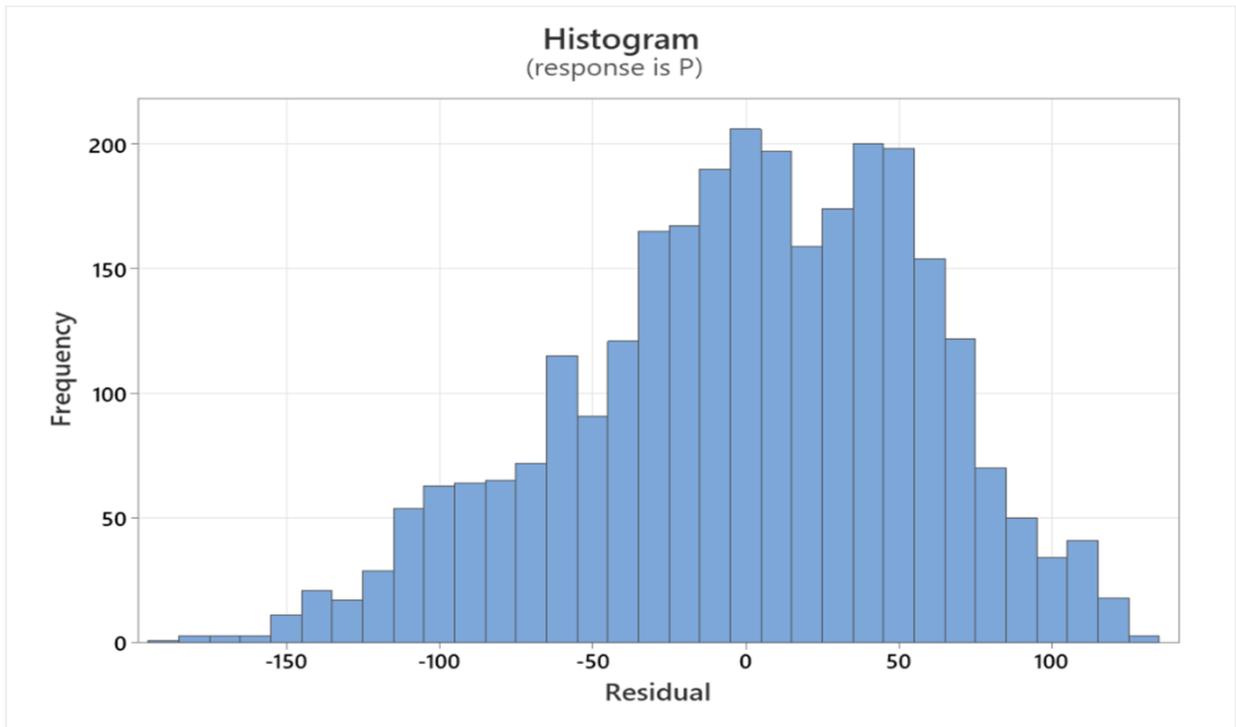


Figure 18: Histogram of residuals for the dataset prediction model

### (1) Ambient temperature effect on power production

This impact is considered to be significant and affects the performance of gas turbines (Tüfekci 2014). Moreover, this is shown in the analysis of AT as the correlation of this independent variable with the power production was found to be (0.28177428), which is a weak correlation based on the weather conditions of the UAE (see Table 8).

As stated by Arrieta and Lora (2005), the operation of a CCPP is affected by ambient parameters where the plant is installed, such as AT, RH and AP. Among these variables, the most substantial effect was the AT effect. However, the AT doesn't significantly affect the power production in this study due to limited data collection. Moreover, the curves of power production, thermal efficiency and heat rate generated in a Arrieta and Lora (2005) study showed a clear tendency for power to shift and be affected by the AT and the complementary firing.

Controlling the AT outside the gas turbine is explained by TMI Staff and Contributors (2017). The high temperature of the surrounding air is correspondent to the low density of the inlet air that enters the gas turbine. Accordingly, the mass flow rate through the gas turbine is minimal, which contributes to low efficiency and power production. Consequently, the AT has an inverse relationship to the mass flow of air that is needed for firing, which means that decreasing the AT increases the mass flow for inlet air. Moreover, Park et al. (2020) highlighted predicting the performance of gas turbines by their inlet parameters, such as AT. If the air temperature at the gas turbine's input was extremely high, the turbine rotor could be destroyed, which affects the power generation.

The analysis of this study showed a directly proportional relationship between AT and P considering the AT effect on P only. However, this relationship is weak due to limited data collection over a short period. In addition, considering the AT effect on P along with other variables has presented an inversely proportional relationship between AT and P, which is compatible with the result of Tüfekci (2014) and Dutta and Ghosh (2021). The relationship is explained by the fact that every unit increase in AT results in decreasing the power production by 9.698 times, while keeping all other variables constant (see equation 47 on page 56).

Figure 19 shows the scatter plot of the AT versus power production and the linear regression line after fitting the data using Minitab (for the purposes of the initial investigation). Note that, using Minitab, the R-squared and adjusted R-squared values found for the P and AT prediction model were higher for week 1 data (24.05% and 23.94%, respectively) and lower for the whole dataset model (7.94% and 7.91%, respectively).

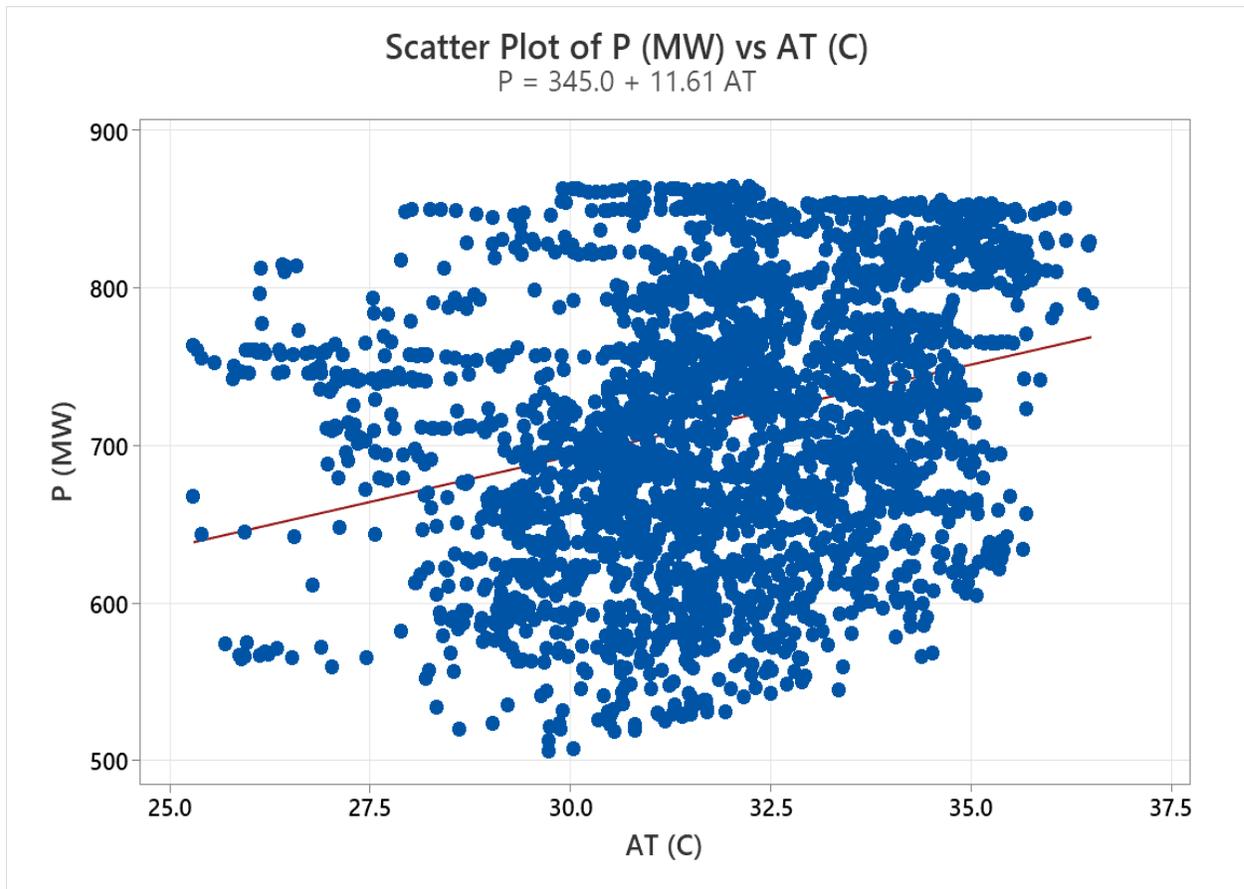


Figure 19: Scatter plot of power production versus ambient temperature

Although it is a weak linear relationship, the model equation (on page 56) represents the predictive model for power in the UAE after performing a hypothesis test for the AT and all the other variables with the power production.

## **(2) Relative humidity effect on power production**

Previous research has shown that, as RH rises, the power production of a steam turbine rises due to an increase in the temperature of a gas turbine's exhaust gas as explained by Tüfekci (2014). However, the model was explicated to be insufficient and needed some adjustments. In this study, the correlation of RH concerning power production was found to be (-0.3007187), which is a moderate negative correlation (see Table 8). The equation illustrates that one unit increase in RH decreases the power production by 2.485, while keeping all the other values constant. This

indicated a presence of an inversely proportional relationship between power production and RH. Additionally, Onoroh, Ogbonnaya and Onochie (2020) disproved the Tüfekci (2014) and Dutta and Ghosh (2021) results for RH and power and confirmed the inversely proportional relationship as illustrated in this research. Several implications of high RH and outer condensation are explained by Dehghani Samani (2018), and include:

- reduction in air compressor's efficiency as the surface of the compressor blades becomes wet;
- reduction in the efficiency of the production as the latent heat of the water droplets raises the inlet gas turbine temperature, decreasing the mass flow of inlet air;
- fouling might occur in the compressor blades due to the decomposition of particles of wet inlet air;
- the pressure instrumentations might get blocked by water droplets and therefore affect the readings of pressure lines.

Correspondingly, the Onoroh, Ogbonnaya and Onochie (2020) study also mentioned other variables simultaneous with highly moisturised air that might change the performance of the gas turbine and, hence, the power production efficiency, which are the dry-bulb temperature, wet bulb temperature and the AP (which will be discussed in the following section).

Dehghani Samani (2018) demonstrated that the RH must not exceed 75%. Nevertheless, this research data ranges from 35.47% to 90.28%, which might affect the power production efficiency.

Another study by Amell and Cadavid (2002) confirmed the results of this research and was performed in Colombia to study electricity production. The researcher confirmed that increasing

the AT of a gas turbine decreases the power production. The inlet air was cooled before entering the gas turbine to resolve this issue. This drives the concerns about the effect of RH. One main result illustrated in that research was that the high RH ratio increased the load of inlet air cooling by 1.5–1.9 times compared to low RH regions of 30% and lower. Owing to that, higher costs correspond to this heavy load of cooling for both techniques: the ice storage and vapour compression cycles.

Figure 20 shows the scatter plot of the RH versus power production and the linear regression line for the initial investigation’s purposes.

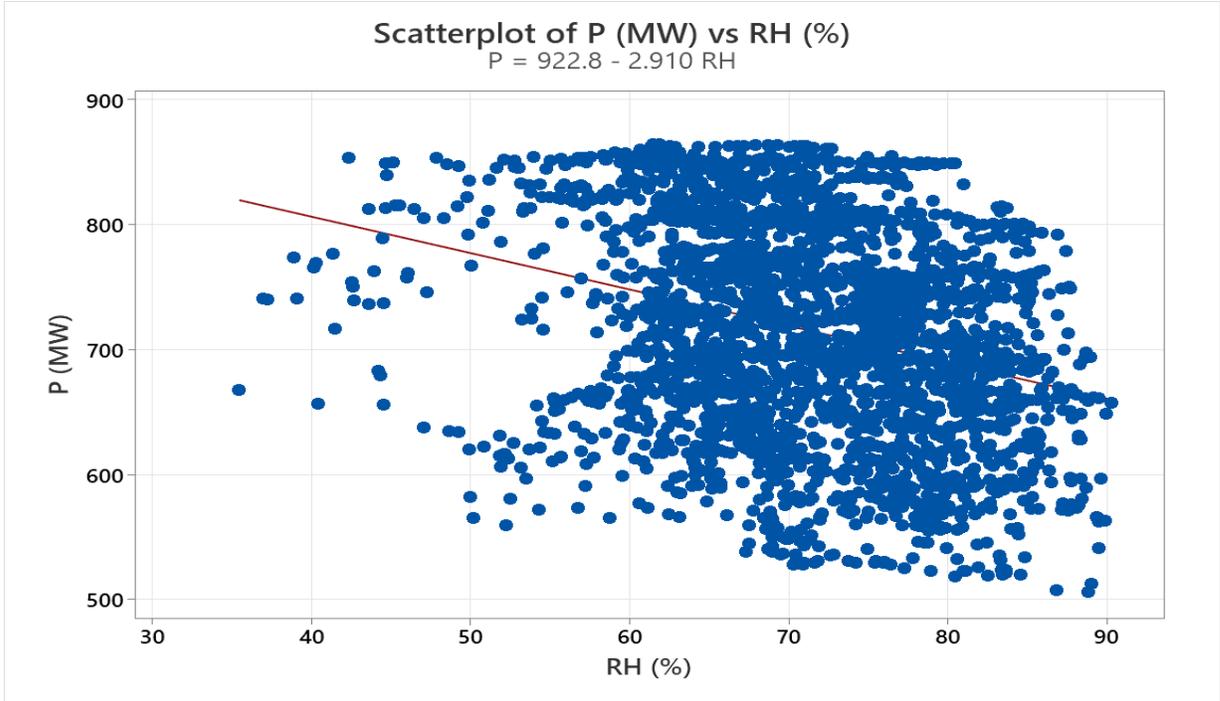


Figure 20: Scatter plot of power production (P) versus relative humidity (RH)

Although it is a weak linear relationship, the model equation on page 56 represents the predictive model after performing a hypothesis test for RH and power production. However, the model needs more data to be a more reliable predictive model.

### **(3) Atmospheric pressure effect on power production**

As stated by Tüfekci (2014), the second most influential variable in the power prediction model is the AP, which was shown to be directly proportional to power. Another view is provided by El Hadik (1990) and confirmed by Tüfekci (2014) showing that the combined effect of AP and the AT corresponds to the change in the density of inlet gas turbine air. Moreover, an El Hadik (1990) study insisted that hot weather conditions and AP in Kuwait and nearby Arab countries drastically influenced gas turbine performance. Furthermore, a critical review from Hashmi, Majid and Lemma (2020) confirmed the influence of various ambient conditions on the compressor at the gas turbine's inlet, leading to decreased efficiency. Such temperature and pressure conditions make the inlet air contaminates form a mass that sticks to the blades of the compressor and causes fouling.

Consequently, the airflow and compressor pressure ratios decline with time. In addition, the previous actions might lead to trigger surging, which causes failure to operate the compressor. As per Park et al. (2020), this degradation must be forecasted using a mathematical model, as it is hard to indicate in a real industry system. Moreover, it must be predicted mathematically before using AI such as ANNs.

In this study, the AP at the inlet of a gas turbine (compressor inlet) is explicitly chosen to be studied, as the external air pressure varies with attitude. As shown in Table 8, the AP has almost no substantial relationship with the target variable (-0.1449736). However, the model prediction equation (on page 56) represents the relationship found between AP and P using Minitab, which

illustrates that every unit change in AP changes power production by -4304 units while keeping all the other variables constant.

Figure 21 shows the scatter plot of the AP versus power production and the linear regression line after fitting the data for investigation.

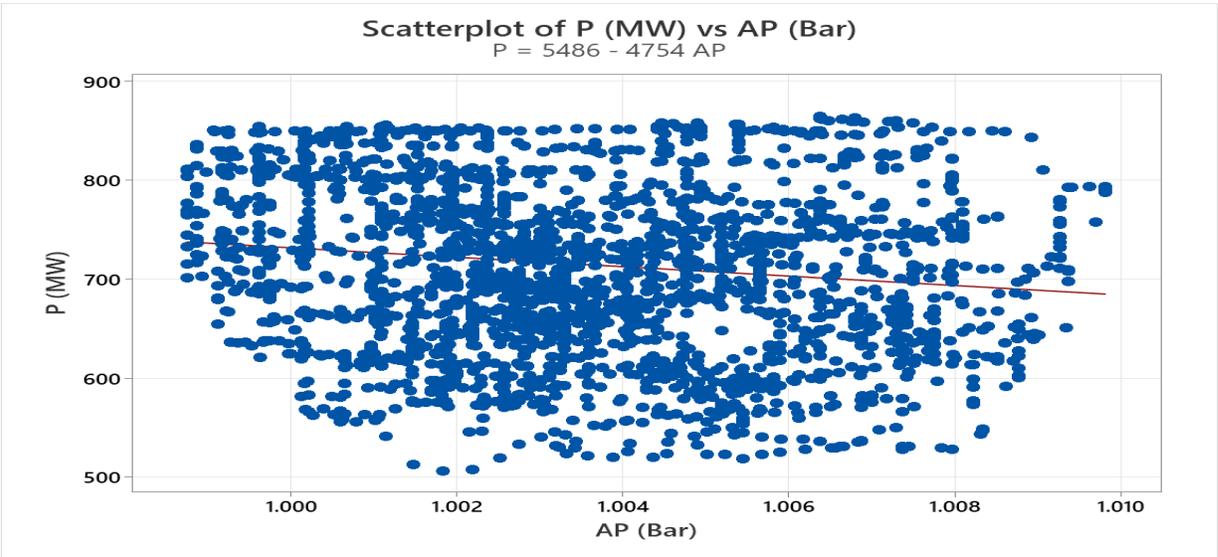


Figure 21: Scatter plot of power production (P) versus atmospheric pressure (AP)

The model equation represents the predictive model after performing a hypothesis testing for the AP and the power production. The line shows an inversely proportional relationship between the two variables, which contradicts Tüfekci (2014) and Dutta and Ghosh (2021), as they showed a directly proportional relationship between the two variables. This means that this study needs more than one month of data to show a more accurate result. On the other hand, Park and Kim (2006) explained the functionality of the hybrid solid oxide fuel cell (SOFC) – a gas turbine system based on ambient pressure conditions (high-temperature conditions) – and compared it with pressurised systems. The study showed that the pressurised system performed well with high-inlet GT temperature. Therefore, higher GT pressure ratios barely increase the system’s efficiency. However, the study found that combining an ambient pressure hybrid system with a GT of high-

pressure ratio is not reasonable, as the efficiency of this ambient condition-based design is lower than the SOFC system efficiency.

