ANALYSIS ON PREDICTION OF RUBBER CROP PRODUCTION USING MACHINE LEARNING

PANG HUI JING

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ANALYSIS ON PREDICTION OF RUBBER CROP PRODUCTION USING MACHINE LEARNING

PANG HUI JING

This report is submitted in partial fulfilment of the requirements for the degree of Bachelor of Electronic Engineering with Honours

Faculty of Electronic and Computer Engineering
Universiti Teknikal Malaysia Melaka

2020/2021

DECLARATION

I declare that this report entitled “Analysis on Prediction of Rubber Crop Production using Machine Learning” is the result of my own work except for quotes as cited in the references.

Signature : ……

Author : Pang Hui Jing ……

Date : 21st June 2021 ……

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Bachelor of Electronic Engineering with Honours.

Signature : .....

Supervisor Name : TS. Dr. Muhammad Noorazlan Shah Bin Zainudin

Date : 21 June 2021

DEDICATION

I would like to dedicate my parents, Pang Foo Chai and Tan Ai Tin, my kind hearted supervisor, TS. Dr. Muhammad Noorazlan Shah Bin Zainudin and thanks to my senior, Kevin Chew and friends.

ABSTRACT

Agriculture sectors are important for the economic growth of countries because it provides job opportunities for majority of population from developing countries. Agriculture provides a significant amount of raw materials to industries like sugar, cotton, palm oil, natural rubber, and so on. Agriculture is the backbone of Malaysia's economy and to date, Malaysia has exported rubber products to various countries. Thus, prediction system of rubber crop production is crucial for making financial decisions earlier when crop production shortage is estimated by system. However, the human based prediction methods are involved time and labour consuming. Crop cut method may conduct measurement errors during weighing process and prone to overestimate the production of crop. In order to solve this problem, a prediction system using machine learning is introduced. This study is implemented and proved the ability to predict the rubber crop production of several states in Melaka, Perak, Pahang and Johor. A various types of machine learning algorithms are applied such as Random Forest, Decision Tree, Linear Regression and Neural Network. The Mean Square Error (MSE) and Mean Absolute Error (MAE) are used as an indicator to evaluate the performance of prediction models. At the end of this study, Linear Regression is able

to provide accurate prediction results with the lowest MAE value in comparison with the abovementioned the machine learning algorithms.

ABSTRAK

Sektor Pertanian sangat penting untuk pertumbuhan ekonomi negara kerana sektor ini menyediakan peluang pekerjaan kepada kebanyakan penduduk dari negara membangun. Sektor pertanian membekalkan jumlah bahan mentah yang banyak untuk industri seperti gula, kapas, minyak sawit, getah asli dan sebagainya. Malaysia telah mengeksportkan produk getah ke beberapa negara. Oleh itu, sistem ramalan pengeluaran tanaman getah sangat penting untuk membuat keputusan kewangan lebih awal semasa apabila kekurangan pengeluaran tanaman dianggarkan oleh sistem. Walau bagaimanapun, kaedah ramalan berasaskan manusia mengambil banyak masa dan perlu memerlukan tenaga kerja yang banyak. Kaedah crop cut akan menyebabkan kesalahan pengukuran dan meramalkan pengeluaran tanaman yang lebih. Untuk menyelesaikan masalah itu, sistem ramalan menggunakan pembelajaran mesin telah diperkenalkan. Sistem ini meramalkan pengeluaran tanaman getah berdasarkan Melaka, Perak, Pahang dan Johor. Random Forest, Decision Tree, Linear Regression dan Neural Network telah digunakan dalam system. Mean Square Error (MSE) dan Mean Absolute Error (MAE) telah dikira untuk menilai prestasi model ramalan. Linear Regression memberikan hasil ramalan yang tepat kerana nilai MAE adalah yang paling rendah antara antara algoritma pembelajaran mesin.

ACKNOWLEDGEMENTS

As the 4th year student in Universiti Teknikal Malaysia Melaka (UTeM), I would like to express my sincere gratitude and appreciation to those who always motivate, support and navigate me along my final year project accomplishment. First of all, I would like to thank my final year project supervisor, TS. Dr. Muhammad Noorazlan Shah Bin Zainudin. Whenever I had doubt, confusing or questions about the research or writing, he consistently steered me to the right path and always enlighten me to achieve better goals of research. Furthermore, I would like to give gratitude to TS. Dr. Nur Fatihah binti Azmi and Dr. Norhidayah binti Mohamad Yatim who as my panels that evaluate my thesis paper. I appreciate their suggestions and comments after the seminar that helps me improve my research. In addition, I also want to thank to my friends who always by my side to help me whenever I faced any problem and troubles. Last but not least, I would like to express my deepest appreciation to my family that always give me strength, confidence and support throughout my university's life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

For examples:

MSE	:	Mean Squared Error
MAE	:	Mean Absolute Error

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

Agriculture plays a major role in economic growth of countries because it provides the job opportunities to vast majority of population from developing countries. When agriculture sector provides the employment opportunities, it will reduce the rate of unemployment and improve the national income level as well as living standards of people. Agriculture is a major source of raw materials for industries such as sugar, cotton, palm oil, natural rubber and so on. Some countries have large exports based on their natural resources. They can export their raw materials and import food requirement to earn foreign exchange. Since Malaysia achieved independence on 1957, agriculture is the backbone of Malaysia's economy. In 2018, agriculture sector contributed 7.3% (RM99.5 billion) to Gross Domestic Product (GDP). Rubber is one of the contributor to the GDP of agriculture sector at 2.8% [1]. The rubber and rubber products from Malaysia are exported to 195 countries and our country is one of the

top 10 rubber exporters and importers. In 2017, our country became the 3rd largest rubber producers in the world [2]. Thus, crop yield prediction is a crucial function in planning for deciding the timely import and export to strengthen the economy of country.

This study aims to develop a system which able to predict the rubber's crop yield. Farmers predict their crop yield by using their previous experience during the past time. It is necessary to incorporate machine learning in prediction of crop yield in order to enhance its efficiency. This work will not only assist farmers to obtain estimated crop yield results but also to analyse the prediction of crop yield using machine learning algorithm.

1.1 Background of Project

Crop yield prediction can be accomplished by using human based method which includes crop cuts, farmer's survey, expert assessment and so on. The crop cuts technique was developed in 1950s in India and it can estimate the crop yield by choosing a random plot of a given size in a selected crop field [3]. This method is required to measure the yield in one or more subplots and the total yield per unit area by dividing the total production of harvested crop by harvested area of the plot. The farmer's survey method is asking farmers to predict or recall the crop yield through interviews. During interviews, farmers will estimate the amount of harvest according to their past predictions experiences by comparing the present crop performance to previous crop performance. Crop yield can be estimated by applying expert assessment method. Expert assessment method is the combination of eye assessment with field measurement and empirical formulas. The field agronomists, extension agents and researchers use the eye assessment which is visually assessing the crop

condition to estimate crop yield [3]. However, human based prediction methods have their own limitations. For example, the weighing process in the crop cut method may conduct measurement errors and tends to overestimate the prediction. Next, the farmers are required to recall their past experience during interviews in order to estimate the crop production. The accuracy of this method is lower if the farmers forget some specific details. For the expert assessment method, accuracy of the prediction is according to the level of expertise by experts.

Machine learning is the non-human based method is currently been used to predict crop yield. One of the models used for crop yield prediction system is using regression analysis. Regression analysis is a common model in machine learning to figure out the relationship between independent variables and dependent variables. Independent variables is referring to the parameters that affect crop yield while dependent variable is crop yield. Prediction by using linear regression algorithm is done by predicting the value of response variable based on the value of explanatory variable [4]. Another prediction system is using random forest algorithm [5]. This algorithm is used for classification for a categorical variable and regression used to predict a continuous response variable. The trained datasets are used to build a randomness number of decision tree. Each tree is split into nodes to enhance the predictability of response variable. Explanatory variables are used to split data while the split point are referring to the predictor variables. Then, the final prediction is based on the mean prediction from all the individual trees.

1.2 Problem Statement

Although the human based method can be used for prediction, this method has several limitations. The crop cut method is considered labour and time consuming as

the enumerators need to weigh the crop when it is ready to be harvested. During the weighing process, there are some problems which are faulty weighing scales and inappropriate way to weigh the crop tend to introduce measurement errors. This may cause overestimation and affect the accuracy of prediction. Farmers' survey method is usually carried out at farmer's house or the at the harvest stored site. This method is likely to yield low accuracy data if the recall data is obtained from several seasons or years. Farmers may forget some specific details because of the long recall periods. When the interviews take longer time, farmer recall data are affected as farmers will get tired and give superficial answers. Enumerator is unable to ask deeper questions and failed to obtain accurate recall data. The expert assessment method is highly depending on the level of expertise from expert because eye estimations require not only practical but also technical knowledge of the yield potential of various crop varieties and their relative performance in different environment. Prediction are more challenging if the experts need to apply their practical and technical expertise while estimating the large range of crops [6]. The prediction of crop yield based on the actual yield grade from the last year can achieve the accuracy up to 51.52% while the prediction by a human expert can achieve the accuracy up to 65.50% [7].

Machine learning can be implemented on prediction system. The input data is received and analysed by machine learning to predict the outcome [8]. There are many algorithms can be used for prediction system. However, choosing a suitable algorithm is crucial in order to obtain high accuracy of prediction results due to its strength and weakness. For example, linear regression is straightforward process to understand the relationship between independent variables and dependent variables. It can prevent overfitting problem by using regularization, cross-validation and dimensionally reduction techniques [9]. Another algorithm used in prediction system is decision tree.