**(4) Exhaust vacuum pressure effect on power production**

To enhance the efficiency of the power generation system the power plant employs a steam turbine, as seen in Figure 3. Relatively, the V pressure plays a significant role in increasing electrical efficiency and affecting the performance of the steam turbine. The data of the V was collected from the steam turbine.

Figure 22 shows the scatter plot of the V versus power production and the linear regression line after fitting the data for the investigation’s purposes.

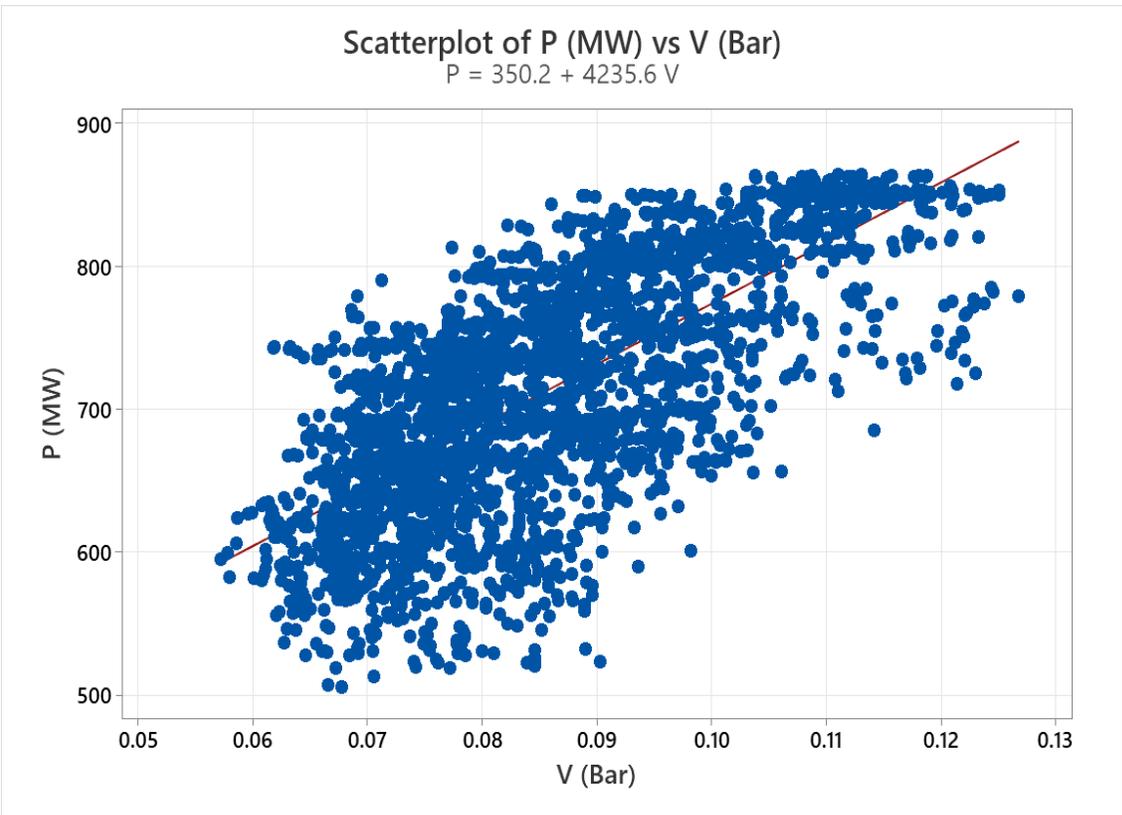


Figure 22: Scatter plot of power production (P) versus exhaust vacuum (V)

As the figure shows, the V and power production have a strong, directly proportional, relationship. The correlation between these two variables was found to be 0.70744484, which is relatively high (see Table 8). As the model equation of power production shows (see page 56), every one unit change in V increases the power production by 4,274 units, while keeping all the other variables constant. This result disproves the results of Tüfekci (2014) and Dutta and Ghosh (2021) for the V, due to a data collection error or the limit of the scope of data in this study. Moreover, Tüfekci (2014) showed that the slope of V and power model equation is greater than the slope of the AP and RH simple linear equations but lower than the AT simple linear model equation. On the other hand, for the multiple linear regression model equation, this study showed that the V model equation's slope magnitude is greater than RH and AT but less than AP.

Moreover, this study also contradicts Dutta and Ghosh (2021), showing an inversely proportional relationship between power and V. In addition, the effect of V is explained by Madu and Nwankwo (2018), in which the optimisations of the steam turbine dynamics were discussed. Usually, the exhaust temperature of the steam turbine is very high. Accordingly, optimising the process is done by improving the water rate and reducing the required energy for operation. This is also done by controlling the steam inlet temperature, turbine exhaust vacuum and inlet pressure of steam. These parameters will affect the performance of the steam turbine, its efficiency and power production. In addition, V or low pressure is required at the end of the steam turbine so that the steam can expand and its kinetic energy increase, leading to higher spin of turbine blades and more power production. As a result, the condenser's vacuum serves to reduce the pressure at the turbine's outlet so that it is below AP. Hence, it aids in getting the most energy possible out of the steam.

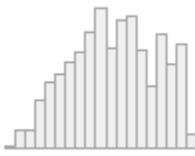
### 4.3 Linear Regression Using Program Coding

In this section, linear regression analysis is performed on the dataset using RStudio software to find the actual and predicted values of power production and compare them with the results of Minitab and ANNs data analysis. This step aims to find the predicted values of the linear regression power model using RStudio and compare it with ANN predicted values of power found using the same software. This step is necessary as the program uses coding for prediction model analysis. Some additional steps were performed to see if the model was valid. The following steps illustrate how the method was applied using RStudio and some additional results:

#### 1- Summary of the data

Considering all the attributes where P is power production, AT is ambient temperature, RH is relative humidity, AP is atmospheric pressure, and V is the exhaust vacuum, Table 13 shows the actual data summary. In addition, Table 14 shows the learning data summary and Table 15 is for the testing data summary, where all the data were randomly distributed to 70% of learning data and 30 % for testing data as a first step in applying ANN. Note that data validation is not required, as the data was taken directly from the power plant transmitters, and no data points were missing.

Table 14: Data summary using RStudio software for the actual value of power production

No	Variable	Stats/Values	Freqs (% of Valid)	Graph	Valid	Missing
1	P [numeric]	Mean (sd): 715.1 (84.1) min ≤ med ≤ max: 506.3 ≤ 717.2 ≤ 864.4 IQR (CV): 126.7 (0.1)	2,834 distinct values		2,881 (100.0%)	0 (0.0%)

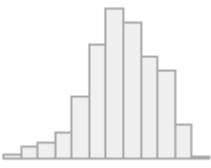
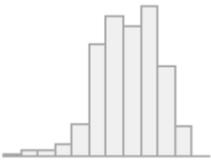
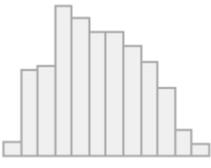
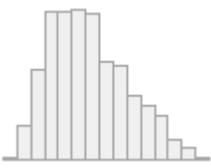
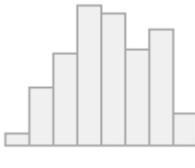
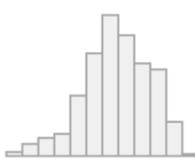
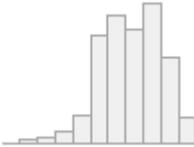
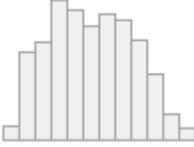
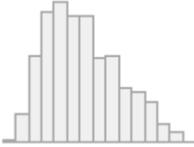
2	AT [numeric]	Mean (sd): 31.9 (2) min ≤ med ≤ max: 25.3 ≤ 31.9 ≤ 36.5 IQR (CV): 2.8 (0.1)	2,693 distinct values		2,881 (100.0%)	0 (0.0%)
3	RH [numeric]	Mean (sd): 71.4 (8.7) min ≤ med ≤ max: 35.5 ≤ 71.6 ≤ 90.3 IQR (CV): 12.4 (0.1)	2,881 distinct values		2,881 (100.0%)	0 (0.0%)
4	AP [numeric]	Mean (sd): 1 (0) min ≤ med ≤ max: 1 ≤ 1 ≤ 1 IQR (CV): 0 (0)	1,235 distinct values		2,881 (100.0%)	0 (0.0%)
5	V [numeric]	Mean (sd): 0.1 (0) min ≤ med ≤ max: 0.1 ≤ 0.1 ≤ 0.1 IQR (CV): 0 (0.2)	2,516 distinct values		2,881 (100.0%)	0 (0.0%)

Table 15: Data summary using RStudio software for the 70% learning data

No	Variable	Stats/Values	Freqs (% of Valid)	Graph	Valid	Missing
1	P [numeric]	Mean (sd): 715.3 (84.8) min ≤ med ≤ max: 506.3 ≤ 718 ≤ 864.4 IQR (CV): 131.5 (0.1)	1,990 distinct values		2,016 (100.0%)	0 (0.0%)
2	AT [numeric]	Mean (sd): 31.9 (2.1) min ≤ med ≤ max: 25.3 ≤ 31.9 ≤ 36.5 IQR (CV): 2.9 (0.1)	1,900 distinct values		2,016 (100.0%)	0 (0.0%)

3	RH [numeric]	Mean (sd): 71.3 (8.7) min ≤ med ≤ max: 37 ≤ 71.4 ≤ 89.9 IQR (CV): 12.5 (0.1)	2,016 distinct values		2,016 (100.0%)	0 (0.0%)
4	AP [numeric]	Mean (sd): 1 (0) min ≤ med ≤ max: 1 ≤ 1 ≤ 1 IQR (CV): 0 (0)	919 distinct values		2,016 (100.0%)	0 (0.0%)
5	V [numeric]	Mean (sd): 0.1 (0) min ≤ med ≤ max: 0.1 ≤ 0.1 ≤ 0.1 IQR (CV): 0 (0.2)	1,810 distinct values		2,016 (100.0%)	0 (0.0%)

## 2- Fitting a linear regression model

To perform this step, the `lm()` function in RStudio was used. In addition, the summary of the analysis and the model fitting was done using the `summary()` function. The following figures 23, 24 and 25 show the residuals, coefficients of the equation and the other values, respectively, after performing hypothesis testing for the actual power production model fitting.

<b>Residuals:</b>				
Min	1Q	Median	3Q	Max
-192.031	-35.435	4.094	43.907	125.372

Figure 23: Residuals of the actual power production model using RStudio

As shown in Figure 23, the residuals' minimum value, median, 1Q and 3Q are almost scattered, which means that the points are randomly placed and, thus, indicate the significance of the model.

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	5210.3196	573.1953	9.090	< 2e-16	***
AT	-9.8082	0.9630	-10.185	< 2e-16	***
RH	-2.5157	0.2143	-11.738	< 2e-16	***
AP	-4358.2537	551.8135	-7.898	4.62e-15	***
V	4303.2164	102.2123	42.101	< 2e-16	***

Figure 24: Coefficients of the actual power production model using RStudio

Figure 24 shows the coefficients of the power production prediction model found using RStudio. In addition to the standard error values, t values and P-values (Pr) for the hypothesis testing.

### 3- Hypothesis test

The same Minitab hypothesis testing was performed here for RStudio output to determine the model's significance and to check and test later the difference between Minitab output, RStudio output and the output of RStudio ANN technique. As shown in Figure 24, the P-values (Pr) for all the variables are less than,  $\alpha = 0.05$ , which represents the model's significance. Accordingly, the following model equation illustrates the prediction model for power using RStudio:

$$P = -9.8082 AT - 2.5157 RH - 4358.2537 AP + 4303.2164 V + 5210.3196 \quad (48)$$

RStudio model equation is slightly different from Minitab's model equation and represents almost the same result. Note that, to perform this hypothesis testing in RStudio, the assumptions made are the normal distribution of data, equal variances and 95% confidence interval.

Residual standard error: 57.76 on 2011 degrees of freedom
Multiple R-squared: 0.5369, Adjusted R-squared: 0.5359
F-statistic: 582.8 on 4 and 2011 DF, p-value: < 2.2e-16

Figure 25: Other values for the actual power production model's hypothesis testing

Figure 25 shows the RSE, which is the actual power values, minus the predicted power values. Generally, the lower the value, the more accurate is the fit of the data. In addition, the R-squared value (0.5369) and the adjusted R-squared value (0.5359) are moderately high, which indicates that the model accounts for approximately 53% of power production variation (same as the Minitab result).

#### **4- Checking linear regression model's adequacy**

Figure A.20 in Appendix A is RStudio output, which shows a randomly scattered residual around the zero axis. This shows that the input variable and the output variable have a linear relationship. In addition, Figure A.21 in Appendix A is the normal Q-Q plot (theoretical quantities for standard normal versus actual quantities of standardised residuals) for the actual data, which shows that the points almost fall in a diagonal line. This results in supporting the assumption of normal distribution of data.

Consequently, Figure A.22 in Appendix A is the scale location plot which displays the square root of the standardised residuals on the y-axis and the fitted values of the linear regression model on the x-axis. As shown, the spread of the residuals is almost random (not perfectly random). In addition, no clear pattern is observed, which supports the significance of the regression model. However, it should be in a better alignment.

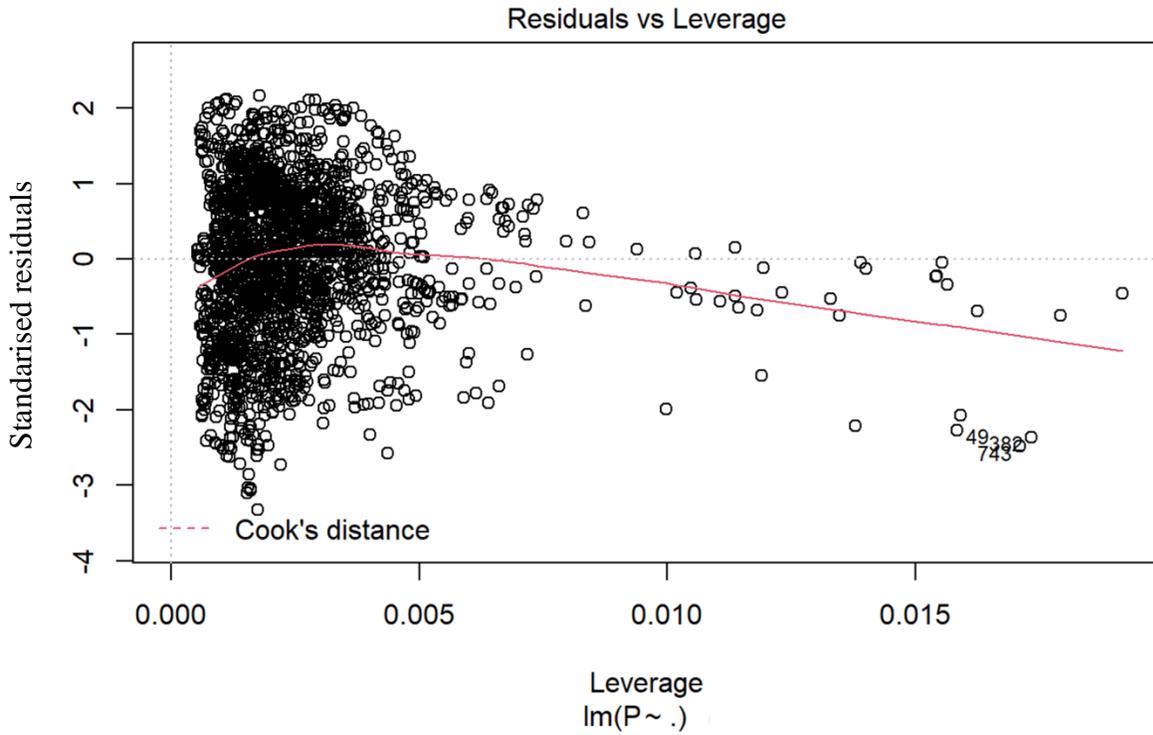


Figure 26: Residuals versus leverage plot for the actual power production

Figure 26 shows the residuals versus the leverage plot, which is used to check how far the data points are from the data line. If one point crosses Cook's distance, there is a need for further investigation, but as shown in the figure, there are no significant outliers.

The function `dfsummary` was used to find the predicted values of power production and fit a regression model for it. The purpose of using this function for the learning and the testing values is to test the similarity between the results. As shown in Table 15, the distribution, mean, median, minimum, maximum and standard deviation values of the testing data for all the five variables are almost the same as the learning data summary (see Table 14 for learning data summary) and the same as the actual data summary (see Table 13 for actual data summary). In addition, the number of cell points in the learning phase is 2,016 out of 2,881, and the number of cell points in the testing phase is 865 out of 2,881 (see Table 15 below). There are no missing values in each test.