Decision tree is able to handle the variables with non-linear relationship. Furthermore, it requires less time for data preparation, data exploration and data cleaning process [10]. Random forest is also can be used in prediction system. It performs effectively on large databases as it contains many trees on the subset of the data and the final output is obtained by averaging the accuracy from the combination of the trees. Besides, it can maintain the accuracy when there are large proportion of missing data [11]. Artificial Neural Networks are capable of detecting nonlinear relationships in input datasets. Neural networks can be trained to approximate each function expressing the dependence of the outputs on the inputs with the correct topology and weights of connections between neurons [12]. Since all the algorithms have their own strength, it is important to determine which algorithm is suitable for prediction system.

1.3 Objectives

The objectives of this project are:

- To explore and develop prediction system for estimating crop yield of rubber using machine learning algorithms
- To evaluate and analyse the prediction result of crop yield using various parameters such as temperature, rainfall, humidity and planted area.
- To compare the accuracy of crop yield prediction using random forest, linear regression, decision tree and neural network.

1.4 Scope of Project

This project will begin by collecting the crop yield rubber data. In this work, the data is collected from official website of Department of Statistic Malaysia. The training subset is used to fit the parameters while the testing subset later are used to predict the crop yield using Python programming language. The results of crop yield

predictions are obtained using different type of parameters such as temperature, rainfall, humidity and planted area in 10 years period. These parameters are chosen since the temperature and rainfall are considered as the primary factors that can give impact to rubber tree growth and its productivity [13]. Then, the predicted results are compared with the actual results in order to determine its accuracy. Furthermore, this project will also be compared the performance of crop yield prediction using 4 types of machine learning models which are Random Forest, Linear Regression, Decision Tree and Neural Network.

1.5 Report Outline

Chapter 1 is the introduction chapter for this project. The terms, background and brief explanation of this project are covered in this chapter. The problem statement based on case studies along with the objectives to achieve would also be highlighted in this chapter along with the scope of work for this project.

In Chapter 2, background study includes a detailed review of researched area. Previous study and relevant history on the problem stated are being studied. Current information and solution regarding the issue were also indicated in this part.

The methodology chapter describes the research methods that are implemented in this project. Flow chart and graph are included to portrait a more detailed steps that have been taken in doing this research. Each method used are categorized in different section for the ease of describing different steps involved in succeeding different objectives.

For Chapter 4, the title of this chapter is result and discussion. The results which are in form of statistical analysis or simulation are presented. The outcome of the system would also be included and each data are discussed.

The final chapter represents the summary of the whole project. The conclusion section will describe the summary of methods implemented and the analysis method together with results obtained. For the future works sections, possible future improvement and further development of the current project are discussed.

CHAPTER 2

BACKGROUND STUDY

2.1 Related Projects

There are several similar research papers are found through research and survey on the previous articles or journals. Those article are related to this work and summary of comparison from all the papers are summarize in this section.

Firstly, “Smart Farming System: Crop Yield Prediction Using Regression Techniques” [14], a crop yield prediction system is introduced. This system made use of regression model which are Support Vector Machine Regression, Random Forest Algorithm and Multivariate Polynomial Regression. The crop yield is predicted according to previous data of crop yield versus temperature, precipitation and rainfall data. Then, the results of prediction, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error and R-squared values in each

regression model are compared. The results showed the Support Vector Machine Regression has the highest accuracy of prediction among the regression model.

Next, article "Crop Yield and Rainfall Prediction in Tumakuru District using Machine Learning" [15], the author developed a system which can predict the values of rainfall to assist and suggest farmers in selecting suitable crops. The rainfall data and price of crops are used as parameter in this crop selection system. Different predictive algorithms are used in the system which are Linear Regression, Support Vector Machine, K-nearest neighbors and Decision Tree. The output showed Support Vector Machine algorithm provide highest efficiency among other algorithms. Figure 2.1 shows the system architecture of the system.

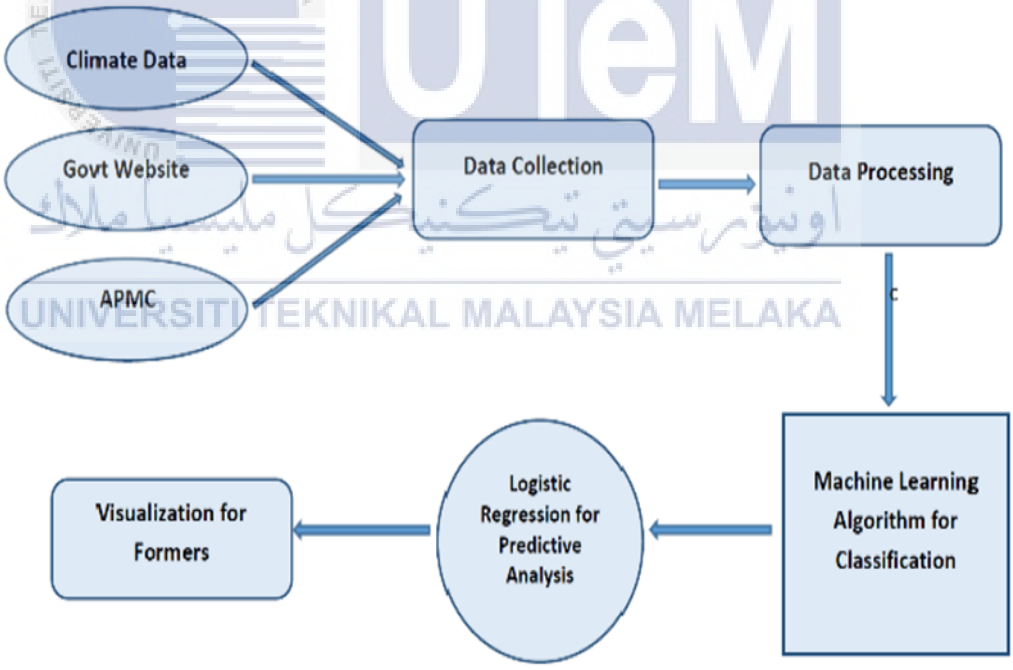

Figure 2.1: System Architecture [15]

In the next article, "Predicting Early Crop Production by Analyzing Prior Environment Factors" [16], author has constructed a system to forecast the production of crops in specific season. Six environmental variables are temperature, rainfall,

humidity, sunshine, cloud coverage and wind speed. Then, the crop yield is predicted according to the listed parameters and Linear Regression and Neural Network are applied in this prediction model. The selected crops in this paper are Aus Rice, Amon Rice, Boro Rice, Jute, Potato, and Wheat. This system is built by using RapidMiner Studio. Firstly, the data is loaded from database and data pre-processing steps is applied to split the data tables into single table. Then, the data is separated based on different regions and passed through machine learning algorithms to train the model. Finally the prediction results and performance testing are done. The Root Mean Squared Error (RMSE) method is used to calculate the error of algorithm.

Next, article "Rice Crop Yield Prediction in India using Support Vector Machines" [17] has introduced a system which can predict the production of rice. The parameters used in the research are precipitation, temperature, planted area and reference crop evapotranspiration. The raw dataset is converted to Microsoft Excel. In order to apply the data mining techniques on the dataset, the unnecessary column are eliminated. Then, the Support Vector Machine is applied to create a model from training data. The prediction results are obtained in WEKA tool. Lastly, the results are evaluated by using different performance testing. The accuracy of the test is done by computing the F1. Furthermore, Mathews Correlation Coefficient was applied to measure the classification quality. Then, the performance of the model is evaluated by computing the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Root Relative Squared Error (RRSE) and Relative Absolute Error (RAE). The F1 score is 0.69 while the Mathews Correlation Coefficient is 0.54. Next, the RMSE is 0.39, MAE is 0.23, RRSE is 82.51% and the RAE is 67.38%.

Aside from these, in the journal “Accurate prediction of sugarcane yield using a random forest algorithm” [18], the author designed a system to predict the sugarcane yield by using Random Forest algorithm. The parameters used in this system are including indices for simulated biomass, some climatic factors which are rainfall, radiation and temperature. Furthermore, the Southern Oscillation Index and Niño 3.4 region SST anomalies are the part of parameters too.

The next article “Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh” [19], the author has developed a system to predict the yields of selected crops. There are three environmental variables which are biotic input attributes and area central input attributes. The environmental variables are including rainfall, humidity, temperature and average sunshine. The biotic input attributes are pH value of specific district’s soil and soil salinity. The area central input attributes are irrigated area and cultivated area of each crop. The selected districts are clustered according to values of related attributes and this method is done by using K-means Clustering algorithm. Then, the prediction results of selected crop yields are obtained by using several classification or regression model which are Linear Regression, K-nearest neighbour algorithm and Neural Network. The range of accuracy results obtained from each algorithm are within 90% to 95%. Then, the Root Mean Squared Error (RMSE) of each algorithm are computed in this paper to evaluate the performance of model.

Next, article “Crop Yield Prediction Using Machine Learning” [5] has introduced a system to predict crop yield based on climatic parameters which are temperature, vapour pressure, precipitation, cloud cover and wet day frequency. The data are collected from different source to prepare datasets. 75% of datasets are extracted for

the training data part while 25% of datasets for testing data part. Then, the prediction model is constructed by implementing Random Forest algorithm. The prediction results of this paper showed the accuracy has achieved more than 75% in all type of crops. Figure 2.2 shows the flowchart of the system.

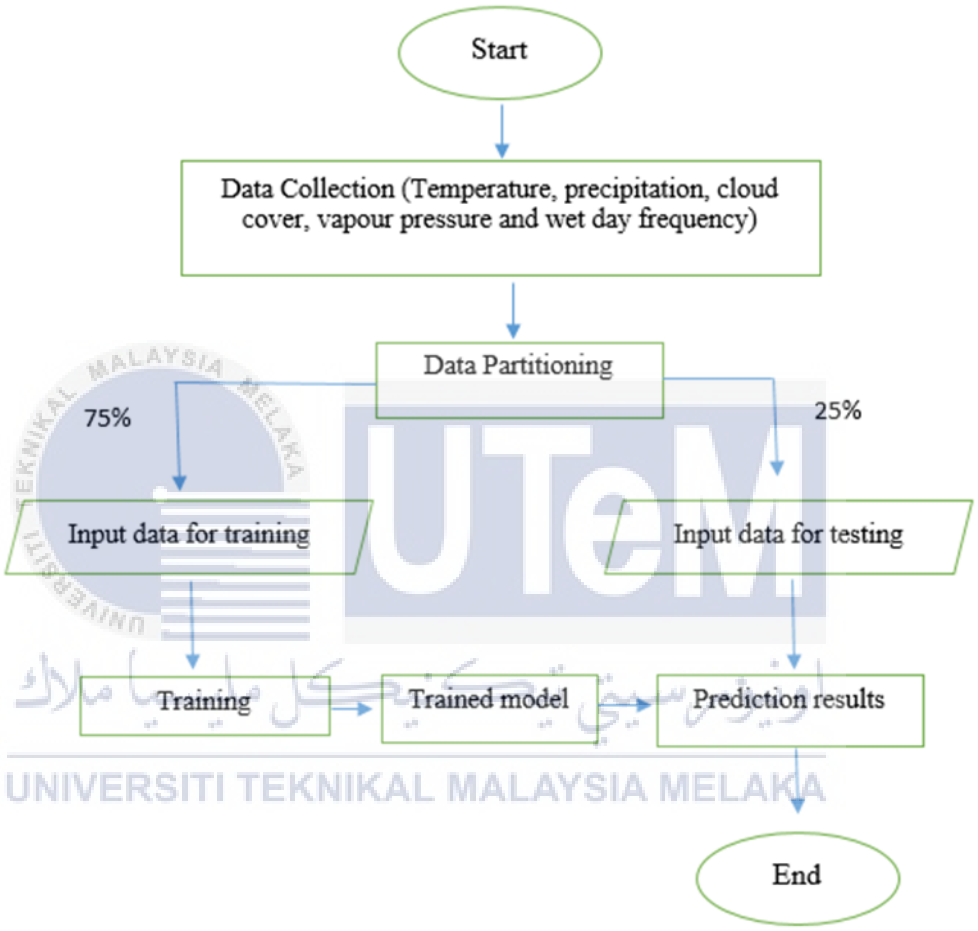

Figure 2.2: Flowchart of the system [5]

The next article "Heuristic Prediction of Crop Yield using Machine Learning Technique" [20] has introduced a system which is able to predict crop yield based on rainfall, temperature, soil moisture and humidity. The dataset consists of four columns with input data and one column with crop yield. Then, the dataset is separated to

training and testing purpose. The K-nearest neighbour algorithm is implemented in this system. Figure 2.3 shows the flowchart of this system.

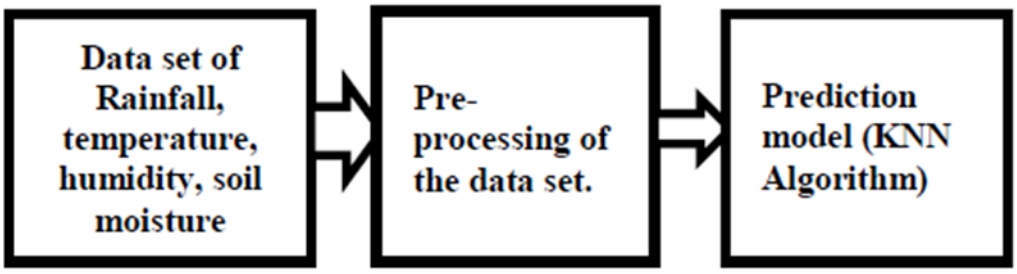

Figure 2.3: Flowchart of the system [20]

In the journal "An Efficient Analysis of Crop Yield Prediction Using Hadoop Framework Based on Random Forest Approach" [21], the author has developed a system which can predict the crop yield based on climate factors and soil parameters. The climate factors are precipitation, temperature, topography and solar energy while the soil parameters are type of soil, pH level, iron, manganese, organic carbon, copper, Sulphur, nitrogen, phosphate and potassium. Random forest algorithm is implemented to classify the dataset. The Hadoop Framework is used to load data into Hadoop Distributed File System (HDFS). The HDFS is a data storage system for Hadoop applications and provides easier access to application data. Then, the MapReduce programming is used to process the dataset. Then, the results are predicted and the next step is continued with evaluating the performance of model. The recall value, precision value, accuracy and F1 score are computed and the results are 0.9453, 0.9380, 0.9143 and 0.9416 respectively. Figure 2.4 shows the flowchart of this system.

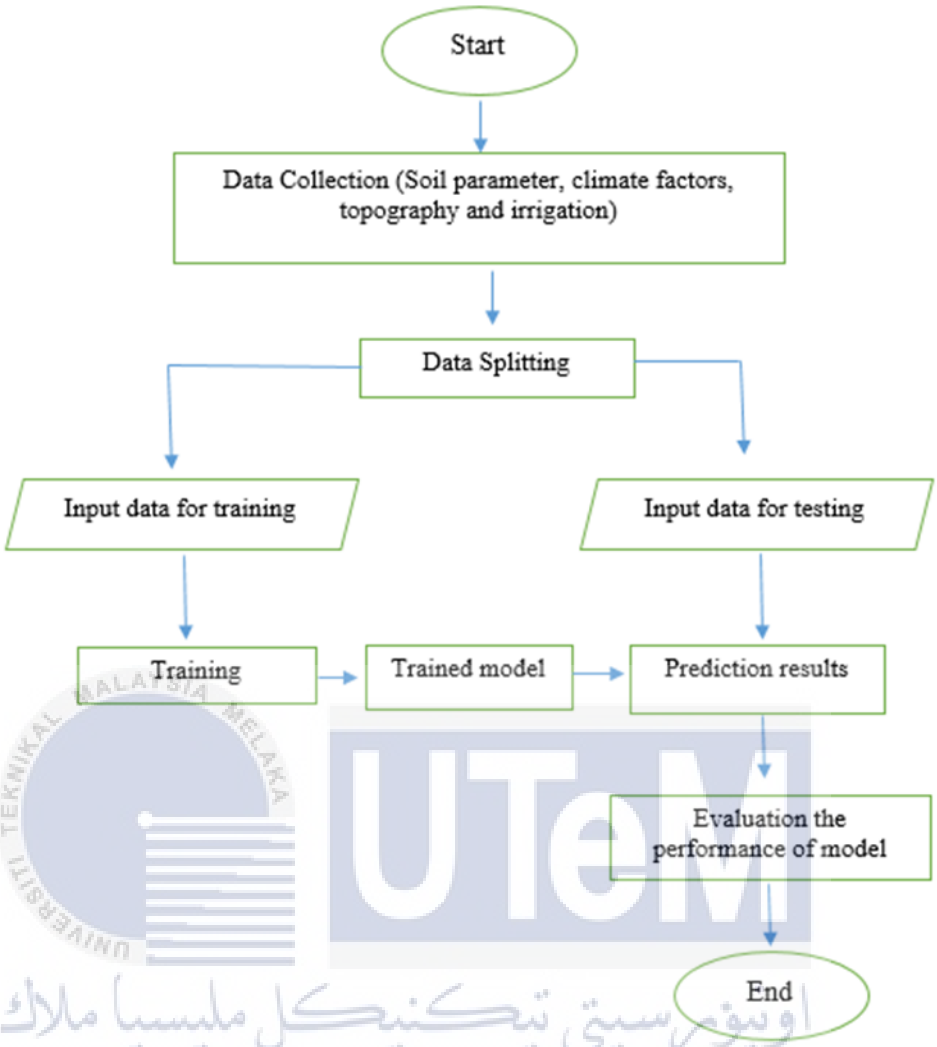

Figure 2.4: Flowchart of the system [21]

The next article "Agricultural Production Output Prediction Using Supervised Machine Learning Techniques" [22], a system is designed to predict various crop yield according to temperature and rainfall in 10 major cities of Bangladesh. Two algorithms are implemented in the system which are K-nearest neighbour regression and Decision Tree algorithm. The analysis of performance by each model is done in this paper and the percentage error of crop yield rate prediction obtained are below 10%. Figure 2.5 shows the implementation of decision tree in this system.

Furthermore, the overall research on related projects regarding to the crop yield prediction is tabulated as shown in Table 2.1.