Table 16: Data summary using RStudio software for the 30% testing data

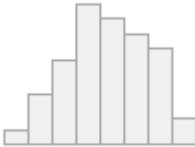
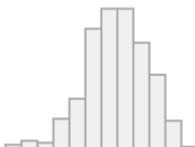
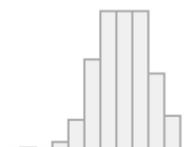
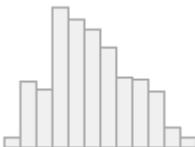
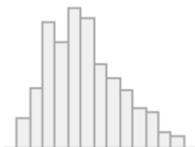
No	Variable	Stats/Values	Freqs (% of Valid)	Graph	Valid	Missing
1	P [numeric]	Mean (sd): 714.6 (82.7) min ≤ med ≤ max: 513.1 ≤ 712.7 ≤ 863.7 IQR (CV): 120.2 (0.1)	859 distinct values		865 (100.0%)	0 (0.0%)
2	AT [numeric]	Mean (sd): 31.9 (1.9) min ≤ med ≤ max: 25.3 ≤ 32 ≤ 36.5 IQR (CV): 2.6 (0.1)	825 distinct values		865 (100.0%)	0 (0.0%)
3	RH [numeric]	Mean (sd): 71.7 (8.7) min ≤ med ≤ max: 35.5 ≤ 71.9 ≤ 90.3 IQR (CV): 12.2 (0.1)	865 distinct values		865 (100.0%)	0 (0.0%)
4	AP [numeric]	Mean (sd): 1 (0) min ≤ med ≤ max: 1 ≤ 1 ≤ 1 IQR (CV): 0 (0)	447 distinct values		865 (100.0%)	0 (0.0%)
5	V [numeric]	Mean (sd): 0.1 (0) min ≤ med ≤ max: 0.1 ≤ 0.1 ≤ 0.1 IQR (CV): 0 (0.2)	784 distinct values		865 (100.0%)	0 (0.0%)

Table 17: Summary table for predicted and actual power production values with the input variables

P (Actual)	AT	RH	AP	V	P (Predicted)
741.1172	30.90449	84.41586	1.000172	0.0878400	713.8255
721.2391	30.58930	85.39021	1.000172	0.0878400	714.4657
657.0855	31.82385	81.26100	1.001151	0.0735904	647.1592
653.2428	31.69922	81.58305	1.001469	0.0748611	651.6537
631.4904	31.69922	79.74304	1.001469	0.0728234	647.5138
576.8560	31.10394	81.31323	1.001875	0.0631968	606.2076

### 5- Plot of actual values versus predicted values

In this step, the plot of the actual power production values versus the predicted ones was found using a specific code, and it is shown in Figure 27 below:

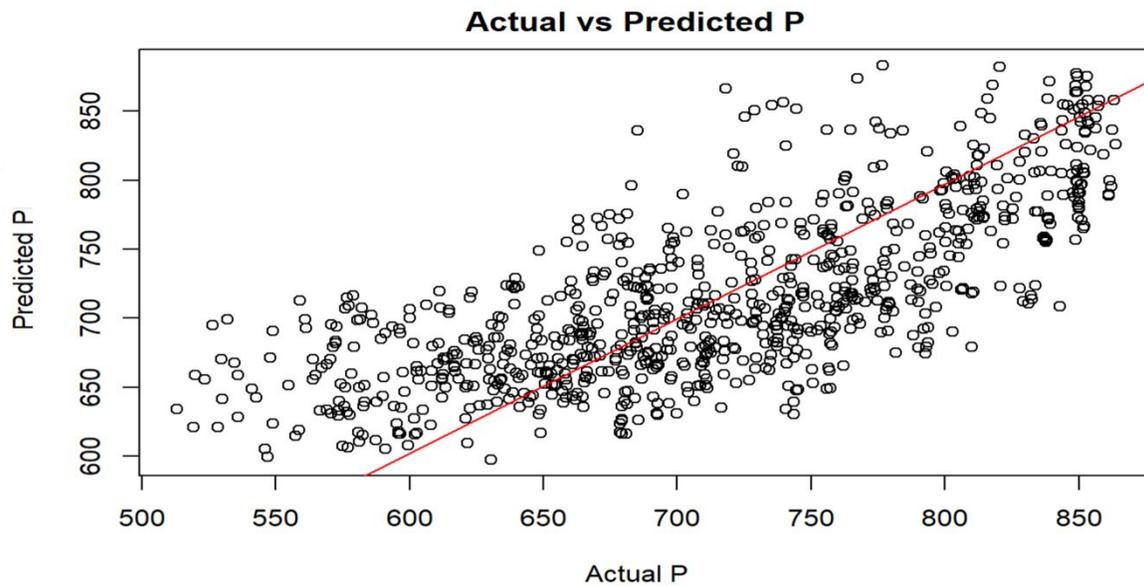


Figure 27: Actual versus predicted values of power production using linear regression by RStudio

## 6- Finding the correlation between actual and predicted power values

```
correlation between actual and predicted values is:  
...{r}  
cor(dat.test$PE,pred)  
...  
[1] 0.7283605
```

Figure 28: The correlation between the actual and predicted power values

As shown in Figure 27, the actual and predicted values of power production are almost aligned with a correlation of 0.7283605 (see Figure 28), which indicates the significance of the prediction model. However, there is a need for better data collection over a more comprehensive period (more than one month) to get better results, better correlation and to improve the model.

### 4.4 Artificial Neural Networks

This section illustrates the use of the machine learning method ANN for analysing the dataset. The advantage of using a neural network is shown in its tremendous structure where there is no need to explain any type of input or output relationships. The AI of neural networks learns the followed relationships from the data provided in this study.

The software used here is RStudio and the analysis involved coding. The following are the steps to get the results:

- 1- The data was divided into two categories: learning and testing, where the same data summary tables (as in the regression part) were found for each. The next step is to find the ANN model where there are many packages used in R to find and fit data to ANN, for example: 'nnet', 'neuralnet', 'MxNet' and 'keras'. In this analysis, the neuralnet package is used.
- 2- Figure 29 represents the ANN model after fitting the CCPP data

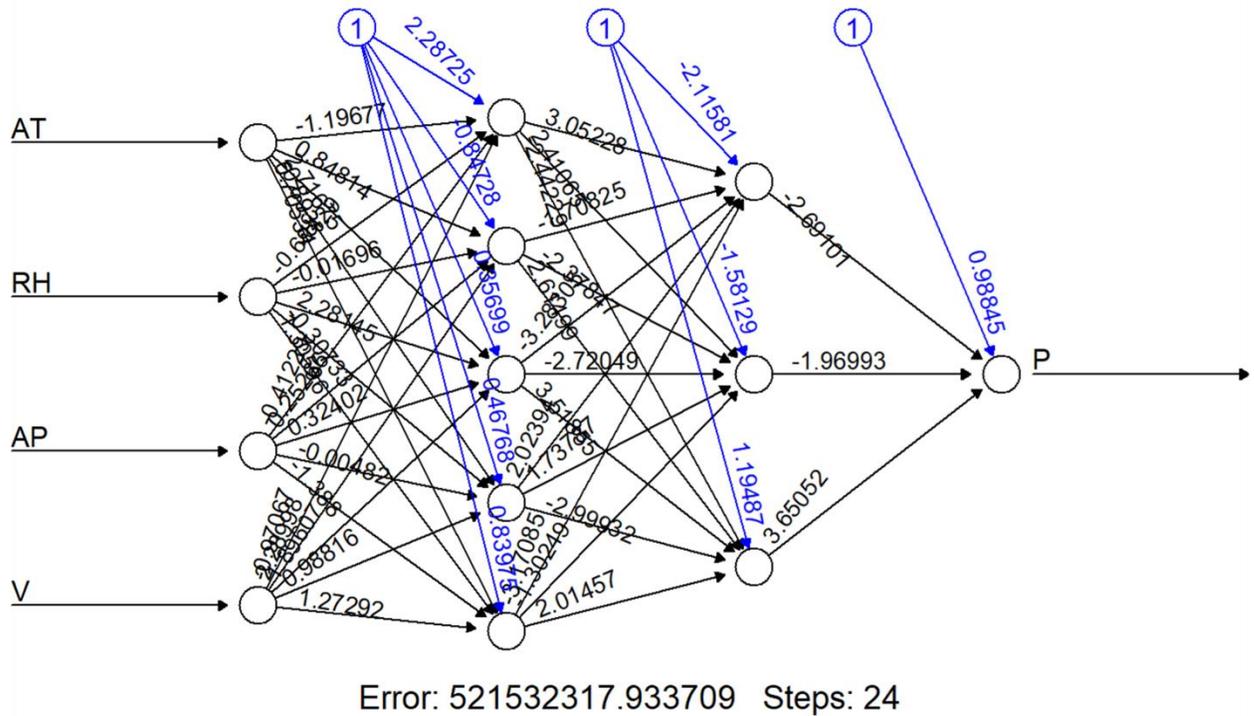


Figure 29: ANN model for the CCPP data

As the figure shows, there is one input layer with four input variables (AT, RH, AP and V), two hidden layers and one output layer, with one output variable (P). The arrows in black with the numbers represent the weights ( $w_i$ ), which shows the contribution of that variable to the next node. In addition, the lines in blue represent the bias weight (b), where all these numbers are used to find the pre-activation equation.

**3-** Since the error value was found to be high (521532317.933709), normalisation and scaling of data to the 0–1 range were done for the actual dataset using the normalise function and the following equation (see Table A.21 in the Appendix for a sample of data after normalisation):

$$z_i = \frac{(x_i - \min(x))}{(\max(x) - \min(x))} \quad (32)$$

4- Figure 30 shows the neural network plot of the CCPP data after normalising the data to the 0–1 range. The error value was reduced to 17.768191.

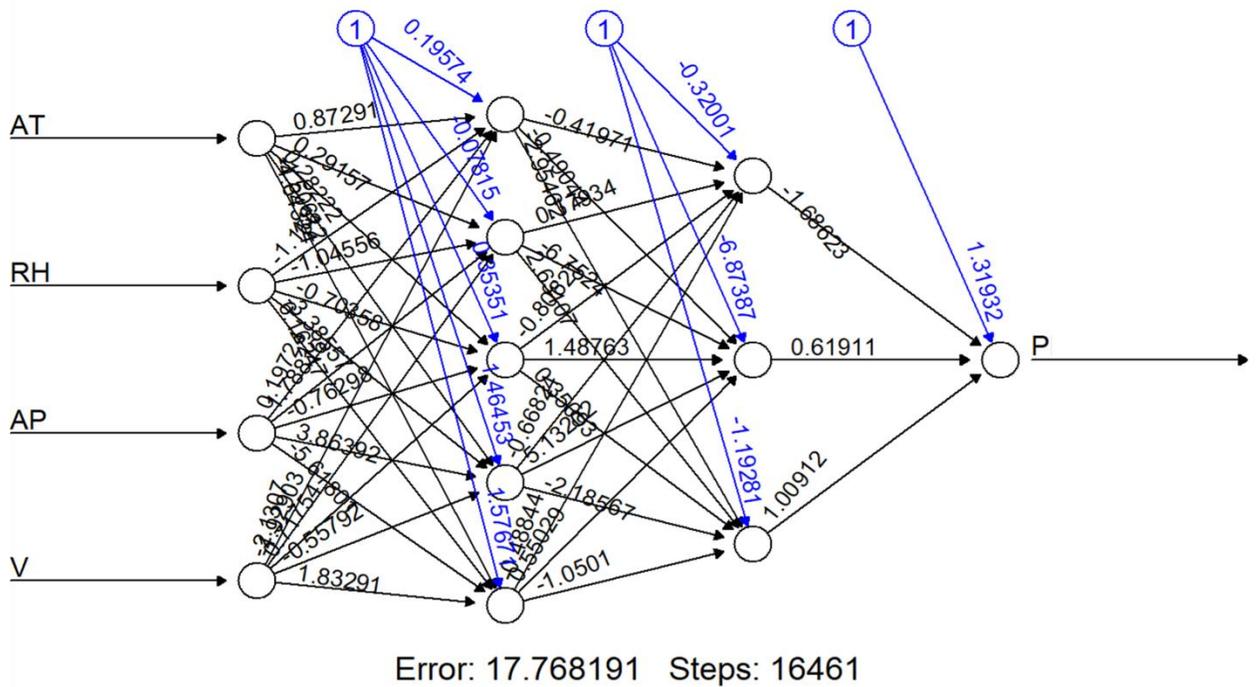


Figure 30: Neural network plot for the dataset after normalisation

5- Figure 31 shows the plot of the normalised predicted power production values versus the actual values of power production

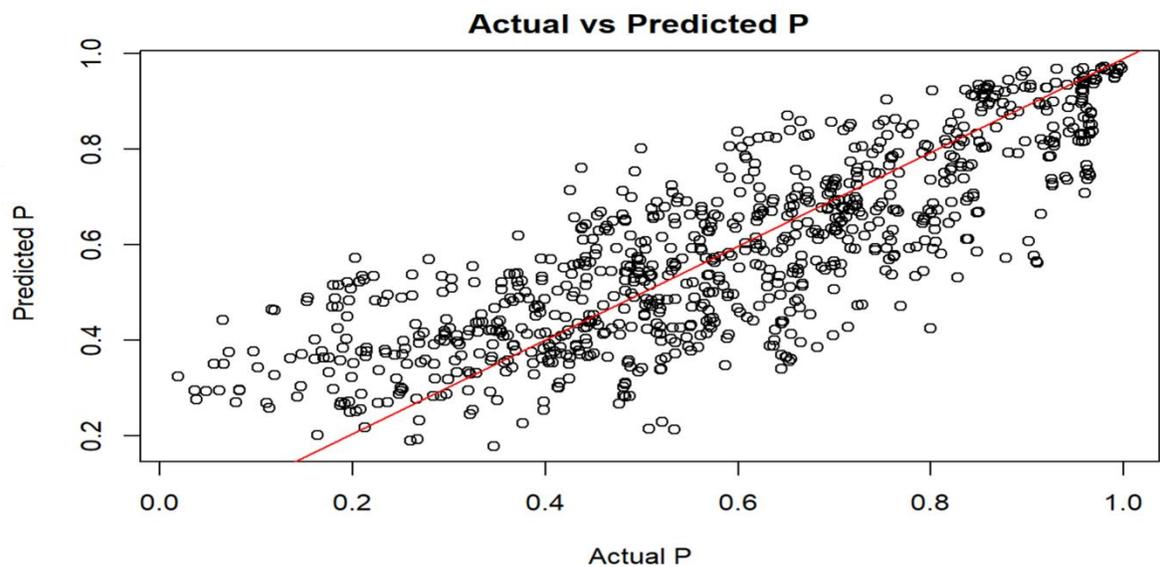


Figure 31: Actual versus normalised predicted values of power production using linear regression

#### 6- Finding the correlation between actual and normalised predicted values:

```
145 cor(dat.test$PE,preds.nn)
146
147
148 * ~~~~
      [,1]
[1,] 0.8143401
```

Figure 32: The correlation between the actual and predicted power values after normalisation

As shown in Figure 31, the actual and normalised predicted values of power production are almost aligned, with a correlation of 0.8143401 (see Figure 32). This indicates the validity of the prediction model, as the normalisation leads to a higher correlation value. However, there is a need for better data collection over a more extended period (more than one month) to get better results, better correlation and to improve the model.

### 4.5 Checking the Model's Accuracy

To check the linear regression model's accuracy and compare it to the neural network model's accuracy, several equations were used, such as MSE, RMSE, MAE, MAPE and RSE. The following calculations represent how these formulas were applied; the program used for calculations is Microsoft Excel.

Table 17 shows the first ten cells in the dataset, and the calculation steps performed on linear regression actual and predicted values of the Minitab Model prediction equation. Table 18 shows the first ten cells and the calculation steps performed on actual and predicted values of the RStudio linear regression prediction model. Conversely, Table 19 shows the first ten cells in the dataset and the calculation steps performed on neural network predicted values.

Table 18: The first ten actual and predicted values and the accuracy calculations performed on the Minitab linear regression prediction model

<b>Actual</b>	<b>Predicted</b>	$(\hat{y}_i - y_i)^2$	$ \hat{y}_i - y_i $	$\frac{ \hat{y}_i - y_i }{y_i}$
759.414	709.517	2489.753	49.89742667	0.065705
749.322	711.321	1444.099	38.00130169	0.050714
741.117	713.2027	779.2184	27.91448412	0.037665
729.388	711.8151	308.8202	17.57328075	0.024093
721.239	713.8382	54.77378	7.40093125	0.010261
727.931	712.8749	226.692	15.0562947	0.020684
713.177	680.9272	1040.048	32.24977686	0.04522
699.658	662.8487	1354.931	36.80939257	0.052611
699.255	672.2112	731.3653	27.0437664	0.038675
704.387	670.8832	1122.474	33.50335046	0.047564

Table 19: The first ten actual and predicted values and the accuracy calculations performed on RStudio linear regression prediction model

<b>Actual P (MW)</b>	<b>Predicted P (MW)</b>	$(\hat{y}_i - y_i)^2$	$ \hat{y}_i - y_i $	$\frac{ \hat{y}_i - y_i }{y_i}$
759.414	710.119	2430.033	49.29536	0.064912
749.322	711.9283	1398.312	37.39401	0.049904
741.117	713.8284	744.678	27.28879	0.036821
729.388	712.4239	287.7958	16.96454	0.023259
721.239	714.4687	45.83851	6.770415	0.009387
727.931	713.4908	208.5244	14.44037	0.019838
713.177	681.3098	1015.518	31.86719	0.044683
699.658	663.0806	1337.914	36.57751	0.052279
699.255	672.5113	715.2242	26.74368	0.038246
704.387	671.163	1103.807	33.2236	0.047167

Table 20: The first ten actual values and the accuracy calculations performed on neural network predicted values

Actual P (MW)	Predicted P (MW)	$(\hat{y}_i - y_i)^2$	$ \hat{y}_i - y_i $	$\frac{ \hat{y}_i - y_i }{y_i}$
759.414	713.8255	2078.348	45.5889	0.060032
749.322	714.4657	1214.983	34.8566	0.046517
741.117	647.1592	8828.106	93.958	0.126779
729.388	651.6537	6042.684	77.7347	0.106575
721.239	647.5138	5435.42	73.7253	0.10222
727.931	606.2076	14816.63	121.7236	0.167219
713.177	607.5342	11160.4	105.6428	0.14813
699.658	605.3715	8889.963	94.2866	0.134761
699.255	599.5757	9935.963	99.6793	0.142551
704.387	639.4315	4219.165	64.9551	0.092215

Table 20 shows the equations, descriptions and sample of calculations performed on linear regression and neural network predicted values.

Table 21: Model accuracy checking for linear regression models and neural network predicted values

	MSE	RMSE	MAD	MAPE	RSE
<b>Description</b>	Mean square error	The Root Mean Square Error	Mean Absolute Deviation	Symmetric Mean Absolute Percent Error	Residual Standard Error
<b>Equation</b>	$\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}$ (39)	$\sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}}$ (40)	$\frac{\sum_{t=1}^n  \hat{y}_t - y_t }{n}$ (41)	$\frac{\sum_{t=1}^n \frac{ \hat{y}_t - y_t }{y_t}}{n} * 100$ (42)	$\sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n-2}}$ (43)
<b>Minitab model accuracy checking</b>	3291.631	57.37273	46.38558	6.71976	57.39266
<b>RStudio model accuracy checking</b>	3291.706	57.37339	46.31997	6.71649	57.39331
<b>Neural Network model</b>	9774.297	98.86505	81.95122	11.64243	98.89938

accuracy checking					
-------------------	--	--	--	--	--

The following is a sample of Microsoft Excel calculations performed on ANN values, where  $n = 2881$ ,  $t$  is the number of the observation,  $y_i$  is the actual value, and  $\hat{y}_i$  is the predicted power production value.

$$- \text{MSE} = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n} = \frac{(713.8255 - 759.414)^2}{2881} + \frac{(711.321 - 749.322)^2}{2881} + \dots = 3291.631$$

As shown in Table 20, the Minitab linear regression prediction model had a lower value of MSE (3291.631) than the ANN model (9774.297) and is almost identical to the RStudio model (3291.706), which means that the Minitab model is the most accurate model (almost the same results obtained for the RStudio model). Accordingly, in the Minitab linear regression model, there is a better match between the actual and predicted values. In addition, RMSE was found to be 57.37273 for the Minitab linear regression prediction model, 57.37339 for the RStudio model and 98.86505 for the ANN prediction model, which leads to the conclusion that the linear regression model fits the data better than the ANN model. In addition, high RMSE values indicate that the model doesn't test the data perfectly, and that if the model was applied to information outside the sample, it would give a low accurate prediction result. Moreover, MAD values were also higher for the ANN model (81.95122) than for both linear regression models, which illustrates that the ANN prediction model has a broader spread of the data points further from the mean value than the linear regression models (46.38558 for Minitab and 46.31997 for RStudio). In addition, MAPE calculations resulted in lower values for the linear regression prediction models compared to 11.64243 for the ANN model. This suggests that the linear regression model is more accurate because the difference between the actual and predicted values is smaller (only 6.7%). Finally, the value of RSE was lower for the linear regression models (57.39266 for Minitab and 57.39331 for RStudio) compared to

98.89938 for the ANN prediction model, indicating that the linear regression models better fit the data and are more accurate than the ANN model. Generally, the linear regression model is found to be better compared to the ANN model. Moreover, Minitab and RStudio provided results that differ only slightly. For MSE, RMSE and RSE calculations, Minitab was slightly lower and more accurate than RStudio. However, for MAD and MAPE calculations, RStudio was found to be slightly lower and more accurate than Minitab. Generally, the values of MSE and RMSE for all the measures are very similar to the Tüfekci (2014) research because the dataset is limited to one month only; the model would be more accurate if the data were collected over a more extended period.

## CHAPTER V

### CONCLUSION AND RECOMMENDATIONS

#### 5.1 Conclusion

A statistical approach is presented to develop a model that estimates the electrical power production of a CCPP in the UAE based on ambient parameters. Unlike thermodynamical modelling, which takes too much effort and time, this study doesn't involve a considerable number of assumptions, which lead to unsatisfactory and unreliable results, and doesn't consider nonlinear modelling. Instead, it uses statistical modelling and ANNs as MLMs to find electrical power production and draw valuable interferences. The thermodynamics system of the study consists of a GT, steam turbine and HRSG. In addition, the data of this paper is limited to one month only with no missing values. However, the sampling variability is low and the model presented robustness to the outliers, which ensures a good prediction.

The purposes of this study include finding the effect of predictors (AT, RH, AP, V), which potentially affect the target variable (P) based on linear regression modelling and ANNs. The main results are as follows:

Firstly, the power data was plotted versus time to check possible trends, and there was an almost steady trend in the overall weekly power production. However, it was increasing slightly during the last weeks.

Secondly, the highest variable effect on power production was shown by V, with a correlation of 0.707. This effect was found to have a directly proportional relationship with power production. The second highest correlation with the power production was found to be with the and, which was -0.3007187 and which had a negative relationship with the power. Moreover, AT

and AP were found to have a weak relationship with the target variable and correlations of 0.28177428 and -0.1449736, respectively. In addition, each input parameter's influence was discussed separately, and a final model equation to predict the electrical power production was obtained with an R-sq value of 53.49%. The best dataset based on a weekly time frame was found to be for week one, which had a value of R-sq (82.16%).

Using RStudio software, ANNs were generated to anticipate power production using AI, and the model prediction equation was found. Accordingly, the correlation between the actual and predicted values of power production was found to be 0.814, which indicates the validity of the prediction model. In addition, the error value of the model was found to be low after normalisation, 17.768191.

Finally, the accuracy of each model (linear regression prediction models and ANN's prediction model) was calculated. Generally, the linear regression models found using Minitab and RStudio were more accurate than the ANN model with MSE (3291.631), RMSE (57.37273), MAD (46.38558) and MAPE (6.719765). Moreover, MSE, RMSE and RSE calculations showed that Minitab is slightly more accurate than RStudio software.

## 5.2 Recommendations

The predictive model developed by this study needs future adjustments to be made more accurate and to increase the precision of its results. This can be done by:

- Expanding the scope of data of this study to at least more than six months so that the R-squared values are higher and the correlation of variables with the power is more accurate.
- This model can be applied to estimate approximately the P of CCPP in the UAE based on future anticipated weather conditions. This requires, as an example, a forecasted AT if indicated precisely by the UAE meteorology institutes. Thus, this model can be used to predict future performance if the ambient parameters are forecasted accurately.
- Future work is needed to predict power production based on ambient parameters for other types of plants, especially nuclear power plant production, as the tendency towards considering these power production plants is higher nowadays.
- More studies can be accompanied to study the effect of the most influential ambient parameter (V in this study) on different types of power plants.
- Another approach that can be conducted in future work is to study the correlation of power production with the atmospheric pressure and relative humidity more intensely, as the association of these input variables with the target variable was found to be low in this study.

## REFERENCES

- Amell, A.A. & Cadavid, F.J. (2002). Influence of the relative humidity on the air cooling thermal load in gas turbine power plant. *Applied Thermal Engineering*. Pergamon, vol. 22 (13), pp. 1529–1533.
- Amozegar, M. & Khorasani, K. (2016). An ensemble of dynamic neural network identifiers for fault detection and isolation of gas turbine engines. *Neural Networks*, vol. 76, pp. 106–121.
- Github, K. (2013). *normalization - How to normalize data to 0-1 range? - Cross Validated* [online]. [Accessed 29 May 2022]. Available at:  
<https://stats.stackexchange.com/questions/70801/how-to-normalize-data-to-0-1-range>.
- Aranda, A., Ferreira, G., Mainar-Toledo, M.D., Scarpellini, S. & Llera Sastresa, E. (2012). Multiple regression models to predict the annual energy consumption in the Spanish banking sector. *Energy and Buildings*, vol. 49, pp. 380–387.
- Arrieta, F. R. P. & Lora, E. E. S. (2005). Influence of ambient temperature on combined-cycle power-plant performance. *Applied Energy*, vol. 80(3), pp. 261–272.
- Azadeh, A., Saberi, M. & Seraj, O. (2010). An integrated fuzzy regression algorithm for energy consumption estimation with non-stationary data: A case study of Iran. *Energy*, vol. 35 (6), pp. 2351–2366.
- Banhidarrah, A.K., Al-Sumaiti, A.S., Wescoat, J.L. & Nguyen, H.T. (2020). Electricity-water usage for sustainable development: An analysis of United Arab Emirates farms. *Energy Policy*, vol. 147, p. 111823.

Barigozzi, G., Perdichizzi, A., Gritti, C. & Guaiatelli, I. (2015). Techno-economic analysis of gas turbine inlet air cooling for combined cycle power plant for different climatic conditions. *Applied Thermal Engineering*, vol. 82, pp. 57–67.

Berend, D. & Kontorovich, A. (2013). A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, vol. 83 (4), pp. 1254–1259.

Byrne, B.M. (2011). *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. New York, United States: Taylor & Francis Group [online]. [Accessed 15 May 2022] .Available at:

<http://ebookcentral.proquest.com/lib/buidae/detail.action?docID=957904>.

Campora, U., Cravero, C. & Zaccone, R. (2018). Marine gas turbine monitoring and diagnostics by simulation and pattern recognition. *International Journal of Naval Architecture and Ocean Engineering*, vol. 10 (5), pp. 617–628.

Chabot, E. & D'Arras, H. (2010). Neural computation and particle accelerators: research, technology and applications LK. *New York: Nova Science Publishers (Neuroscience Research Progress Series)* [online]. [Accessed 03May 2022]. Available at:

<http://site.ebrary.com/id/10659167>.

Chai, T. & Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, vol. 7 (3), pp. 1247–1250.

Charles, H., Eccleston, C.H. & March, F. T. (2014). Global environmental policy: concepts,

principles, and practice. [online]. [Accessed 11 April 2022]. Available at:

<http://www.crcnetbase.com/isbn/9781439896068>.

Che, J., Wang, J. & Wang, G. (2012). An adaptive fuzzy combination model based on self-organizing map and support vector regression for electric load forecasting. *Energy*, vol. 37 (1), pp. 657–664.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in medicine*, vol. 19 (22), pp. 3127–3131.

De, R.K., Mandal, D.P.& Ghosh, A. (2010). Machine interpretation of patterns: image analysis and data mining. *World Scientific (Statistical Science and Interdisciplinary Research)* [online]. [Accessed 10 April 2022]. Available at: <http://site.ebrary.com/id/10479715>.

Dehghani Samani, A. (2018). Combined cycle power plant with indirect dry cooling tower forecasting using artificial neural network. *Decision Science Letters*, vol. 7 (2), pp. 131–142.

Dutta, S. & Ghosh, S. (2021). Predicting electrical power output in a combined cycle power plant – a statistical approach. *International Journal of Energy Engineering*, vol 11 (2), pp. 16–26.

Eckardt, D. T. (2014). Gas Turbine Powerhouse: the Development of the Power Generation Gas Turbine at BBC - ABB - Alstom [online]. [Accessed 21 April 2022]. Available at: <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=753982>.

Ekonomou, L. (2010). Greek long-term energy consumption prediction using artificial neural

networks. *Energy*, vol. 35 (2), pp. 512–517.

Franco, A. & Casarosa, C. (2002). On some perspectives for increasing the efficiency of combined cycle power plants. *Applied Thermal Engineering*, vol. 22 (13), pp. 1501–1518.

Goodwin, P. & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, vol. 15 (4), pp. 405–408.

Granit, J. & Löfgren, R. (2010). *Water and Energy Linkages in the Middle East*, pp. 1–20.

Hadi, A.S. & Chatterjee, S. (2012). *Regression Analysis by Example*. Somerset, United States: John Wiley & Sons, [online]. [Accessed 11 April 2022]. Available at:  
<http://ebookcentral.proquest.com/lib/buidae/detail.action?docID=918623>.

El Hadik, A.A. (1990). The impact of atmospheric conditions on gas turbine performance. *Journal of Engineering for Gas Turbines and Power*, vol. 112 (4), pp. 590–596.

Han, J., Kamber, M. & Pei, J. (2012). *Classifications: Basic Concepts and Methods*. Data mining concepts and techniques, vol. 4 (1), pp. 270–280.

Hashmi, M.B., Majid, M.A.A. & Lemma, T.A. (2020). Combined effect of inlet air cooling and fouling on performance of variable geometry industrial gas turbines. *Alexandria Engineering Journal*, vol. 59 (3), pp. 1811–1821.

Ho, M., Obbard, E., Burr, P.A. & Yeoh, G. (2019). A review on the development of nuclear power reactors. *Energy Procedia*. Elsevier, vol. 160, pp. 459–466.

Hocking, R.R. (2013). *Methods and Applications of Linear Models: Regression and the Analysis*

*of Variance* [online]. [Accessed 02 February 2022]. Available at:

<http://ebookcentral.proquest.com/lib/buidae/detail.action?docID=1224710>.

Holcomb, Z. (2016). *Fundamentals of Descriptive Statistics*, Routledge. pp. 1–11.

Hub, C. (2022). *Hectopascals to Bars Conversion (hPa to bar) - Inch Calculator* [online].

[Accessed 30 May 2022]. Available at: <https://www.inchcalculator.com/convert/hectopascal-to-bar/>.

Ibrahim, T.K., Kamil, M., Awad, O.I., Rahman, M.M., Najafi, G., Basrawi, F., Abd Alla, A.N. & Mamat, R. (2017). The optimum performance of the combined cycle power plant: A comprehensive review. *Renewable and Sustainable Energy Reviews*, vol. 79, pp. 459–474.

Kunming, C. & Zhou, X. (2012). International Conference on Manufacturing Engineering and Process. *Manufacturing engineering and process: selected, peer reviewed papers from the 2012*, Kunming, China LK [online]. [Accessed 05 May 2022]. Available at:<https://buid.on.worldcat.org/oclc/849739486>.

Jitesh, J.T. (2021). *Structural Equation Modelling*. Application for research and practice with amos and R. Springer. pp. 1-124.

Kaplan, S. (2009). *Power Plant Characteristics and Costs* [online]. [Accessed 22 April 2022]. Available at: <http://ebookcentral.proquest.com/lib/buidae/detail.action?docID=3018748>.

Kavaklioglu, K. (2011). Modeling and prediction of Turkey's electricity consumption using support vector regression. *Applied Energy*, vol. 88 (1), pp. 368–375.

Kaviri, A.G., Jaafar, M.N.M., Lazim, T.M. & Barzegaravval, H. (2013). Exergoenvironmental optimization of heat recovery steam generators in combined cycle power plant through energy and exergy analysis. *Energy Conversion and Management*, vol. 67, pp. 27–33.

Kehlhofer, R., Rukes, B., Hannemann, F. & Stirnimann, F. (2009). *Combined-Cycle Gas & Steam Turbine Power Plants*. PennWell Books [online]. [Accessed 12 May 2022]. Available at: <https://books.google.ae/books?id=aLcfEAAAQBAJ>.

Kesgin, U. & Heperkan, H. (2005). Simulation of thermodynamic systems using soft computing techniques. *International Journal of Energy Research*, vol. 29 (7), pp. 581–611.

Khair, U., Fahmi, H., Hakim, S. & Rahim, R. (2017). Forecasting error calculation with mean absolute deviation and mean absolute percentage error. *Journal of Physics: Conference Series* [online]. vol. 930 (1). [Accessed 23 February 2022]. Available at: <https://iopscience.iop.org/article/10.1088/1742-6596/930/1/012002/meta>

Kloeckner, J., Machado, P.L., Rodrigues, Á.L. & Costa, J. (2019). Covariance table: a fast automatic spatial continuity mapping. *Computers & Geosciences*. vol. 130, pp. 94–104.

Kong, M., Li, D. & Zhang, D. (2019). Research on the application of improved least square method in linear fitting. *IOP Conference Series: Earth and Environmental Science* [online]. vol. 252 (5). [Accessed 11 February 2022]. Available at: <https://iopscience.iop.org/article/10.1088/1755-1315/252/5/052158/meta>

Konno, H. & Koshizuka, T. (2005). Mean-absolute deviation model. *IIE Transactions*, vol. 37 (10), pp. 893–900.

Kotowicz, J. & Brzęczek, M. (2018). Analysis of increasing efficiency of modern combined cycle power plant: a case study. *Energy*, vol. 153, pp. 90–99.

Kotowicz, J., Job, M. & Brzeczek, M. (2015). The characteristics of ultramodern combined cycle power plants. *Energy*, vol. 92, pp. 197–211.

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, vol. 26 (2), pp. 195–197.

Krose, B. & Smagt, P. (2011). An introduction to neural networks, University of Amsterdam. pp. 18–21.

Krugel, F., Pélérini-Issac, M. & Benali, H. (2002). Estimating the effective degrees of freedom in univariate multiple regression analysis. *Medical Image Analysis*, vol. 6 (1), pp. 63–75.

Leung, P.C.M. & Lee, E.W.M. (2013). Estimation of electrical power consumption in subway station design by intelligent approach. *Applied Energy*, vol. 101, pp. 634–643.

Li, M. (2015). Moving beyond the linear regression model. *Journal of Management*, vol. 41 (1), pp. 71–98.

Cao, L., Sai, L., Fu, J., and Duan, X. (2020). Artificial neural network potential for gold clusters. *Chinese Physics B* [online]. [Accessed 02 April 2022]. Available at: <http://cpb.iphy.ac.cn>.

Reese, R., & Bhatia, A. (2018). Natural Language Processing with Java: Techniques for Building Machine Learning and Neural Network Models for NLP, 2nd Edition.

<https://buid.on.worldcat.org/oclc/1048799933>. Birmingham: Packt Publishing Ltd [online].

[Accessed 15 April 2022]. Available at:

<https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5485028>.

Madu, K.E. & Nwankwo, E.I. (2018). Evaluation of the impact of high exhaust temperature in steam turbine operation [online]. vol. 1(1). [Accessed 11 February 2022]. Available at:

[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3272624](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3272624)

Maier, H.R. & Dandy, G.C. (1996). The use of artificial neural networks for the prediction of water quality parameters. *Water Resources Research*, vol. 32 (4), pp. 1013–1022.

Marmolin, H. (1986). Subjective MSE Measures. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16 (3), pp. 486–489.

Nelder, J.A. & Wedderburn, R.W. (2019). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, vol. 135 (3), pp. 370–384.

Mohseni, O., Stefan, H.G. & Erickson, T.R. (1998). A nonlinear regression model for weekly stream temperatures. *Water Resources Research*, vol. 34 (10), pp. 2685–2692.

Morano, P. & Tajani, F. (2014). Least median of squares regression and minimum volume ellipsoid estimator for outliers detection in housing appraisal. *International Journal of Business Intelligence and Data Mining*, vol. 9 (2), pp. 91–111.

Motawa, I. & Oladokun, M.G. (2015). Structural equation modelling of energy consumption in buildings. *International Journal of Energy Sector Management*, vol. 9 (4), pp. 435–450.

Mustafa, A. (2012). Clean energies development in built environment. *World Journal of Science*,

*Technology and Sustainable Development*, vol. 9 (1), pp. 45–63.

Najah, A., Binti, F., Abdulmohsin, H., Khaleel, R., Ming, C., Shabbir, M., Ehteram, M. & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, vol. 578, p. 124084.

Nick, T.G. (2007). Descriptive Statistics. *Topics in Biostatistics: Methods in Molecular Biology*, vol. 404 (Humana Press) [online]. [Accessed 13 January 2022]. Available at: [https://doi.org/10.1007/978-1-59745-530-5\\_3](https://doi.org/10.1007/978-1-59745-530-5_3).

Onoroh, F., Ogonnaya, M. & Onochie, U.P. (2020). Modeling and simulation of the effect of moisture content and ambient temperature on gas turbine power plant performance in Ughelli, Nigeria. *Nigerian Journal of Technology*, vol. 39 (1), pp. 182–188.

Park, S.K. & Kim, T.S. (2006). Comparison between pressurized design and ambient pressure design of hybrid solid oxide fuel cell–gas turbine systems. *Journal of Power Sources*, vol. 163 (1), pp. 490–499.

Park, Y., Choi, M., Kim, K., Li, X., Jung, C., Na, S. & Choi, G. (2020). Prediction of operating characteristics for industrial gas turbine combustor using an optimized artificial neural network. *Energy*, vol. 213, p. 118769.

Pérez-Vicente, S. & Expósito Ruiz, M. (2009). Descriptive statistics. *Allergologia et Immunopathologia*, vol. 37 (6), pp. 314–320.