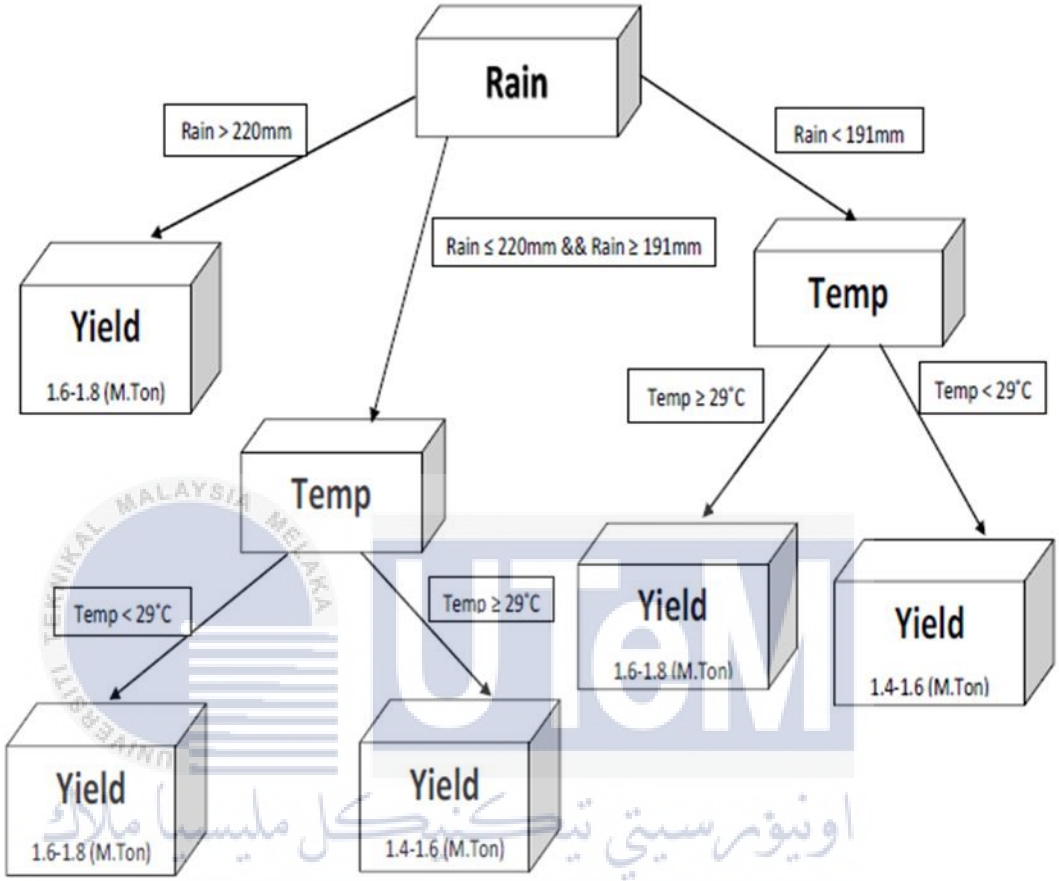

Figure 2.5: Implementation of Decision Tree [22]

Table 2.1: Research on Crop Yield Prediction

No	Author	Type of parameter	Type of model	Type of evaluation metrics
1	A. Shah, A. Dubey, V. Hemnani, D. Gala, and D. R. Kalbande (2018)	Temperature, precipitation and rainfall	Support Vector Machine Regression, Random Forest Algorithm and Multivariate Polynomial Regression	Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error, R-

				squared values
2	L. Girish, S. Gangadhar, T. R. Bharath, K. S. Balaiji, K. T. Abhishek (2018)	rainfall data and price of crops	Linear Regression, Support Vector Machine, K-nearest neighbors and Decision Tree	Classification accuracy
3	T. Osman, S. Shahjahan Psyche, M. Rafik Kamal, F. Tamanna, F. Haque, and R. M.Rahman (2017)	temperature, rainfall, humidity, sunshine, cloud coverage and wind speed	Linear Regression and Neural Network	Root Mean Squared Error (RMSE)
4	N. Gandhi, O. Petkar, L. J.Armstrong, and A. Kumar Tripathy (2016)	precipitation, temperature, planted area and reference crop evapotranspiration	Support Vector Machine	F1 score, Mathews Correlation Coefficient, the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Root Relative Squared Error (RRSE) and Relative Absolute Error (RAE)
5	Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber (2016)	indices for simulated biomass, rainfall, radiation and temperature	Random Forest algorithm	Classification accuracy
6	A. T. M. Shakil Ahamed et al (2015)	rainfall, humidity, temperature, average sunshine, pH value of soil, soil salinity	K-means Clustering algorithm, Linear Regression, K-nearest neighbour algorithm and Neural Network	Root Mean Squared Error (RMSE)
7	M. Champaneri, D. Chachpara, C.	temperature, vapour pressure, precipitation,	Random Forest algorithm	Classification accuracy

	Chandvidkar, and M. Rathod (2020)	cloud cover and wet day frequency		
8	S. Pavani, S. B. P. Augusta (2019)	rainfall, temperature, soil moisture and humidity	K-nearest neighbour algorithm	Classification accuracy
9	S. Sahu, M. Chawla, and N. Khare (2017)	precipitation, temperature, topography, soil parameters and solar energy	Random forest algorithm	Recall value, precision value, accuracy and F1 score
10	M. Tahmid Shakoor, K. Rahman, S. Nasrin Ratya, and A. Chakrabarty (2017)	temperature and rainfall	K-nearest neighbour regression and Decision Tree algorithm	Percentage error

2.2 Theory

2.2.1 Prediction System

The prediction system using machine learning algorithms can be applied in various sector. In the journal titled "Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia" [23], the author has developed a system to forecast the house price. The type of data implemented in this system are numerical and categorical. However there are some missing values in the original dataset. If there are missing values up to 55% from that column, it eliminated from dataset. Then, imputation is applied to other input variables to solve the problem with mild amount of missing values. The machine learning models used in this system are Polynomial Regression, Support Vector Machine, Linear Regression, Regression Tree and Neural Network. Then, the performance of models are evaluated by computing the Mean Squared Error (MSE). The next article titled "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction" [24] has introduced a system to predict

the rainfall. There are some missing values in the dataset and the mean average mechanism can deal with this problem. This system is using Random Forest, Support Vector Machine, Decision Tree, Neural Network and Naïve Bayes as prediction model. Lastly, the models used in this system are evaluated by computing the information retrieval metrics which are recall, precision and f-measure [24].

2.2.2 Decision Tree Algorithm

Decision trees are one of the methods for data mining to extract useful information from large datasets [25]. This algorithm can be used in variable selection as choosing the most related input variables. The decision tree model is formed from the relevant input variables. After the relevant variables are identified, some variables are prioritized according to the reduction of model accuracy. The importance of variable is depending on the impact of variable. The decision tree analysis is able to solve the missing values. The missing values are classified into separate category that can be analyzed with another categories. When there are lots of missing value, decision tree is built and set the missing value as target variable. Then, prediction is computed and missing values are replaced with predicted value. Decision tree model can be derived from historical data and make a prediction. Furthermore, decision tree models deal with the extra categories of one categorical variable by deciding to collapse categorical variables into manageable number of categories.

The main parts of decision tree model are nodes and branches. The type of nodes are root node, internal nodes and leaf nodes. Root node is the top decision node and it represents the whole population and population is separated depends on various features. The possible choices available at that point in tree structure are internal nodes. Internal nodes, also known as chance nodes. The final output of decision trees

is called leaf nodes. The sub section from internal nodes and root nodes is called branches. There are several steps to build a model which are splitting, stopping and pruning. The most important input variables are identified during the progress of building model. The splitting process is required to separate records at the root node into two or more sub nodes. Generally, only part of the input variables are used to build the decision tree model and some certain input variable can be used several times at different levels of decision tree. Over fitted problem may exist when building decision tree model and affect the prediction results. Thus, stopping step is applied to avoid the decision tree model to become more complicated. This step can be done by limiting the number of records in a leaf, number of records in an important node for splitting and the depth of any leaf from root node [25]. Sometimes the stopping step may not work as expectation thus pruning step is taken. The best possible sub tree is chosen from multiple candidates. The other method is use the validation dataset. These methods help to optimize the size of tree. Pruning step is categorized into pre-pruning and post-pruning. Chi-square tests are applied in pre-pruning to prohibit non-significant branches [25]. After a complete decision tree is generated, post-pruning is used to eliminate branches in order to enhance the accuracy of the output. Figure 2.6 shows the example of decision tree.

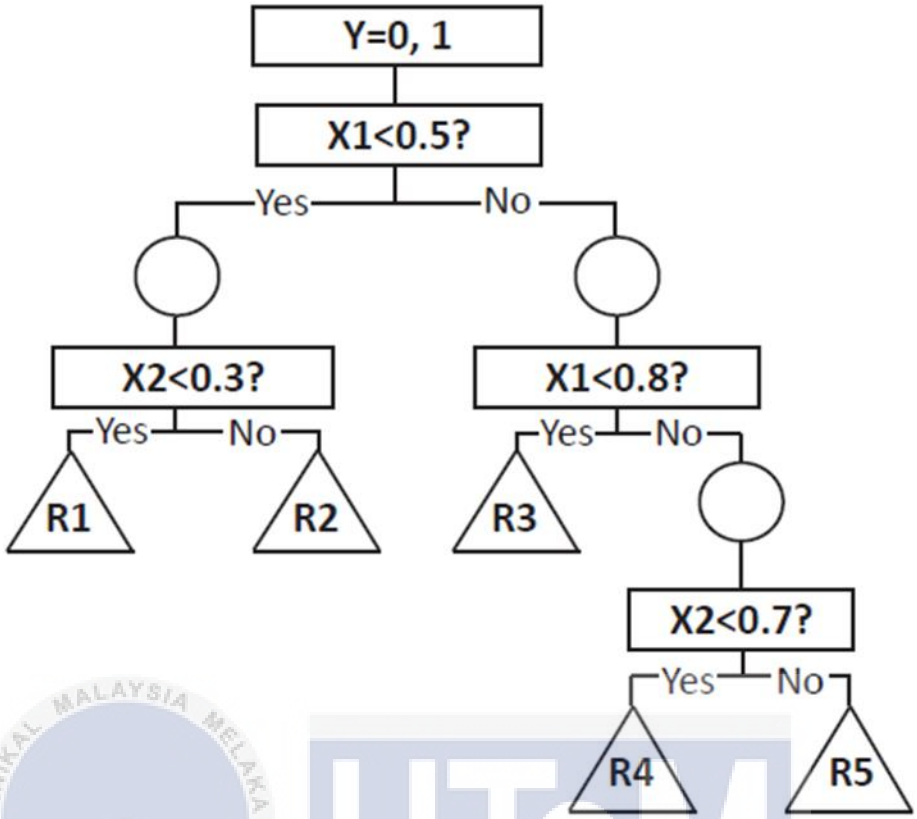

Figure 2.6: Example of Decision Tree [25]

2.2.3 Random Forest Algorithm

Random forests are constructed from multiple tree predictors and each tree contain predictor variables on its lead nodes and other dependent variables on internal nodes [26]. The decision tree from the random forests generate the decision and it is considered as vote for that predictor [27]. Then, the random forests will consider the decision receiving the major votes from all of the trees.