Petridis, G. & Nicolau, D. (2011). *Nuclear Power Plants*. Hauppauge, United States: Nova

Science Publishers, Incorporated [online]. [Accessed 09 April 2022]. Available at:

<http://ebookcentral.proquest.com/lib/buidae/detail.action?docID=3019813>.

Pugesek, B., Tomer, A. & Eye, A. (2003). *Structural equation modeling: applications in ecological and evolutionary biology*. Cambridge ; Cambridge University Press [online].

[Accessed 05 January 2022]. Available at: <https://doi.org/10.1017/CBO9780511542138>.

Raja, A.K., Srivastava, A.P. & Dwivedi, M. (2006). *Power plant engineering*. New Delhi: New Age International (P) Ltd., Publishers [online]. [Accessed 10 January 2022]. Available at:

<http://site.ebrary.com/id/10323349>.

Al Rashdi, M.R., El Tokhi, M., Alaabed, S., El Mowafi, W. & Arabi, A.A. (2020). Rare earth elements around the barakah nuclear power plant, UAE. *Natural Resources Research*, vol. 29 (6), pp. 4149–4160.

Rhinehart, R.R. (2016). *Nonlinear Regression Modeling for Engineering Applications: Modeling, Model Validation, and Enabling Design of Experiments*. Chicester, United Kingdom [online].

[Accessed 15 January 2022]. Available at:

<http://ebookcentral.proquest.com/lib/buidae/detail.action?docID=4622919>.

Rönkkö, M., Aalto, E., Tenhunen, H. & Aguirre-Urreta, M.I. (2022). Eight simple guidelines for improved understanding of transformations and nonlinear effects. *Organizational Research Methods*, vol. 25 (1), pp. 48–87.

Saadeh, R., Burqan, A. & El-Ajou, A. (2022). Reliable solutions to fractional Lane-Emden equations via Laplace transform and residual error function. *Alexandria Engineering Journal*,

vol. 61 (12), pp. 10551–10562.

Saghafifar, M. & Gadalla, M. (2016). Thermo-economic analysis of conventional combined cycle hybridization: United Arab Emirates case study. *Energy Conversion and Management*, vol. 111, pp. 358–374.

Said, M.I., Steiner, M., Siefer, G. & Arab, A.H. (2020). Maximum power output prediction of HCPV FLATCON® module using an ANN approach. *Renewable Energy*, vol. 152, pp. 1274–1283.

Sankhye, S. & Hu, G. (2020). Machine learning methods for quality prediction in production. *Logistics*, vol. 4 (4), p. 35.

Searle, S.R. & Gruber, M.H. (2016). *Linear Models*. Newark, United States [online]. [Accessed 20 January 2022]. Available at:

<http://ebookcentral.proquest.com/lib/buidae/detail.action?docID=4696766>.

Sushil, S. & Tandon, D.J. (2019). A study on the impact of transport and power infrastructure development on the economic growth of United Arab Emirates (UAE). *Journal of Management*, vol. 6 (2).

Talukder, P. & Soori, P.K. (2015). Integration of parabolic trough collectors with natural gas combined cycle power plants in United Arab Emirates. *2015 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE)*, pp. 62–69.

Tayman, J. & Swanson, D.A. (1999). On the validity of MAPE as a measure of population

forecast accuracy. *Population Research and Policy Review*, vol. 18 (4), pp. 299–322.

Thatcher, M.J. (2007). Modelling changes to electricity demand load duration curves as a consequence of predicted climate change for Australia. *Energy*, vol. 32 (9), pp. 1647–1659.

Thompson, P.A. (1990). An MSE statistic for comparing forecast accuracy across series. *International Journal of Forecasting*, vol. 6 (2), pp. 219–227.

TMI Staff & Contributors. (2017). *How ambient temperature affects gas turbine types* [online]. [Accessed 30 May 2022]. Available at: <https://www.turbomachinerymag.com/view/how-ambient-temperature-affects-gas-turbine-types>.

Treyer, K. & Bauer, C. (2016). The environmental footprint of UAE's electricity sector: combining life cycle assessment and scenario modeling. *Renewable and Sustainable Energy Reviews*, vol. 55, pp. 1234–1247.

Tso, G.K. & Yau, K.K. (2007). Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. *Energy*, vol. 32 (9), pp. 1761–1768.

Tu, J.V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*. Pergamon, vol. 49 (11), pp. 1225–1231.

Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power and Energy Systems*, vol. 60, pp. 126–140.

Verdict. (2020). *Cascade Combined-Cycle Gas Turbine (CCGT) Power Plant, Canada* [online]. [Accessed 15 June 2022]. Available at: <https://www.power-technology.com/projects/cascade-combined-cycle-gas-turbine-ccgt-power-plant-alberta/>.

Wallach, D. & Goffinet, B. (1989). Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling*, vol. 44 (3–4), pp. 299–306.

Walter, S. (2013). Generalized linear mixed models: modern concepts. *Methods and Applications* [online]. [Accessed 01 June 2022]. Available at: <http://www.crcnetbase.com/isbn/9781439815137>.

Whittle, P. (2010). Neural nets and chaotic carriers. *Advances in Computer Science and Engineering* [online]. vol. 5. [Accessed 11 May 2022]. Available at: <http://site.ebrary.com/id/10479953>.

Willmott, C.J. & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, vol. 30 (1), pp. 79–82.

Zhang, G. & Chow, S. (2010). Standard error estimation in stationary multivariate time series models using residual-based bootstrap procedures. *Individual Pathways of Change: Statistical Models for Analyzing Learning and Development*. pp. 169–182.

# APPENDICES

## Appendix A: Data Analysis Details

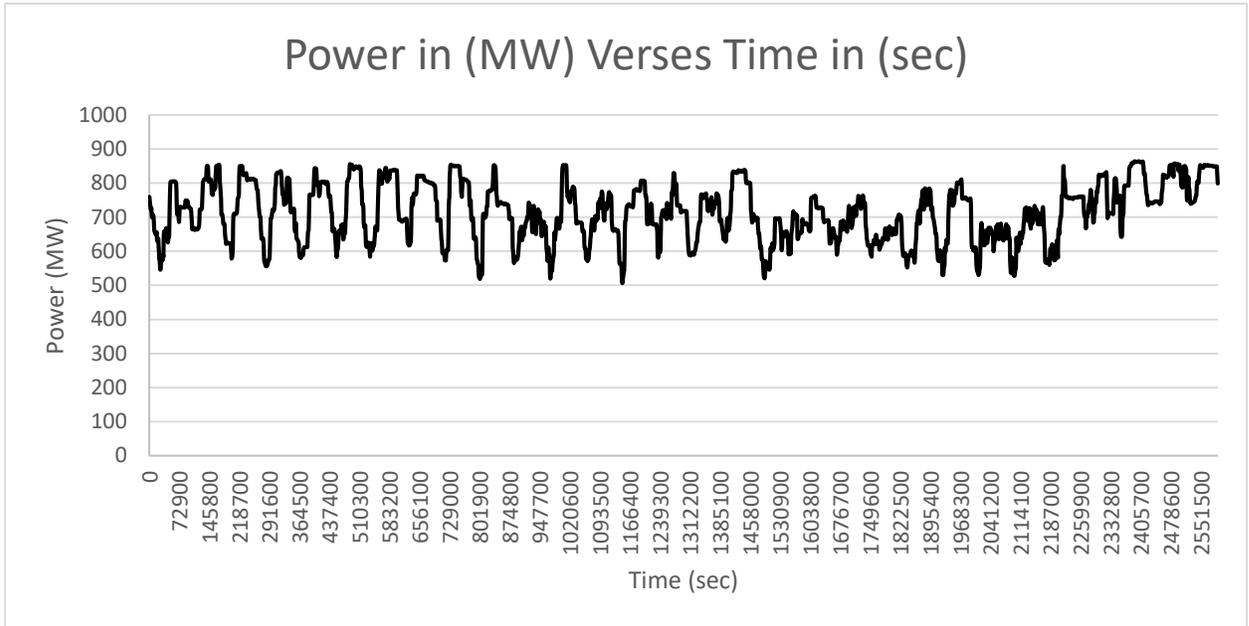


Figure A. 1: Power in (MW) versus time in (seconds)

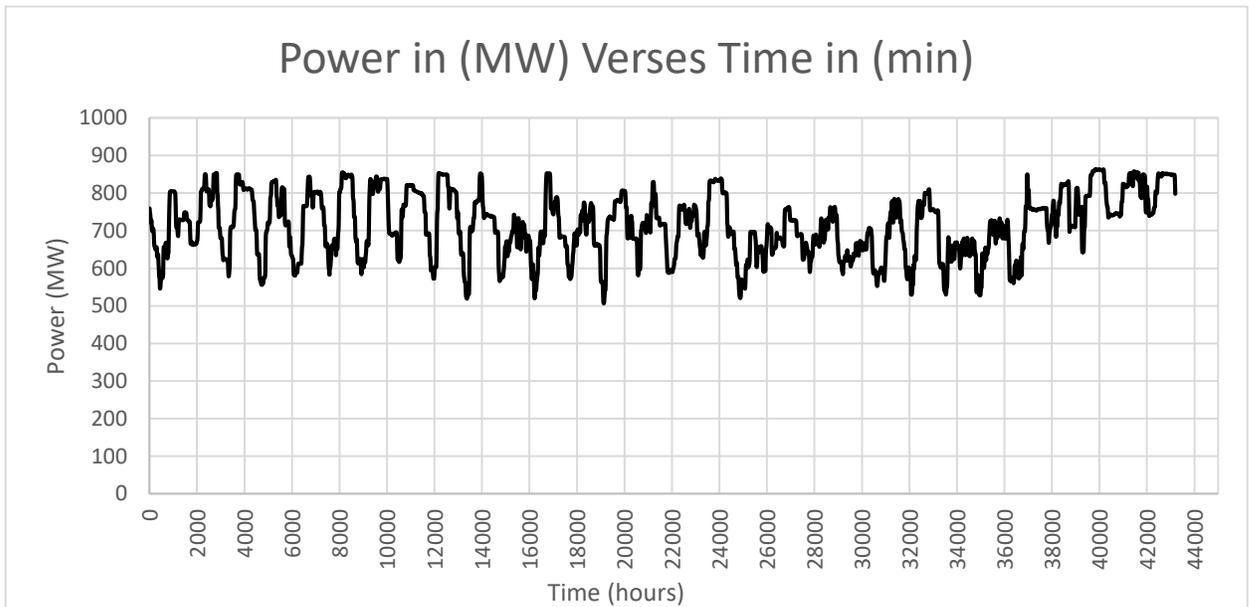


Figure A. 2: Power in (MW) versus time in (min)

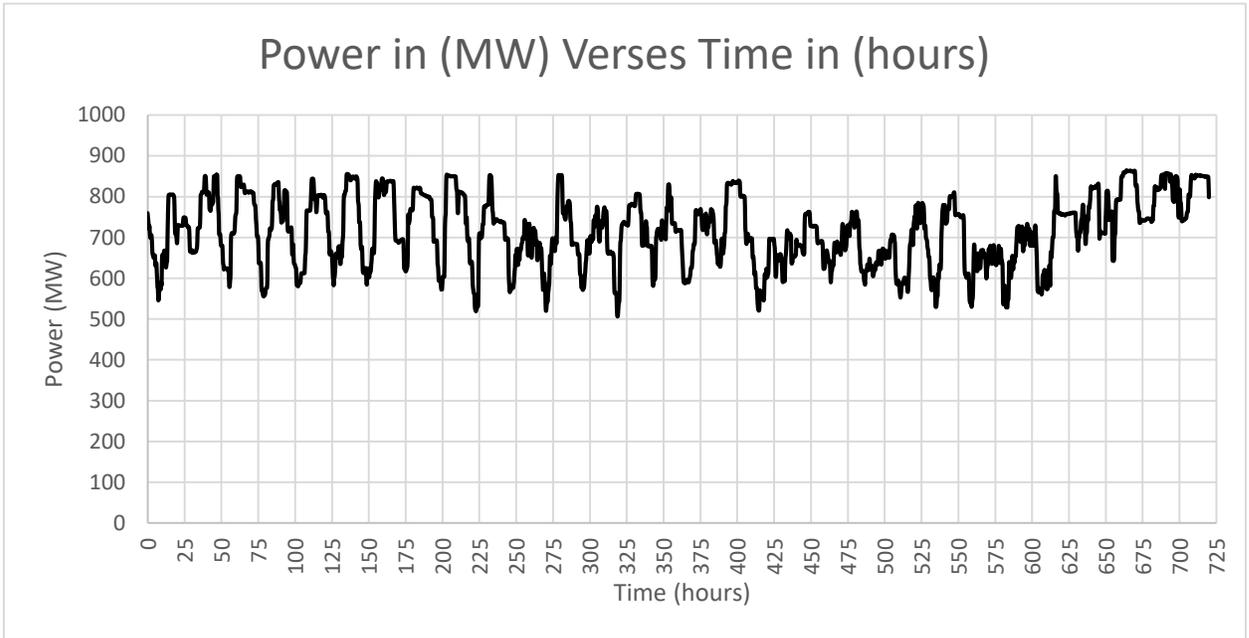


Figure A. 3: Power in (MW) versus time in (hours)

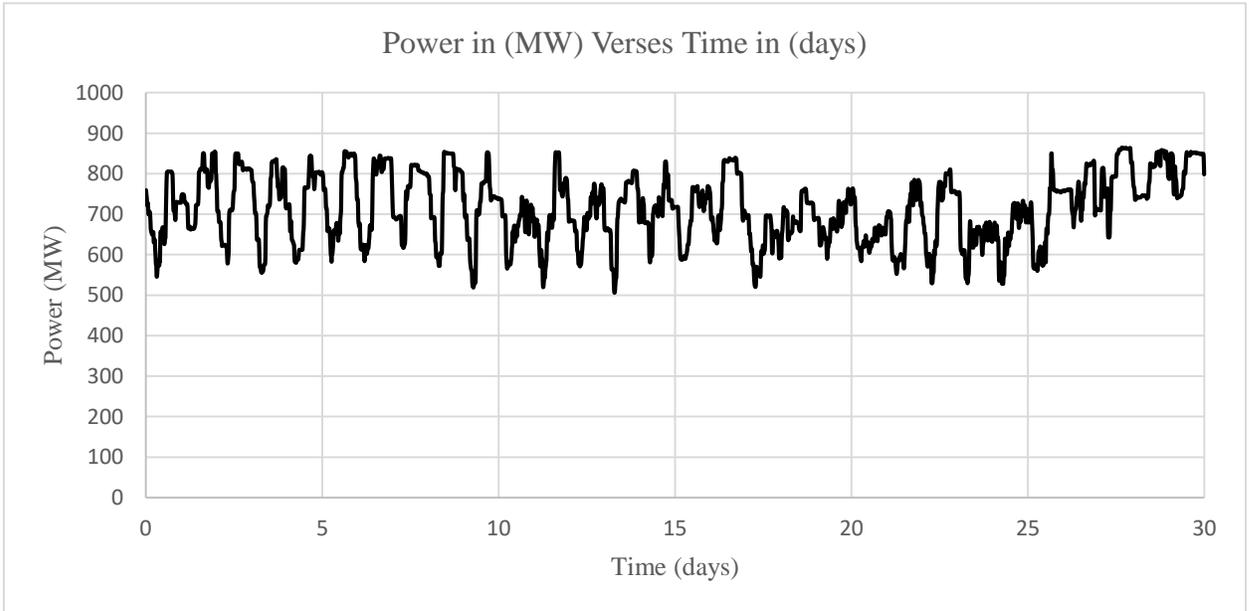


Figure A. 4: Power in (MW) versus time in (days)

Table A. 1: Week 1 power prediction model equation

<b>Week 1 linear regression model equation</b>
$P1 = 5609 - 2.49 AT1 - 1.443 RH1 - 5080 AP1 + 4519 V1$

Table A. 2: Week 1 model summary using Minitab

S	R-sq	R-sq(adj)	R-sq(pred)
36.4831	82.16%	82.05%	81.92%

Table A. 3: Coefficients table for week 1 prediction model

Term	Coef	SE Coef	T-Value	P-Value	VIF
<b>Constant</b>	5609	1145	4.90	0.000	
<b>AT1</b>	-2.49	1.08	-2.30	0.021	2.12
<b>RH1</b>	-1.443	0.197	-7.33	0.000	1.78
<b>AP1</b>	-5080	1155	-4.40	0.000	1.25
<b>V1</b>	4519	112	40.48	0.000	1.59

Table A. 4: Analysis of variance for the week 1 prediction model

Source	DF	Adj SS	Adj MS	F-Value	P-Value
<b>Regression</b>	4	4088193	1022048	767.87	0.000
<b>AT1</b>	1	7070	7070	5.31	0.021
<b>RH1</b>	1	71439	71439	53.67	0.000
<b>AP1</b>	1	25741	25741	19.34	0.000
<b>V1</b>	1	2181388	2181388	1638.88	0.000
<b>Error</b>	667	887790	1331		