Random forests are considered as an ensemble learning method for classification and regression [27].Classification is a technique applied on class variables to predict the decision values for categorical or qualitative types of variables using calculation which involves one or more predictor and variables. This technique is commonly applied in variety of fields including computer science, botany, medicine, and

psychology. One of its strength is that it can be visualized into a graphical manner and it aids user to interpret data in a simpler method compare to numerical interpretation .There are lots of classification techniques or classifier which can be employed to predict a qualitative response. There are a few classifier which are commonly used. These classifiers includes k-nearest neighbour, Bayes' theorem, logistic regression and more. The attention on classification trees are rising due to few of its features. These features includes the hierarchical property of the technique and its flexibility. Regression technique is applied during the random forest generate multiple trees for quantitative or numerical values of the class variable [27]. In contrast, classification regression is applied when generating trees for qualitative class variables. Thus, the final result value from Random Forest regression is numerical. The mean-squared error can be computed by this formula, $E_{X,y}(y - X)^2$ [27]. X is the predictor variable while y is the class variable.

2.2.4 Linear Regression

Linear regression in one of the predictive model to figure out the relationship between independent and dependent variables. There are two types of linear regression which are simple linear and multiple linear regression. The linear regression's equation is $y = x\beta + \varepsilon$ [28]. The independent variable is represented by y and the values of y is in continuous or categorical. The dependent variable is represented by x and it is continuous value. Simple linear regression separate the independent variables and dependent variables to determine the relationship between two different variables as similar to correlation [28]. Figure 2.7 shows the prediction process using univariate regression analysis because only one independent variable is involved. The equation of univariate regression is $y = a + bx + \varepsilon$ [28].

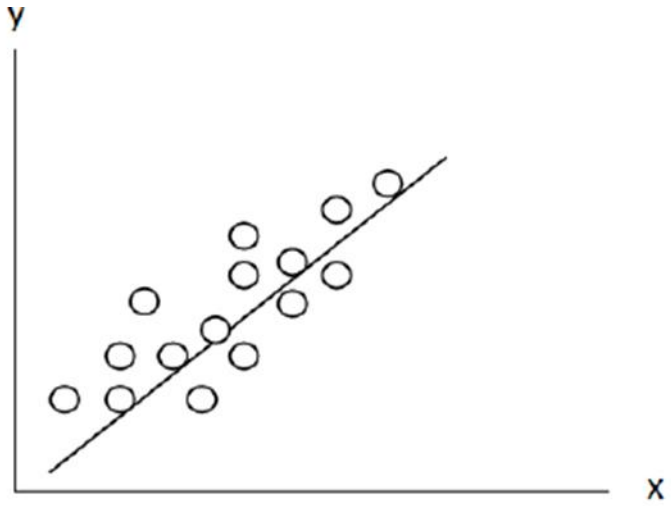

Figure 2.7: Simple Linear Regression [28]

For the multiple linear regression, the prediction process is using multivariate analysis because more than one independent variables are involved. The equation of multivariate analysis is $y = a + b_1x + b_2x + \cdots + b_nx + \varepsilon$ [28]. Figure 2.8 shows the multiple linear regression.

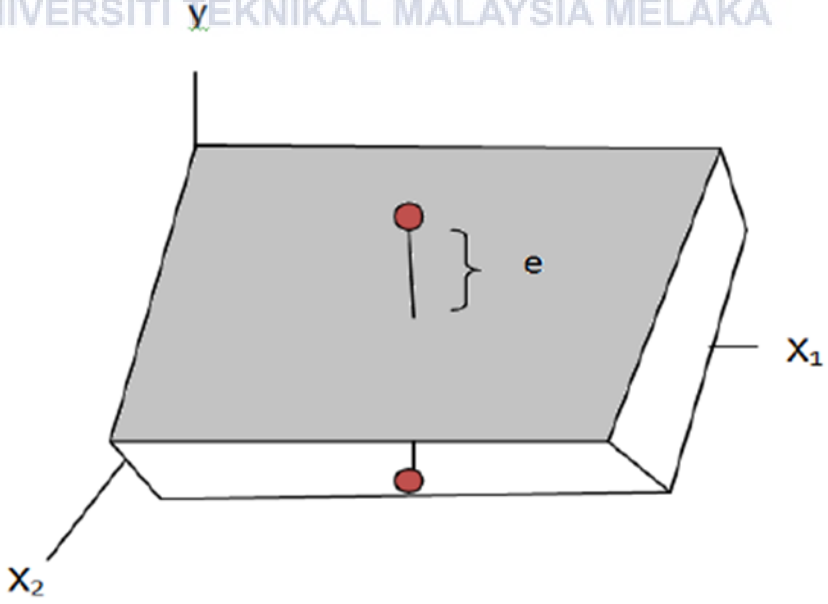

Figure 2.8: Multiple Linear Regression [28]

2.2.4.1 Regularisation

Regularisation is crucial in machine learning algorithm. This is a type of regression in which the coefficient estimates are constrained, regularised, or shrunk towards zero. To avoid the problem of overfitting, regularisation is introduced to measure the complexity of the model and it helps to prevent the learning of complex model. The first example of regularisation is ridge regression. When the number of predictor variables are more than the number of observations, it tends to overfit a model and ridge regression is used to create a suitable model by minimising the estimated coefficients toward zero to solve this problem [29]. The ridge coefficients are defined as $\hat{\beta} = \arg\min \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$ [30]. From this equation, $\frac{1}{n}\|Y - X\beta\|_2^2$ is the loss function while β is the coefficient estimates for different types of variables [30]. Next, $\lambda\|\beta\|_2^2$ is the penalty term and the λ is tuning parameter to adjust the strength of penalty term [30]. Thus, the value of λ is crucial because when $\lambda = 0$, the penalty term will become zero and ridge regression is equal to loss function. When $\lambda = \infty$, the impact of minimisation penalty will increase and coefficient estimates in ridge regression will approach to zero. The ridge regression utilised the L_2 regularisation method. The next example of regularisation is LASSO regression. LASSO estimator is defined as $\hat{\beta} = \arg\min\left\{\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\}$ [31]. The difference between LASSO and ridge regression is the $\|\beta\|$ as its penalty term as known as L_1 norm. The λ is tuning parameter to adjust the strength of penalty term. The LASSO regression will same as loss function when $\lambda = 0$. Again, the LASSO coefficient will approach to zero when $\lambda = \infty$. However, the L_1 penalty is structured in such a way that some coefficients are reduced to zero. Thus, it has the ability to select variables in a linear model. When λ increase, more coefficients are set to zero tends to choose fewer variables. Besides, more shrinkage is used among the non-zero coefficients [31].

2.2.5 Artificial Neural Network

An artificial neural network can be divided into three parts which are input layers, hidden layers and output layer. The input neurons are also known as input layer, it is used to feed input patterns into the rest of the network. This layer is in charge of receiving data, signals, features, or measurements from the external environment [32]. These are generally normalised within the activation function's limit values. The normalisation will improve the numerical precision of the network's mathematical operations. The following layer are intermediate layers of units, which are hidden layers. These layers are made up of neurons that are in charge of extracting patterns related to the process or system. These layers handle the majority of a network's internal processing. The hidden layers are then followed by a final output layer. This layer is made up of neurons as well. It is in charge of generating and displaying the final network outputs, which are the result of the processing done by the neurons in the previous layers.

Multilayer feedforward networks is one of the main architecture from artificial neural networks. One or more hidden neural layers create feedforward networks with multiple layers. The main network using multiple-layer feedforward architectures in this system is Multilayer Perceptron (MLP) [32]. Multilayer Perceptron is the most commonly used model in neural network applications utilising the back-propagation training algorithm. It is the origin of Perceptron model which is proposed by Rosenblatt in the late 1950s [33]. In multilayer perceptron, the number of hidden layers and the number of neurons in each layer can be adjusted according to specific problem. The direction of connections are constant which is from lower layers to upper layers. Meanwhile, the neurons are not interconnected in same layer [33]. Each unit is linked to all units in the following layer while each unit receives input from all units

in the previous layer. Each connection has weight, which represents the influence of the unit on the response of the unit in the following layer. The outcome of a multilayer perceptron is determined by the input as well as the strength of the unit connections. As information is provided to a multilayer perceptron by activating neurons in the input layer, the information is then processed layer by layer before the output layer is activated. Multilayer perceptron is able to approximate virtually any function to any desired accuracy with provided sufficient hidden units and data. Figure 2.9 shows the a feedforward network with multiple layers consisting of an input layer with n sample signals, two hidden neural layers with n_1 and n_2 neurons, and one output neural layer with m neurons representing the problem's respective output values [32].