<b>Total</b>	671	4975983			
--------------	-----	---------	--	--	--

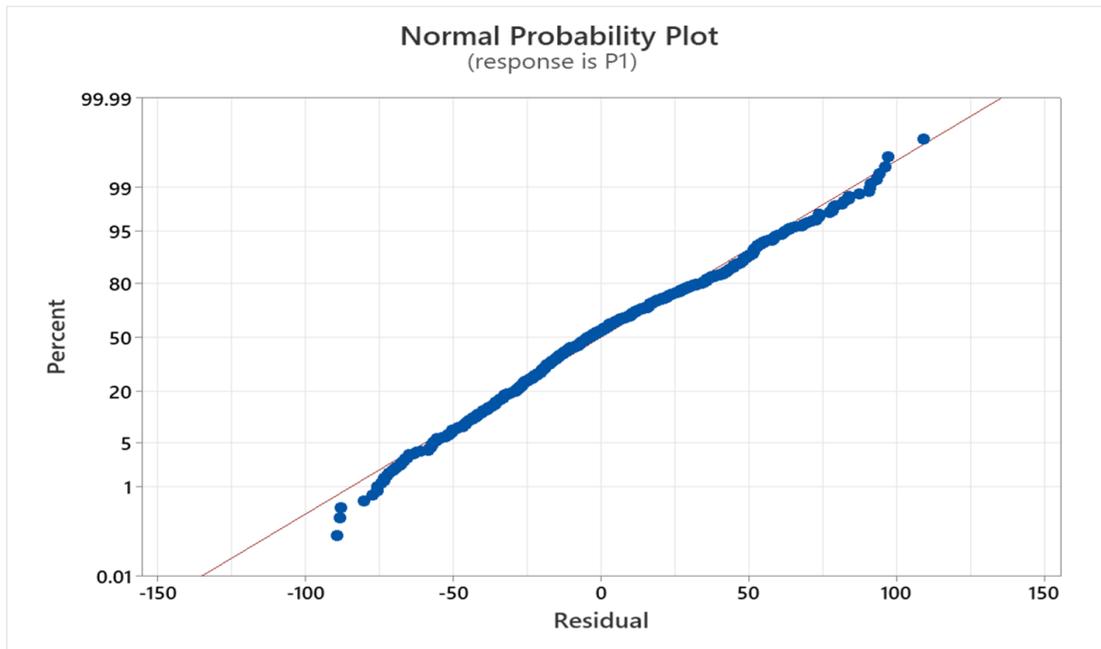


Figure A. 5: Normality plot for the week 1 prediction model

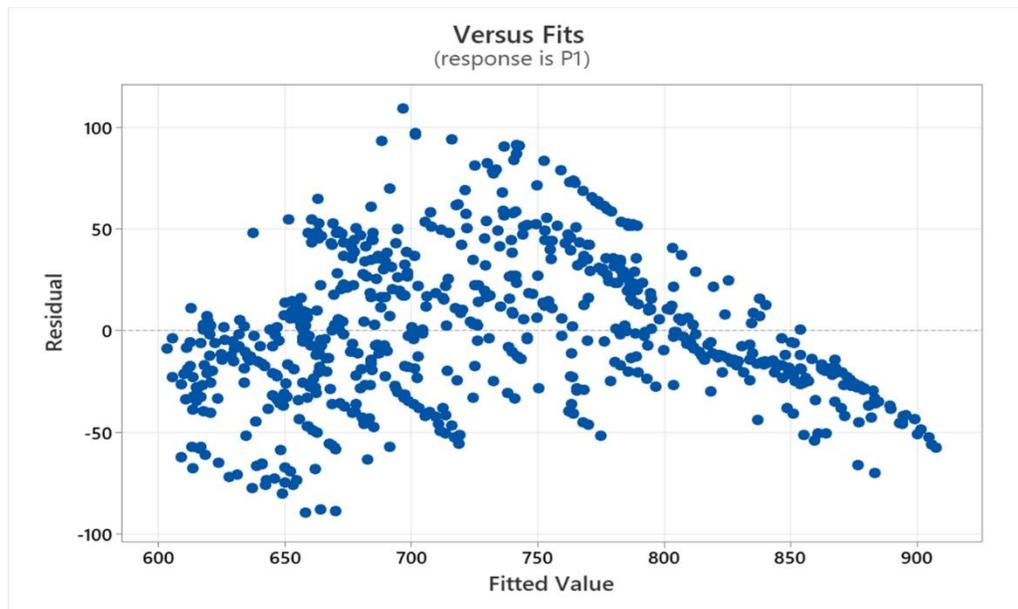


Figure A. 6: Residuals plot versus fitted values for the week 1 prediction model

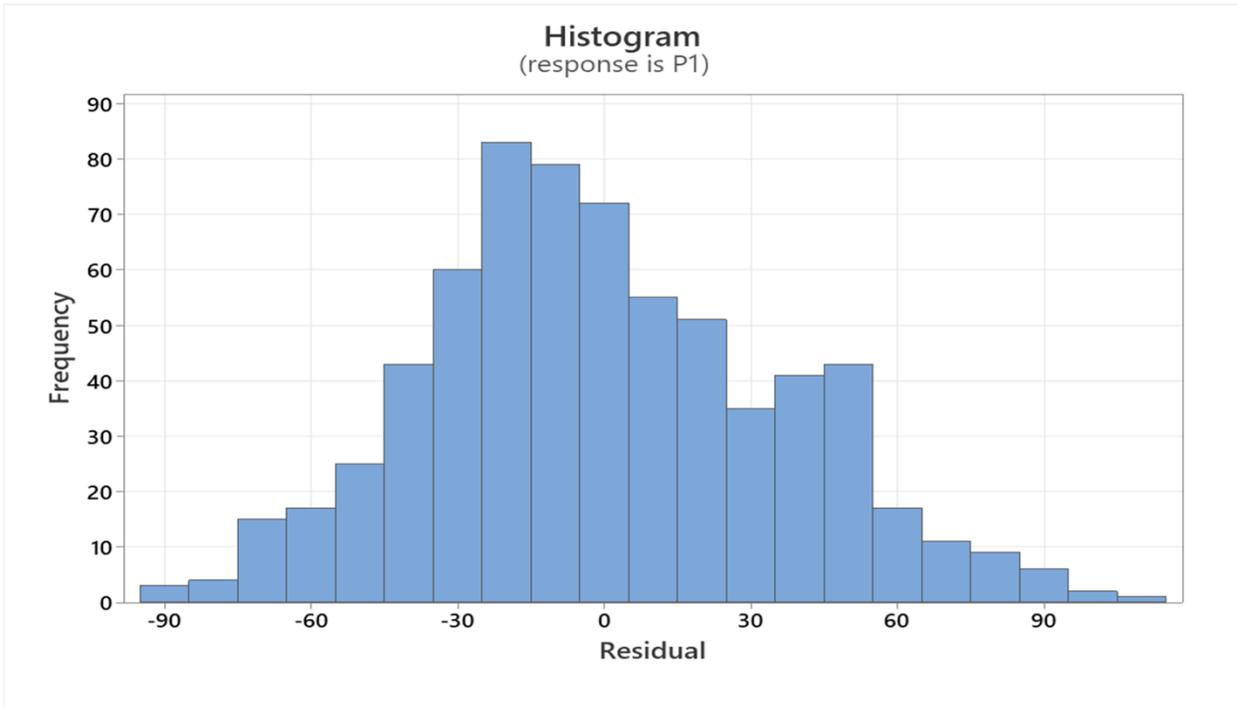


Table A. 5: Week 2 power prediction model equation

<b>Week 2 linear regression model equation</b>
$P2 = 12217 + 23.72 AT2 + 1.460 RH2 - 12584 AP2 + 2828 V2$

Table A. 6: Week 2 model summary using Minitab

<b>S</b>	<b>R-sq</b>	<b>R-sq(adj)</b>	<b>R-sq(pred)</b>
54.6057	54.67%	54.43%	54.14%

Table A. 7: Coefficients table for week 2 prediction model

<b>Term</b>	<b>Coef</b>	<b>SE Coef</b>	<b>T-Value</b>	<b>P-Value</b>	<b>VIF</b>

<b>Constant</b>	12217	1020	11.97	0.000	
<b>AT12</b>	23.72	2.51	9.43	0.000	4.57
<b>RH2</b>	1.460	0.502	2.91	0.004	4.36
<b>AP2</b>	-12584	989	-12.72	0.000	1.08
<b>V2</b>	2828	180	15.73	0.000	1.27

Table A. 8: Analysis of variance for week 2 data subset

Source	DF	Adj SS	Adj MS	F-Value	P-Value
<b>Regression</b>	4	2657638	664409	222.82	0.000
<b>AT12</b>	1	265410	265410	89.01	0.000
<b>RH2</b>	1	25198	25198	8.45	0.004
<b>AP2</b>	1	482455	482455	161.80	0.000
<b>V2</b>	1	737671	737671	247.39	0.000
<b>Error</b>	739	2203535	2982		
<b>Total</b>	743	4861173			

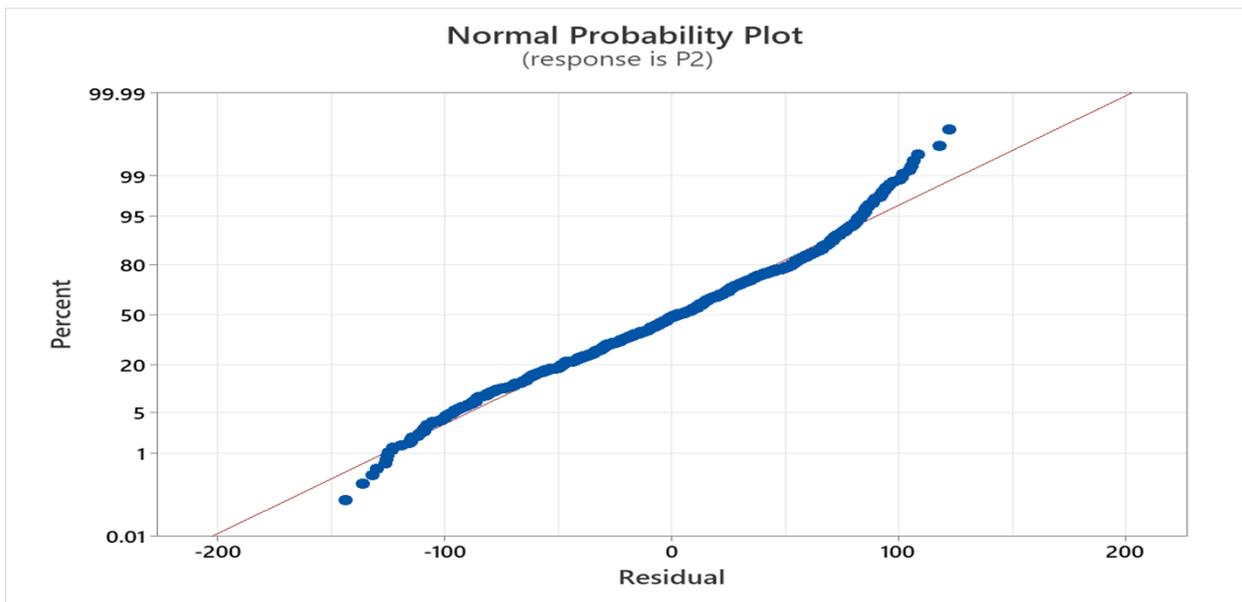


Figure A. 8: Normality plot for week 2 prediction model

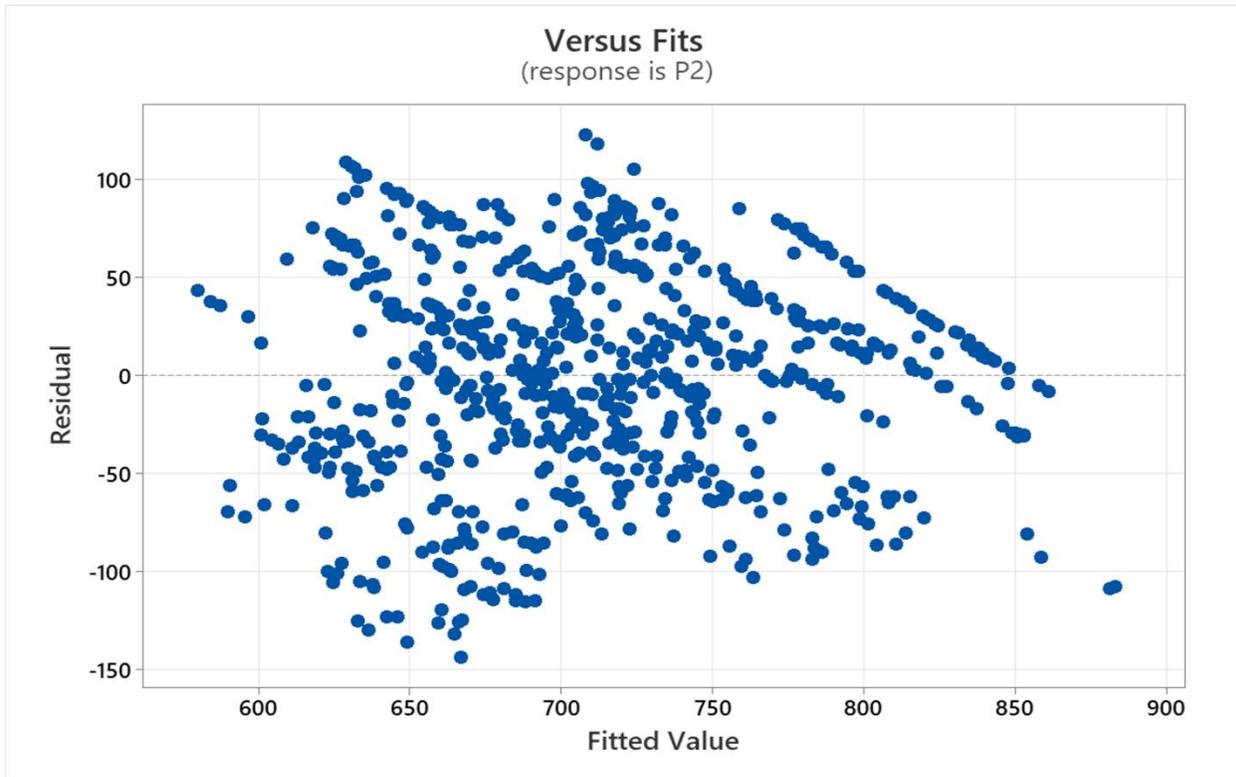


Figure A. 9: Residuals plot versus fitted values for week 2 prediction model

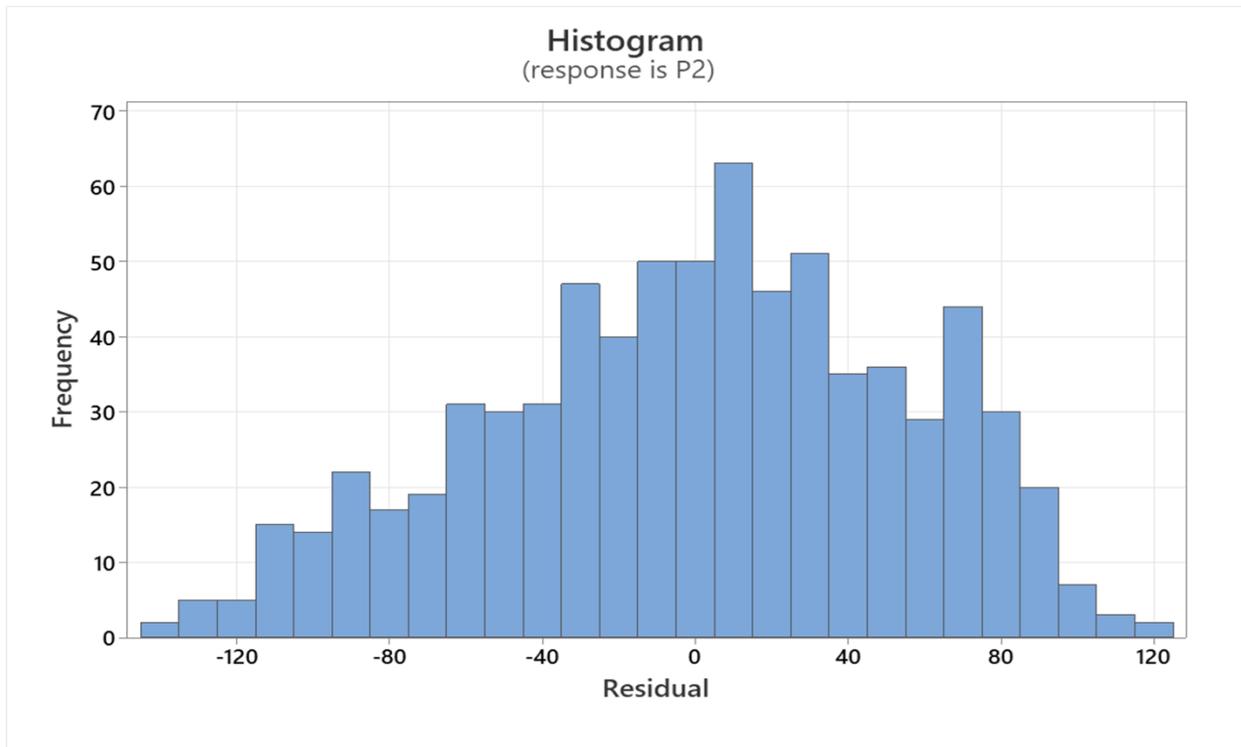


Figure A. 10: Histogram of residuals for week 2 data subset prediction model

Table A. 9: Week 3 power prediction model equation

<b>Week 3 linear regression model equation</b>
$P3 = 10407 + 11.25 AT3 + 1.870 RH3 - 10467 AP3 + 3336 V3$

Table A. 10: Week 3 model summary using Minitab

<b>S</b>	<b>R-sq</b>	<b>R-sq(adj)</b>	<b>R-sq(pred)</b>
52.0316	42.97%	42.63%	42.25%

Table A. 11: Coefficients table for week 3 prediction model

<b>Term</b>	<b>Coef</b>	<b>SE Coef</b>	<b>T-Value</b>	<b>P-Value</b>	<b>VIF</b>
<b>Constant</b>	10407	1509	6.90	0.000	
<b>AT3</b>	11.25	3.31	3.40	0.001	8.37
<b>RH3</b>	1.870	0.710	2.63	0.009	7.58
<b>AP3</b>	-10467	1510	-6.93	0.000	1.14
<b>V3</b>	3336	224	14.88	0.000	1.43

Table A. 12: Analysis of variance for week 3 data subset

<b>Source</b>	<b>DF</b>	<b>Adj SS</b>	<b>Adj MS</b>	<b>F-Value</b>	<b>P-Value</b>
<b>Regression</b>	4	1360469	340117	125.63	0.000
<b>AT3</b>	1	31240	31240	11.54	0.001
<b>RH3</b>	1	18784	18784	6.94	0.009
<b>AP3</b>	1	130006	130006	48.02	0.000
<b>V3</b>	1	599735	599735	221.53	0.000
<b>Error</b>	667	1805758	2707		
<b>Total</b>	671	3166227			

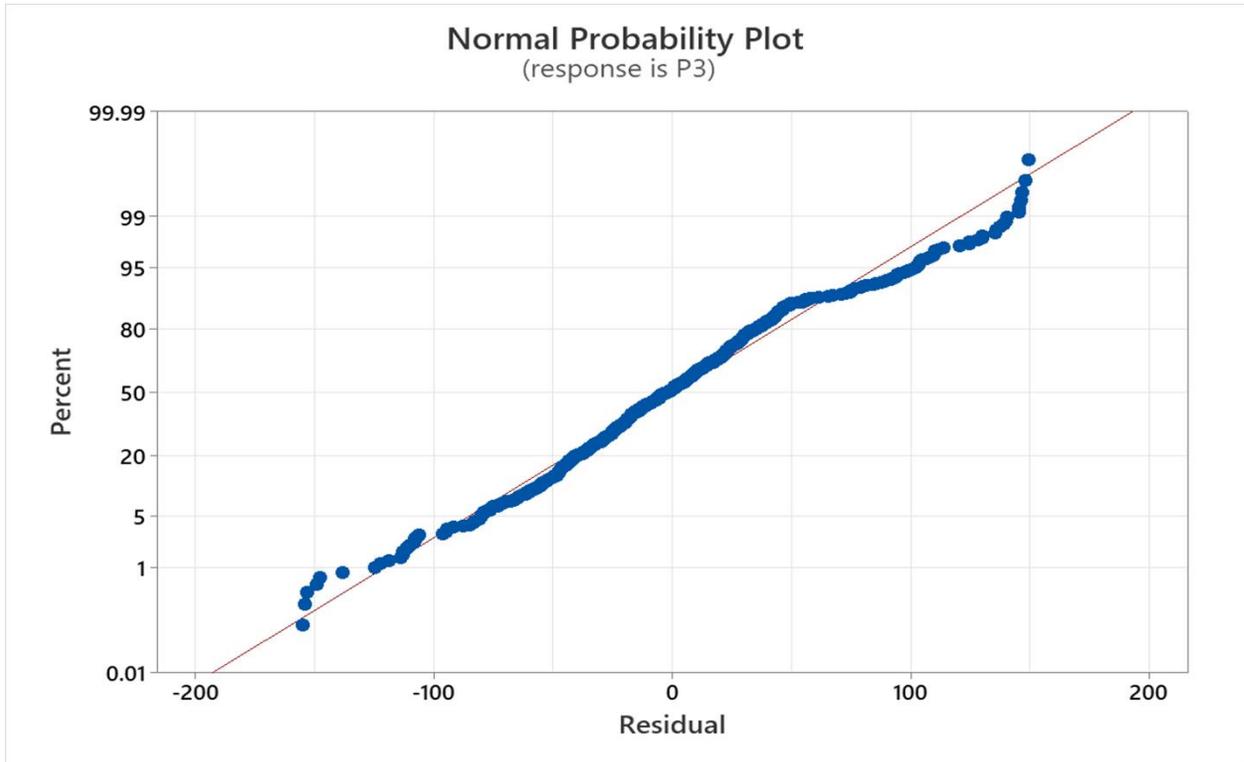


Figure A. 11: Normality plot for week 3 prediction model

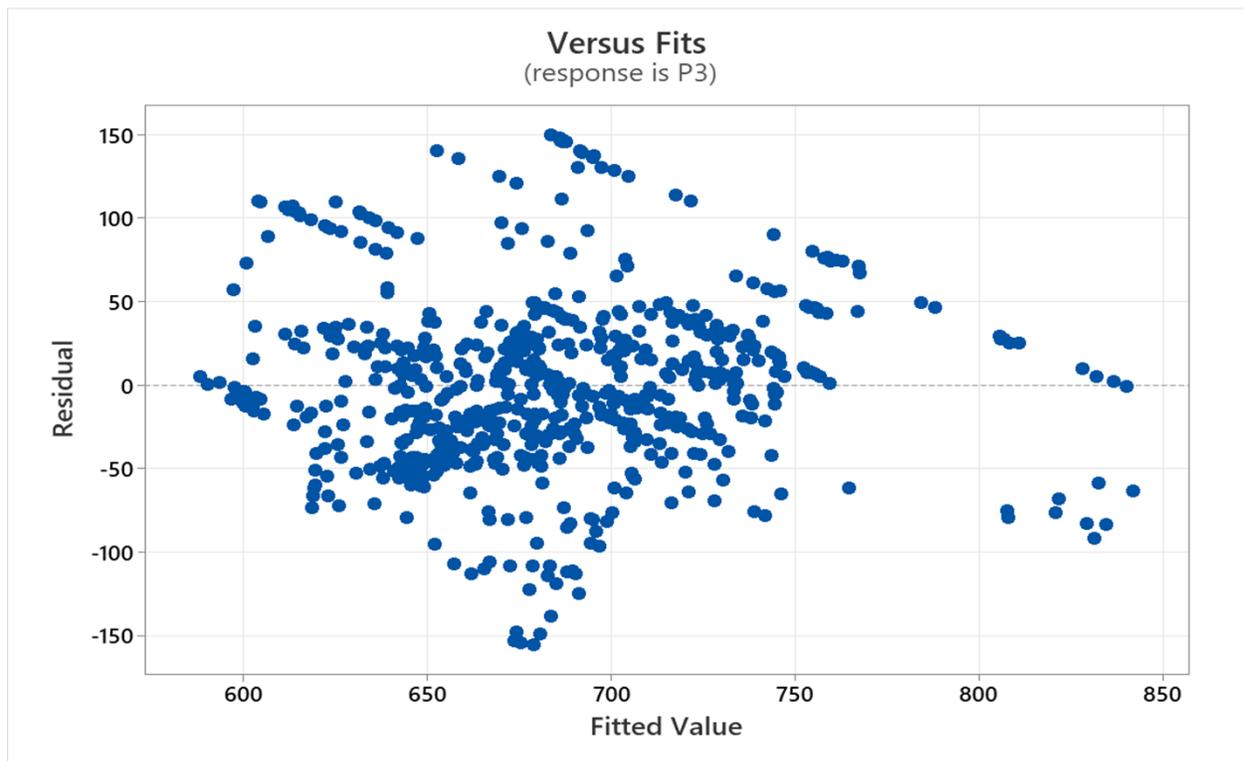


Figure A. 12: Residuals plot versus fitted values for week 3 prediction model

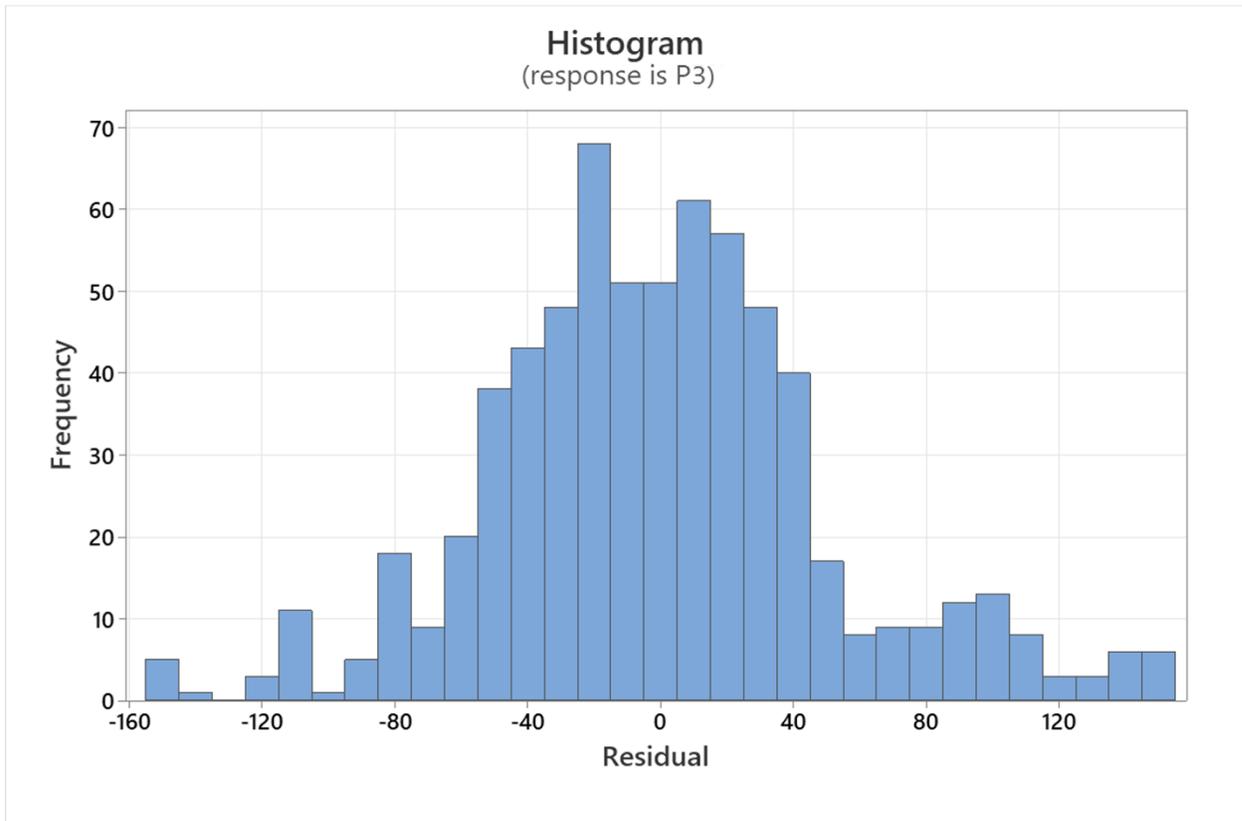


Figure A. 13: Histogram of residuals for week 3 data subset prediction model

Table A. 13: Week 4 power prediction model equation

<b>Week 4 linear regression model equation</b>
$P4 = -11992 - 13.67 AT4 - 0.860 RH4 + 12671 AP4 + 5196 V4$

Table A. 14: Week 4 model summary using Minitab

<b>S</b>	<b>R-sq</b>	<b>R-sq(adj)</b>	<b>R-sq(pred)</b>
52.6483	62.27%	62.05%	61.61%

Table A. 15: Coefficients table for week 4 prediction model

Term	Coef	SE Coef	T-Value	P-Value	VIF
<b>Constant</b>	-11992	1348	-8.90	0.000	
<b>AT4</b>	-13.67	1.67	-8.17	0.000	3.26
<b>RH4</b>	-0.860	0.456	-1.89	0.060	2.84
<b>AP4</b>	12671	1303	9.72	0.000	1.49
<b>V4</b>	5196	157	32.99	0.000	1.47

Table A. 16: Analysis of variance for week 4 data subset

Source	DF	Adj SS	Adj MS	F-Value	P-Value
<b>Regression</b>	4	3051960	762990	275.26	0.000
<b>AT4</b>	1	185224	185224	66.82	0.000
<b>RH4</b>	1	9858	9858	3.56	0.060
<b>AP4</b>	1	261943	261943	94.50	0.000
<b>V4</b>	1	3017523	3017523	1088.63	0.000
<b>Error</b>	667	1848822	2772		
<b>Total</b>	671	4900781			

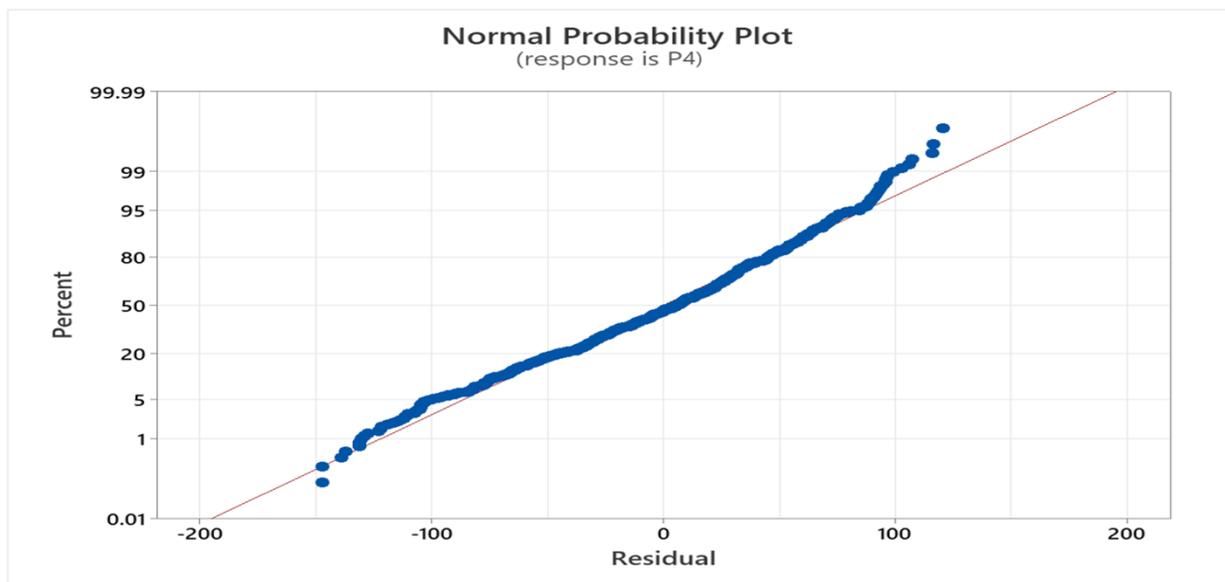


Figure A. 14: Normality plot for week 4 prediction model

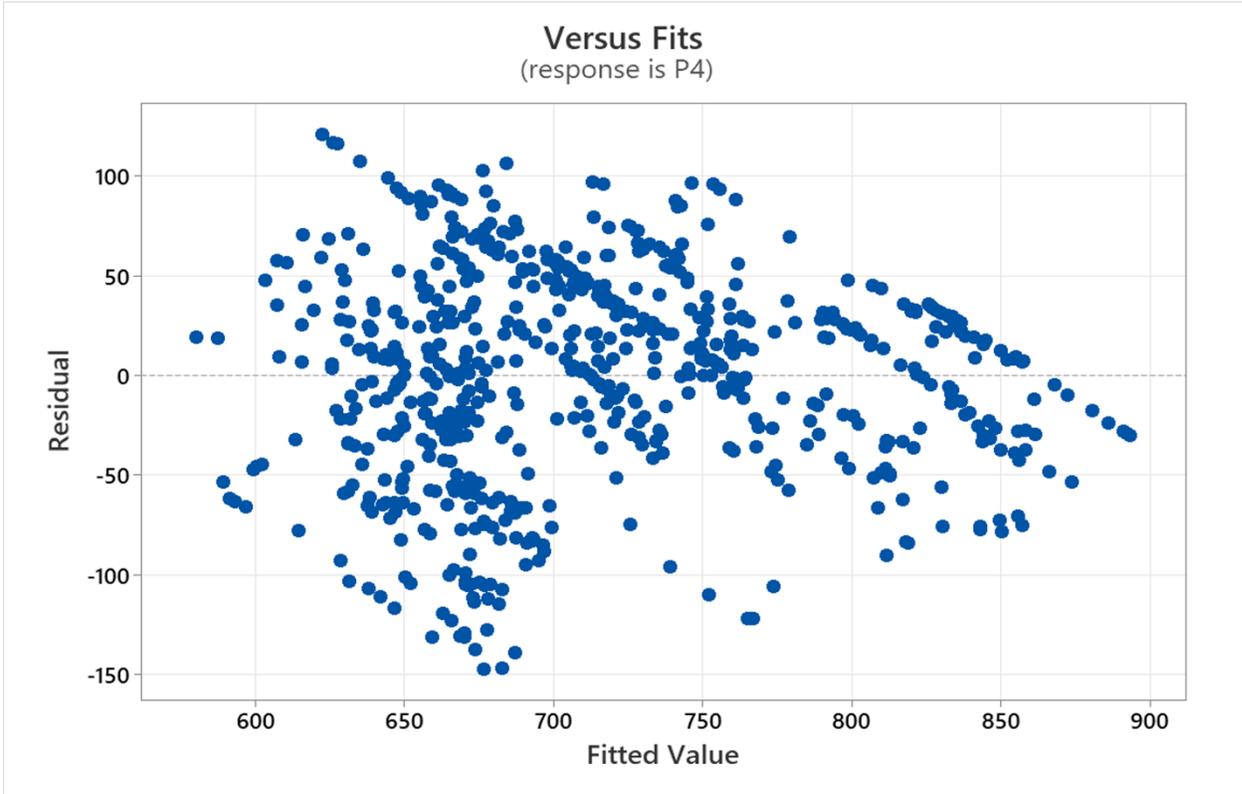


Figure A. 15: Residuals plot versus fitted values for week 4 prediction model

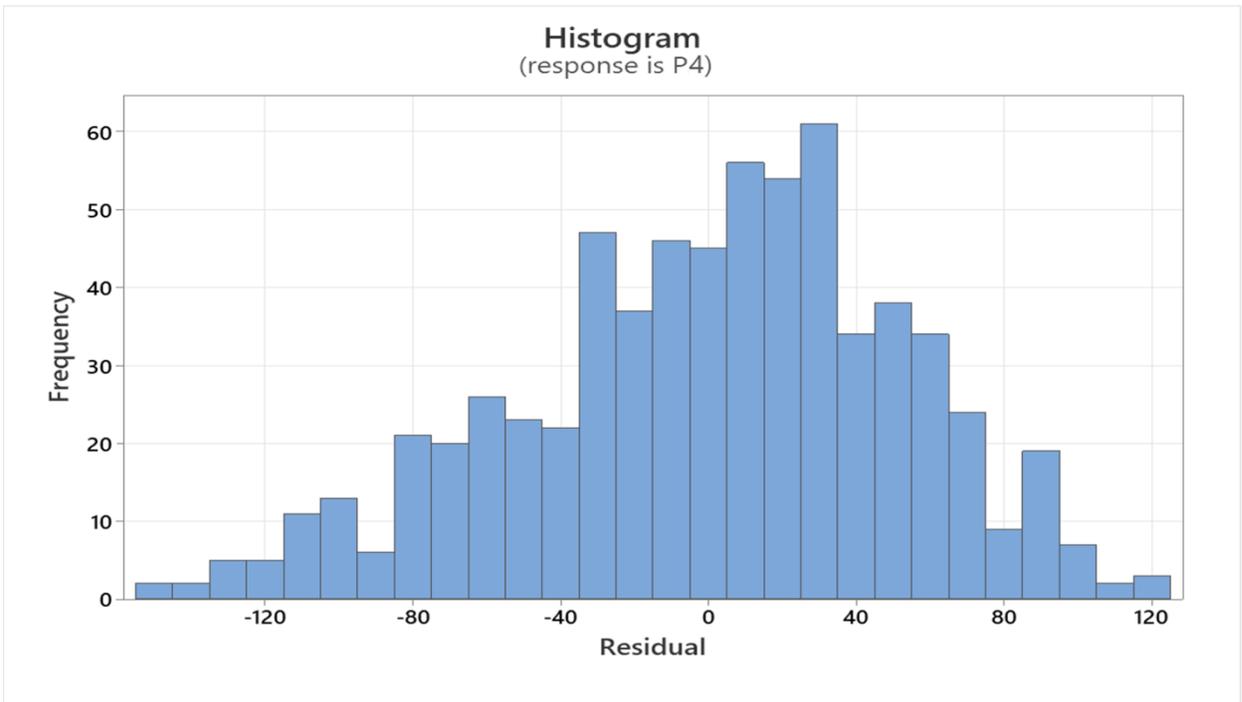


Figure A. 16: Histogram of residuals for week 4 data subset prediction model

Table A. 17: Week 5 power prediction model equation

<b>Week 5 linear regression model equation</b>
$P5 = 6211 - 1.26 AT5 - 0.408 RH5 - 5586 AP5 + 3012 V5$