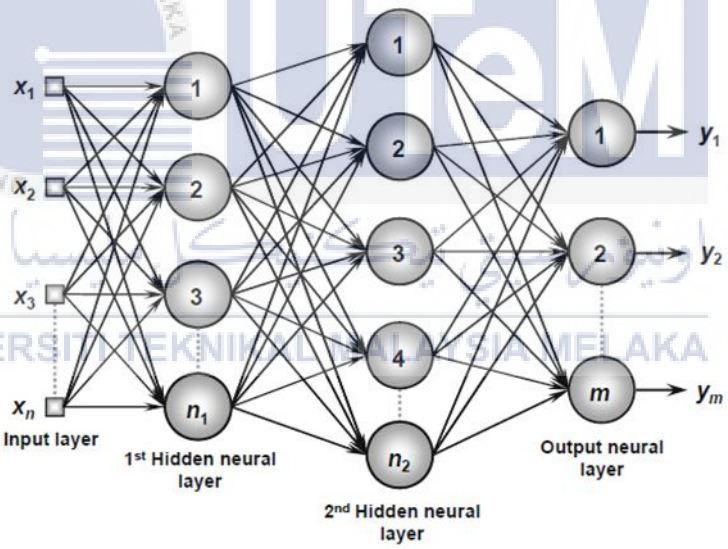

Figure 2.9: Multiple Linear Regression [32]

2.2.6 Data Standardisation

Data standardisation is also known as z-score normalization. In this process, the data is converted into distribution with a mean of 0 and a standard deviation of 1 [34]. The standardized value can be computed by this formula, $v' = \frac{v-\bar{A}}{s_A}$ [34]. V represents

the original data value and $\bar{A}$ is the mean of the data values. $\bar{A}$ can be computed by this formula, $\bar{A} = \frac{1}{n}\sum_{i=1}^{n} v_i$[34]. Apart from that, s_A is mean absolute deviation and it is computed by this formula, $s_A = \frac{1}{n}\sum_{i=1}^{n} |v_i - \bar{A}|$[34]. Generally, the range of z-score is from -3.00 to 3.00 in a standard normal distribution. This method is applied when the range of input data's feature is large. The feature of input data with different measurement unit is the main factor to cause the large difference of range.

2.2.7 Evaluation Metrics

The performance of the regression model is evaluated by computing Mean Absolute Error (MAE) and Mean Squared Error (MSE). Both of the evaluation metrics are able to determine the goodness of regression model by identifying the average prediction error. The calculation of MAE is done by applying the formula, $MAE = \frac{1}{n}\sum_{j=1}^{n}(|A_j - P_j|)$ [35]. The A_j represent the actual values, P_j represent the predicted values and n is the size of dataset. The average of summation of absolute error between the actual values and predicted values is calculated from this formula. Lower value of MAE represents a better performance of machine leaning model. Next, the calculation of MSE is done by applying the formula, $MSE = \frac{1}{n}\sum_{j=1}^{n}(A_j - P_j)^2$ [35]. This formula is quite similar with MAE but the difference is MSE is used to calculate the average of sum of square of error. MSE value is always positive and the error is the mean deviation from the true values [36]. If lower value of MSE that closer to 0 is calculated from the final output in prediction model, it means the performance of model is good and provide more accurate prediction result.

CHAPTER 3

METHODOLOGY

In this section, the evaluation of prediction system using machine learning is presented. The overview flow charts of the proposed system are including software implementation, database collection and preparation, data splitting, build prediction model, data training and data testing on trained model. Apart from that, the environment and sustainability of this proposed system is discussed in detail.

3.1 Overall Project Methodology

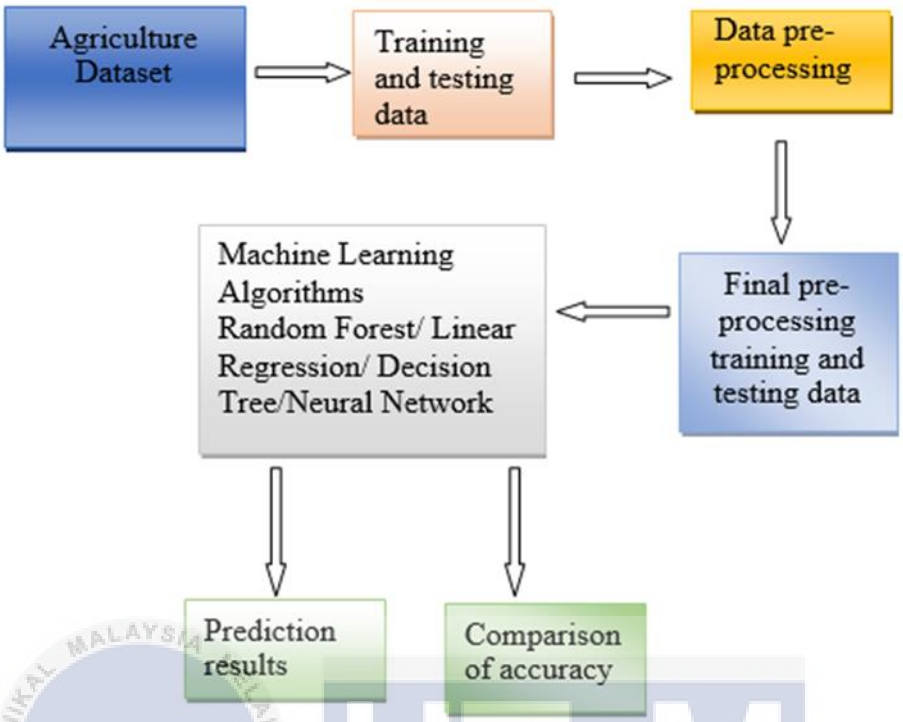

Figure 3.1: Block diagram of system

Figure 3.1 illustrate the block diagram of this work. The agriculture database is collected from official website of Department of Statistic Malaysia and website of Open Government Data Malaysia. The dataset is split into 2 parts which are training subset and testing subset. Then, the dataset will undergo data-pre-processing step if there is missing data. Training data is a sample of data to fit the model. Random Forest algorithm, Decision Tree algorithm, Linear Regression algorithm and Neural Network are applied in this work. Testing subset is used to evaluate the performance of the model fit on the training subset. After that, the prediction results are computed from the system and the accuracy of each model are compared.

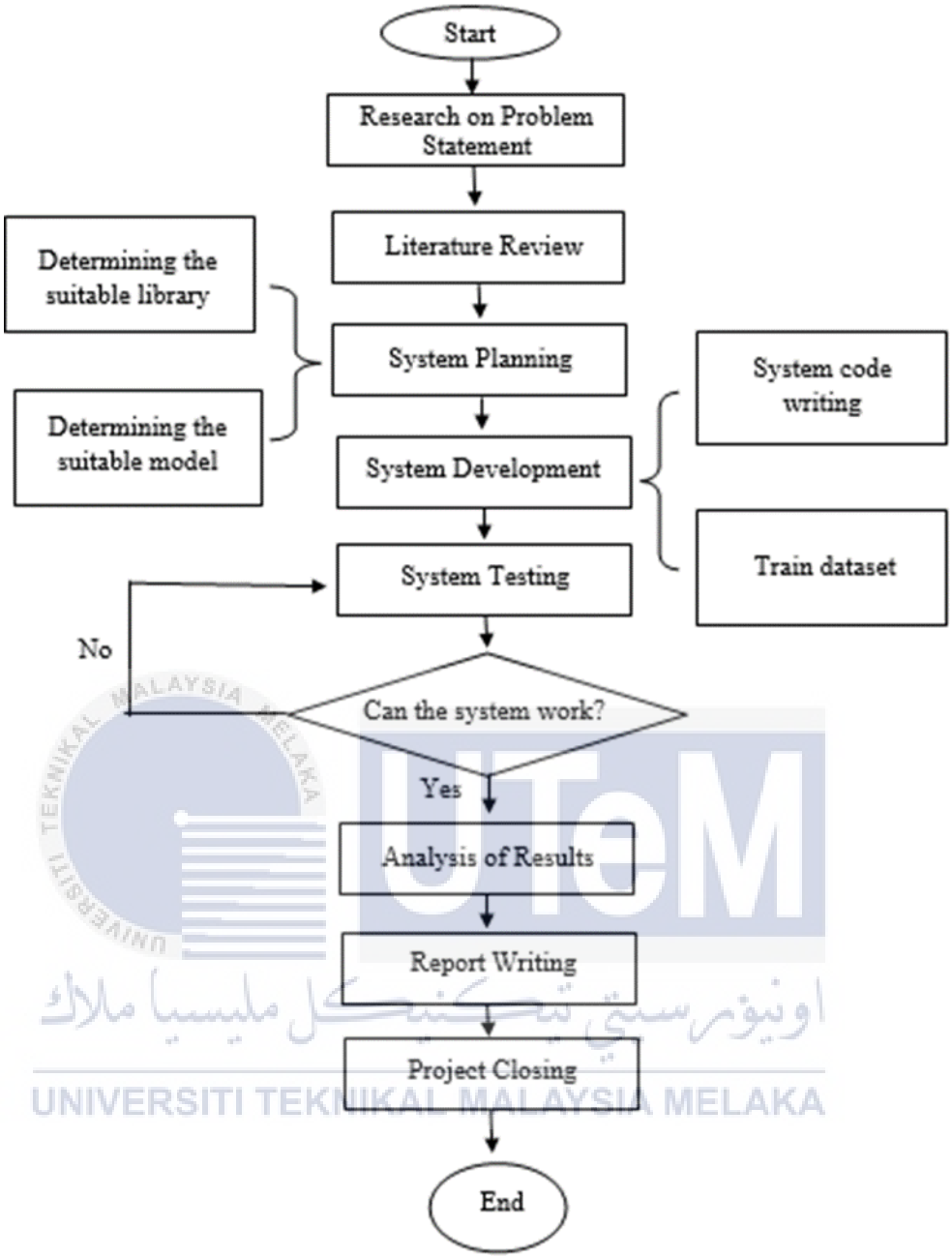

Figure 3.2: Flowchart of project

Figure 3.2 shows the procedure flow of the whole work. Various research are conducted before starting of the work in order to get a better understanding about the objectives and to identify the problem. After reviewing the research, the next step is continue with literature review of similar work. The purpose of literature review is for understanding and comparing the similar research to identify which and what method

is the best choice for the work. After all research are done, the next step is system planning. The suitable library used in the system are determined. For this system, Scikit-learn is chosen because it is a machine learning library for Python language. It consists of various algorithms and also supports scientific libraries and Python numerical which are SciPy and NumPy. Furthermore, Pandas library is used to import dataset while Matplotlib is used for graphical plotting. Then, the suitable models for prediction are chosen; Random Forest, Decision Tree, Neural Network and Linear Regression. Next, system development is follow-up when things are done here. For the development part, the first part to do is writing the program for the system. The system is wrote using Python Programming in the PyCharm IDE. Then, the second part is training and testing the dataset. After finish writing the program, the system is tested and run through troubleshooting process. When everything is done, the result is analyzed for its accuracy. Lastly, the report is written according to this work.

3.2 Software Implementation and Data Preparation

In this work, the software used is PyCharm IDE. PyCharm is an integrated development environment (IDE) and it is used for computer programming which is Python language. Pip is used in PyCharm to manage project packages. Some scientific tools are available in this software which are Matplotlib, NumPy Pandas, and Scikit-learn. Set of Python plotting libraries are provided by Matplotlib while basic package for scientific computing with Python is provided by NumPy. Moreover, Pandas is a tool for loading data from a variety of file formats into in-memory data objects. Scikit-learn provides multiple types of machine learning algorithm. Those libraries available in PyCharm IDE help to accomplish this proposed system. All the additional library can be downloaded from PyCharm IDE Python Interpreter menu. Figure 3.3 shows the PyCharm IDE icon.

Figure 3.3: PyCharm IDE 2020

The aim of this work is to develop a crop yield prediction system. Thus, all the dataset are collected from official website of Department of Statistic Malaysia and website of Open Government Data Malaysia. The planted area of rubber tree data and production of rubber data are extracted from Annual Rubber Statistics Yearbook 2011. Then, these datasets are sorted with temperature data, rainfall data and humidity data in CSV file.

3.3 Program Implementation

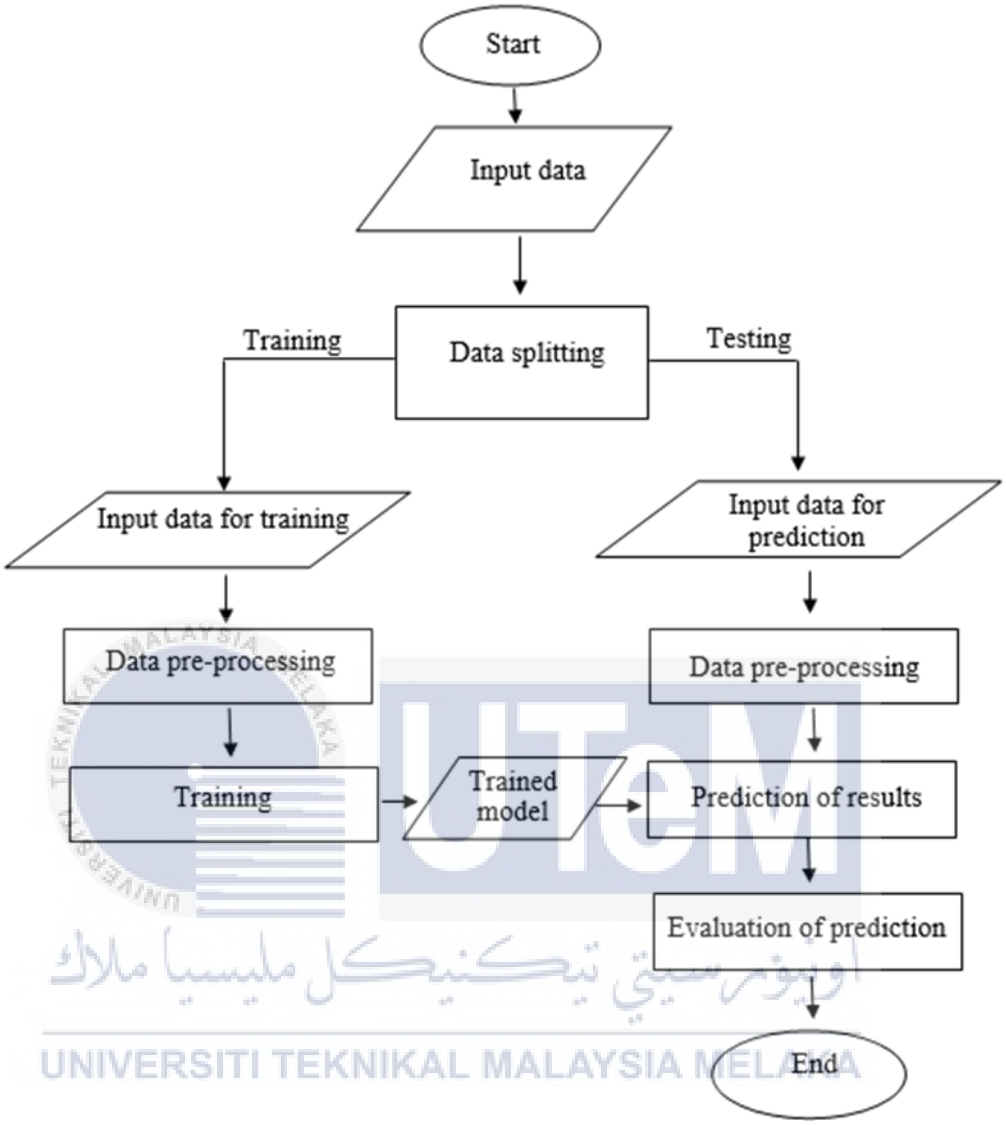

Figure 3.4: Flowchart of system

Figure 3.4 shows the flowchart of prediction system. First, the system started with input data which contain temperature data, rainfall data, humidity data, planted area of rubber tree data and production of rubber data. Then, this dataset will split into training subset and testing subset. The training and testing subset will undergo the next step with data pre-processing. The data are trained and trained model is composed. Then, the testing subset is used for evaluating the trained model and the prediction

results are computed. After that, the prediction results by using various types of machine learning algorithms are evaluated.

3.3.1 Data Splitting

In Pycharm, Pandas library is used because it is a software library especially for the Python programming language in data analysis and data manipulation. The CSV files datasets are imported to Python by using 'pandas.read_csv' because this command can read a CSV file into DataFrame. The next step is extracting the independent variables and dependent variable. For instance, the independent variables of dataset are temperature data, rainfall data, humidity data and planted area of rubber tree. The dependent variable is the total production of rubber data. Then, the train_test_split is imported from scikit learn library and the dataset is split into 80:20. In consequence, 80% of dataset is used for training subset and 20% of dataset is used for testing subset. The year 2000 to 2008 is chosen for training subset while the year 2009 to 2011 is used for testing subset. The training subset is used to train the model while the testing subset is used to test the accuracy of the final model that fits the training model.

3.3.2 Data Pre-processing

Data pre-processing is the procedure for preparing a suitable raw data in a machine learning model. Feature scaling is the crucial step for data pre-processing. First, the relevant Python library is imported in Pycharm which is StandardScaler from scikit learn library. StandardScaler is used in data standardisation to perform feature scaling. The training and testing datasets are standardised in such a way that the variables of the dataset lie within a specific range.

3.3.3 Training, Testing and Evaluating Process

During the process of training, the input data of this system are prepared along with desired output. Hyperparameter optimisation is used to select the suitable hyperparameter for machine learning algorithm. A hyperparameter is a value for a parameter that is used to manage the learning process of algorithm. GridSearchSV is used in fine tuning the prediction model and figuring out optimal hyperparameters. The parameters chosen for Random Forest algorithm are 'bootstrap', 'max_depth', 'max_features', 'min_samples_leaf', 'min_samples_split' and 'n_estimators'. Besides, the parameters selected for Decision Tree algorithm are 'splitter', 'max_depth', 'max_features', 'min_samples_leaf', 'max_leaf_nodes' and 'min_weight_fraction_leaf'. After that, regularisation method is applied on Linear Regression algorithm and fine tuning the 'alpha' value. Last but not least, the parameter chosen for fine tuning in Neural Network are 'hidden_layer_sizes', 'activation' and 'solver'. After fine tuning, the ideal values of parameters are obtained and fit into machine learning algorithm.

Then, the input data are fit into machine learning model because a computational algorithm can optimally configures itself along with the training inputs. The algorithm can select and weight the input data to provide the most decisive outcomes as it have variable numerical parameter that are adjusted through iterative optimization [37]. There are lots of computational pathways are generated during the training process. After the training subset is fit into model, the trained model is produced. Then, the trained model is evaluated from the testing subset and the final prediction output are obtained. Lastly, the performance of the prediction model are evaluated by computing the regression metrics which are Mean Squared Error (MSE) and Mean Absolute Error (MAE).

3.4 Environment and Sustainability

The purpose of this proposed prediction system is to predict the rubber production. Machine learning helps the prediction system to achieve better accuracy. This project is user friendly because Python language is implemented and this is one of the popular programming language. Thus, this system has high sustainability because the code can be easily read and modified when there is error occurred. Furthermore, the code can be improved if this project needs enhancement. This project can predict other types of crops instead of rubber production if there are sufficient and available datasets. Besides, this prediction system does not require labour force thus it is cost effective and high sustainability. Moreover, good farm planning will provide efficient utilization of all the available environment resources.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

In this section, the details of the result are discussed while all the results are presented in table and graph. The prediction results are computed by using four types of machine learning algorithms which are Random Forest, Decision Tree, Linear Regression and Neural Network. Different prediction models are done to evaluate the performance of machine learning algorithms. The datasets are prepared from four different states which are Melaka, Perak, Pahang and Johor. These datasets are used and analysed in each prediction model to compare the accuracy of machine learning algorithms.