Table A. 18: Week 5 model summary using Minitab

S	R-sq	R-sq(adj)	R-sq(pred)
12.0862	91.26%	90.96%	90.44%

Table A. 19: Coefficients table for week 5 prediction model

Term	Coef	SE Coef	T-Value	P-Value	VIF
<b>Constant</b>	6211	1443	4.30	0.000	
<b>AT5</b>	-1.26	1.37	-0.91	0.362	8.82
<b>RH5</b>	-0.408	0.240	-1.70	0.092	6.28
<b>AP5</b>	-5586	1445	-3.87	0.000	1.77
<b>V5</b>	3012	136	22.10	0.000	2.45

Table A. 20: Analysis of variance for week 5 data subset

Source	DF	Adj SS	Adj MS	F-Value	P-Value
<b>Regression</b>	4	176911	44227.7	302.77	0.000
<b>AT5</b>	1	122	122.1	0.84	0.362
<b>RH5</b>	1	422	421.8	2.89	0.092
<b>AP5</b>	1	2183	2183.2	14.95	0.000

<b>V5</b>	1	71376	71376.2	488.62	0.000
<b>Error</b>	116	16945	146.1		
<b>Total</b>	120	193856			

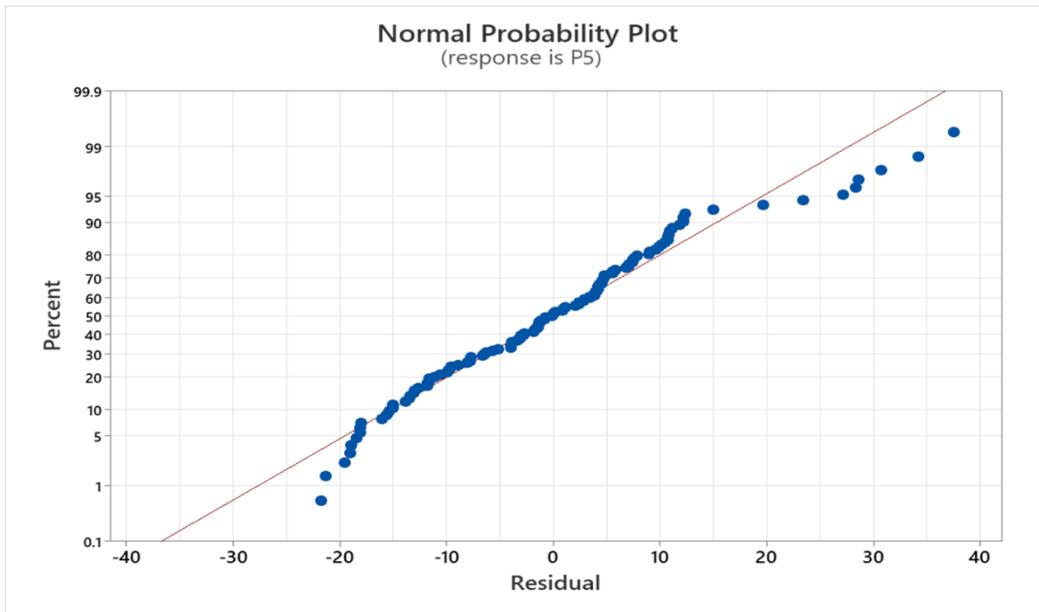


Figure A. 17: Normality plot for week 5 prediction model

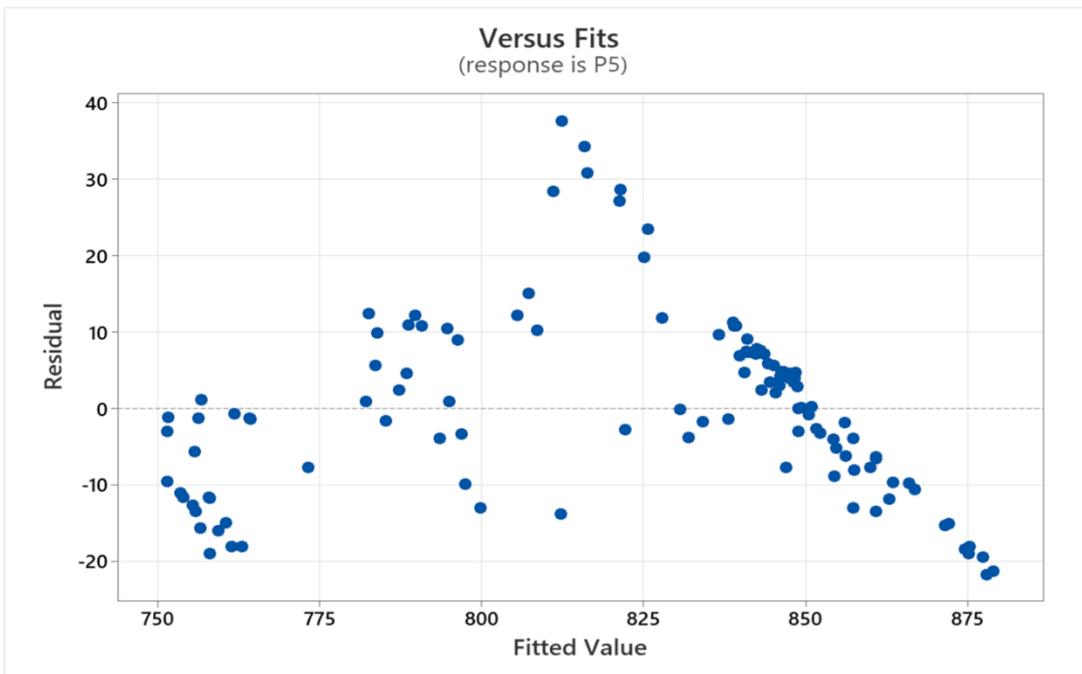


Figure A. 18: Residuals plot versus fitted values for week 5 prediction model

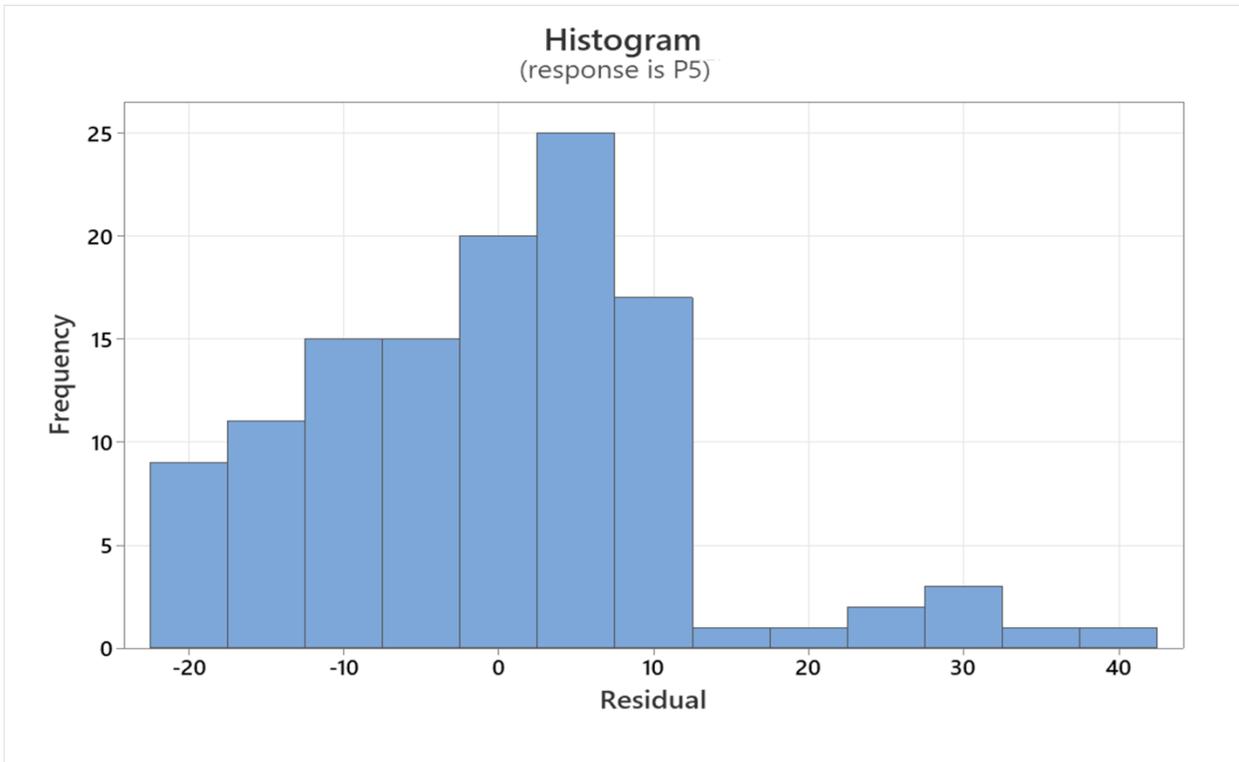


Figure A. 19: Histogram of residuals for week 5 data subset prediction model

Table A. 21: A sample of the actual data after normalisation

	<b>P</b>	<b>AT</b>	<b>RH</b>	<b>AP</b>	<b>V</b>
<b>1</b>	0.7067173	0.7067173	0.7067173	0.7067173	0.7067173
<b>2</b>	0.6785367	0.6785367	0.6785367	0.6785367	0.6785367
<b>3</b>	0.6556253	0.6556253	0.6556253	0.6556253	0.6556253
<b>4</b>	0.6228745	0.6228745	0.6228745	0.6228745	0.6228745
<b>5</b>	0.6001189	0.6001189	0.6001189	0.6001189	0.6001189
<b>6</b>	0.6188055	0.6188055	0.6188055	0.6188055	0.6188055
<b>7</b>	0.5776068	0.5776068	0.5776068	0.5776068	0.5776068
<b>8</b>	0.5398574	0.5398574	0.5398574	0.5398574	0.5398574
<b>9</b>	0.5387318	0.5387318	0.5387318	0.5387318	0.5387318

<b>10</b>	0.5530610	0.5530610	0.5530610	0.5530610	0.5530610
-----------	-----------	-----------	-----------	-----------	-----------

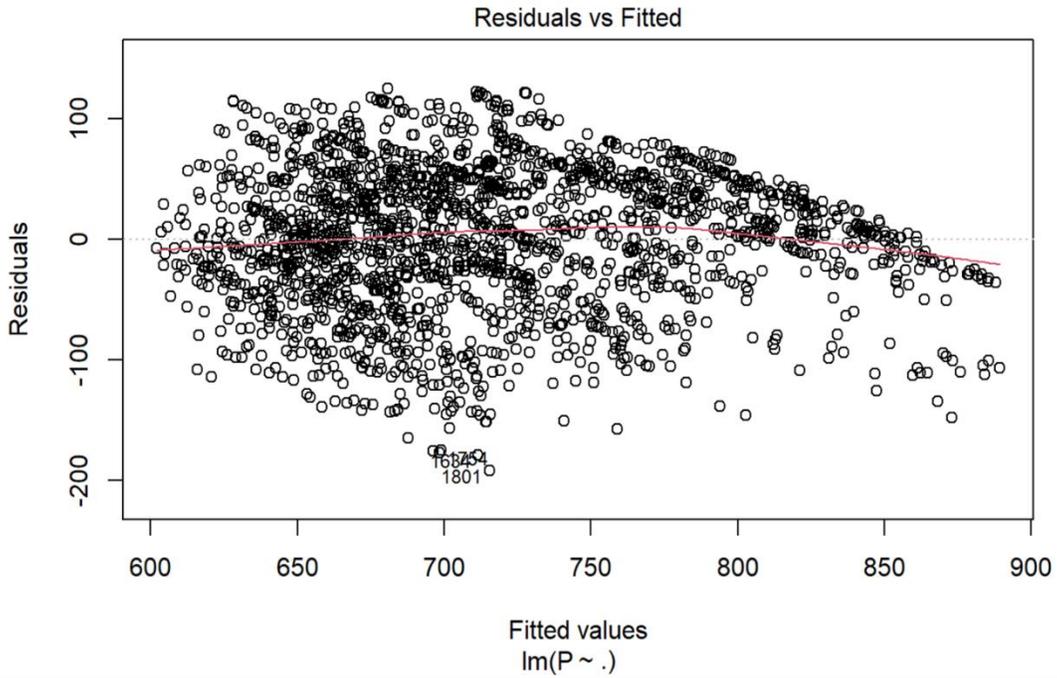


Figure A. 20: Residuals versus fitted values plot for the actual power production

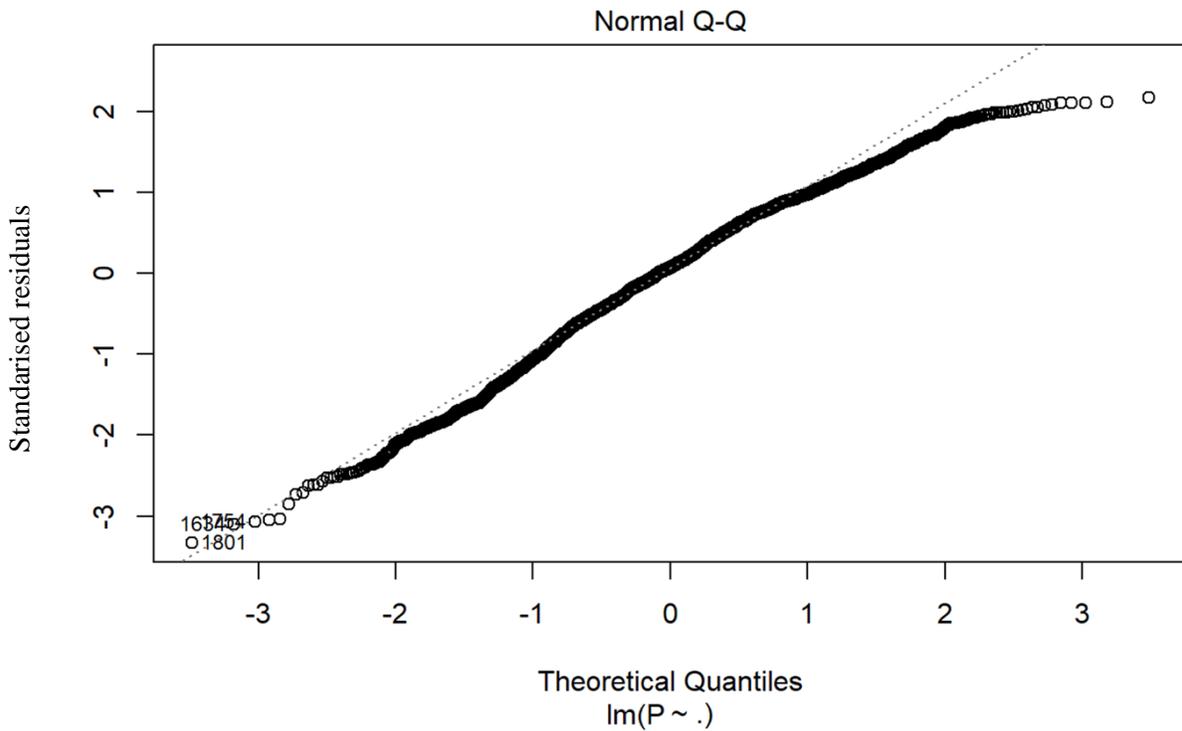


Figure A. 21: Normal probability plot using RStudio for the actual power production

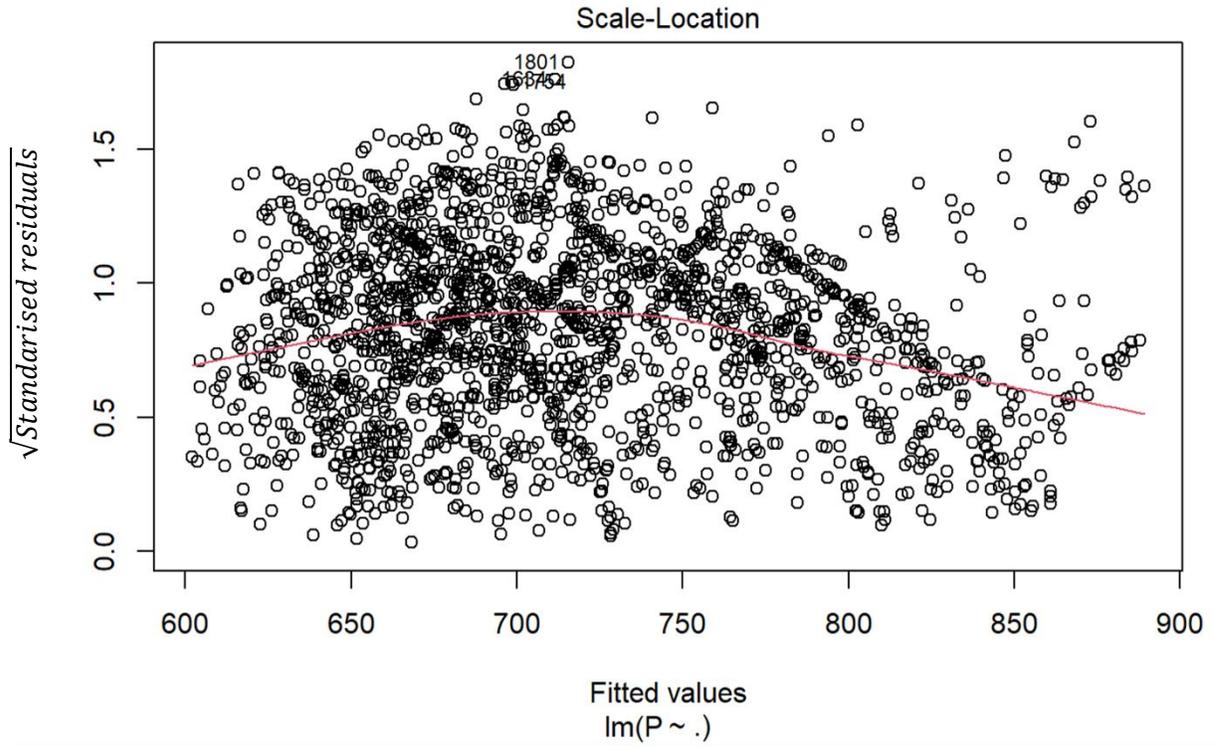


Figure A. 22: Scale location plot for the actual power production