4.2 Initial Prediction Results

The initial prediction results of prediction of rubber production from year 2009 to 2011 are obtained from Random Forest, Decision Tree, Linear Regression and Neural Network. The initial results before applying the optimisation is shown in this section. In this experiment, default parameter is used for each algorithms. The prediction results of Melaka's rubber production with Mean Squared Error (MSE) and Mean Absolute Error (MAE) values are tabulated.

Table 4.1: Results before hyperparameter optimisation

Type of model	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Random Forest	2009	2003	2130.24	0.4262	0.5176
	2010	1387	2159.47		
	2011	1479	3295.9		
Decision Tree	2009	2003	1966.67	0.2334	0.4009
	2010	1387	2467		
	2011	1479	2467		
Linear Regression	2009	2003	1960.65	0.0179	0.1015
	2010	1387	1287.18		
	2011	1479	1869.46		
Neural Network	2009	2003	1988.27	0.0666	0.2126
	2010	1387	1989.02		
	2011	1479	1978.04		

From Table 4.1, it is clearly shown that the differences between predicted and actual results obtained from Random Forest and Decision Tree are relatively high since highest value of MSE and MAE are obtained. In contrast, MSE and MAE values from Linear Regression are the lowest. The improvement of prediction results can be done with optimisation by tuning the hyperparameters of each prediction models. Table 4.2 shows the prediction results after hyperparameter optimisation. The

comparison of results of before and after hyperparameter tuning are tabulated in Table 4.3 and 4.4.

Table 4.2: Results after hyperparameter optimisation

Type of model	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Random Forest	2009	2003	1919.5	0.3946	0.4782
	2010	1387	2020.25		
	2011	1479	3272.6		
Decision Tree	2009	2003	1919.5	0.0794	0.2383
	2010	1387	2114		
	2011	1479	1919.5		
Linear Regression	2009	2003	2072.94	0.0089	0.0880
	2010	1387	1182.33		
	2011	1479	1666.08		
Neural Network	2009	2003	1988.27	0.0446	0.1571
	2010	1387	1989.02		
	2011	1479	1978.04		

Table 4.3: Comparison between the MSE values

Type of model	Year	MSE value before tuning	MSE value after tuning	Difference between before and after tuning
Random Forest	2009	0.4262	0.3946	0.0316
	2010			
	2011			
Decision Tree	2009	0.2334	0.0794	0.1540
	2010			
	2011			
Linear Regression	2009	0.0179	0.0089	0.0009
	2010			
	2011			
Neural Network	2009	0.0666	0.0446	0.0220
	2010			
	2011			

The optimal hyperparameters are defined and selected through tuning process until the most appropriate hyperparameters are applied with prediction models. Table 4.2 shows that some of the predicted results are much closer to the actual results. The MSE value in each model also decreasing. By referring the Table 4.3, the value of MSE in Decision Tree decreases from 0.2334 to 0.0794 and it has been reduced by 0.1540. The MSE values of Random Forest, Linear Regression and Neural Network are recorded the lowest by 0.0316, 0.0009 and 0.0220 respectively.

Table 4.4: Comparison between the MAE values

Type of model	Year	MAE value before tuning	MAE value after tuning	Difference between before and after tuning
Random Forest	2009	0.5176	0.4782	0.0394
	2010			
	2011			
Decision Tree	2009	0.4009	0.2383	0.1626
	2010			
	2011			
Linear Regression	2009	0.1015	0.0880	0.0135
	2010			
	2011			
Neural Network	2009	0.2126	0.1571	0.0555
	2010			
	2011			

Moreover, the value of MAE in each model also decreases when it is comparing with the value before tuning the parameter. Table 4.4 shows that the Decision Tree MAE values decrease from 0.4009 to 0.2383, it has been reduced by 0.1626. The MAE value for Random Forest, Decision Tree and Neural Network has the smallest by 0.0394, 0.0135 and 0.0555 respectively.

4.3 Final Prediction Results

After debugging the program in PyCharm, the prediction results are obtained from the system and Random Forest, Decision Tree, Linear Regression and Neural Network are implemented in this program. The datasets used in the system which include climate factors, planted area of rubber tree and rubber production in Melaka, Perak, Pahang and Johor. The time period of dataset is 12 years which is collected from 2000 to 2011. Figure 4.4 shows the interface of dataset in CSV file.

MeanTempMax	MeanTempMin	Rainfall	MeanHumid	PlantedArea	Production
32	23.9	2156.7	82.4	6355	7688
31.9	23.9	2220.1	82.4	4252	3883
32.8	24.1	2082.3	80.1	3810	2706
32.2	24.1	1824.2	81.4	3122	2855
32.5	24.2	1615.5	80.5	2380	2202
32.5	24.4	1875.5	77.6	2014	2061
32.1	24.1	2178.1	79.7	1753	2079
31.9	24.1	2078.6	79.3	1717	1894
31.8	24.2	2194.4	78.7	1585	1945
31.8	24.3	1878	78.9	1452	2003
32.1	24.4	1992.8	79.7	1119	1387
31.6	24.1	1845.2	82	1146	1479

Figure 4.1: Interface of dataset

The measurement unit of temperature data is degree Celsius (°C), while the rainfall is millimetres per hour (mm h–1). Apart from that, the relative humidity is expressed as percentage (%), measurement unit of planted area of rubber tree is hectares and lastly the measurement unit of rubber production is tonnes. The data from 2000 to 2008 are selected for training while data from 2009 to 2011 are selected for testing. The predicted output versus actual output of rubber production are illustrated in next section along with MAE and MSE values.

4.3.1 Results from Random Forest Algorithm

Random Forest algorithm is applied and analysed. As a result, different prediction performance of rubber production is generated and computed according to the state of each dataset.

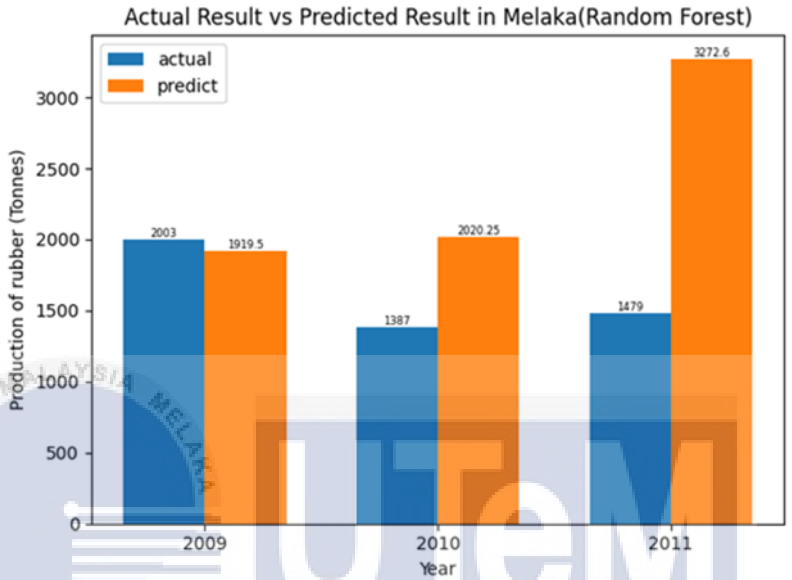

Figure 4.2: Prediction results in Melaka using Random Forest

Table 4.5: Results in Melaka using Random Forest

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Random Forest	Melaka	2009	2003	1919.5	0.3946	0.4782
		2010	1387	2020.25		
		2011	1479	3272.6		

From the Figure 4.2, the blue colour bars represent the actual rubber production while the orange colour bars represent the predicted rubber production in 2009, 2010 and 2011. In 2009, the predicted result in Melaka is less than actual result yet the difference between both results is relatively low. In 2010, the predicted result is more

than actual result and there is gap between both results. In 2011, the predicted result is greater than actual result and the differences between both results are relatively high. The MSE value obtained is 0.3946 while MAE value computed is 0.4782.

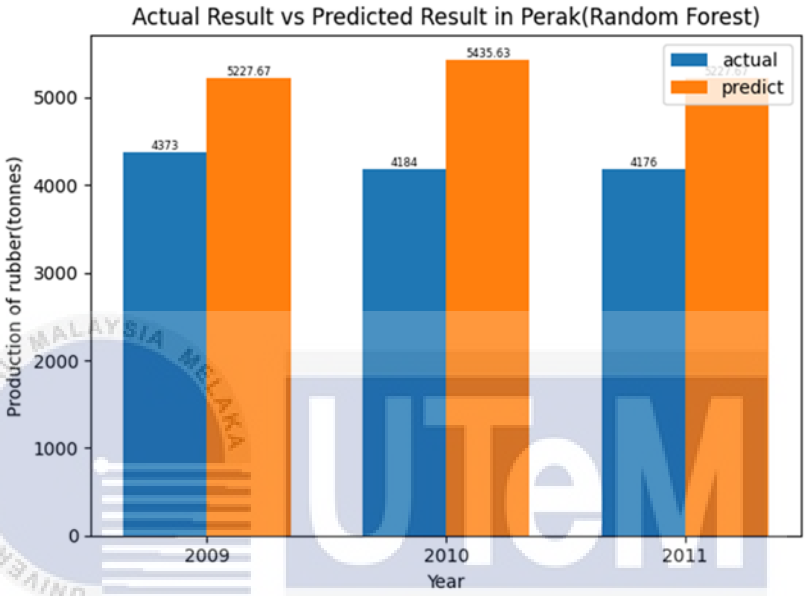

Figure 4.3: Prediction results in Perak using Random Forest

Table 4.6: Results in Perak using Random Forest

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Random Forest	Perak	2009	4373	5227.67	0.1222	0.3454
		2010	4184	5435.63		
		2011	4176	5227.67		

From year 2009 to 2011, the predicted rubber production in Perak results are greater than actual results as shown in Figure 4.3. The differences between both results are

relatively high in 2009, 2010 and 2011. The MSE value obtained is 0.1222 while MAE value computed is 0.3454.

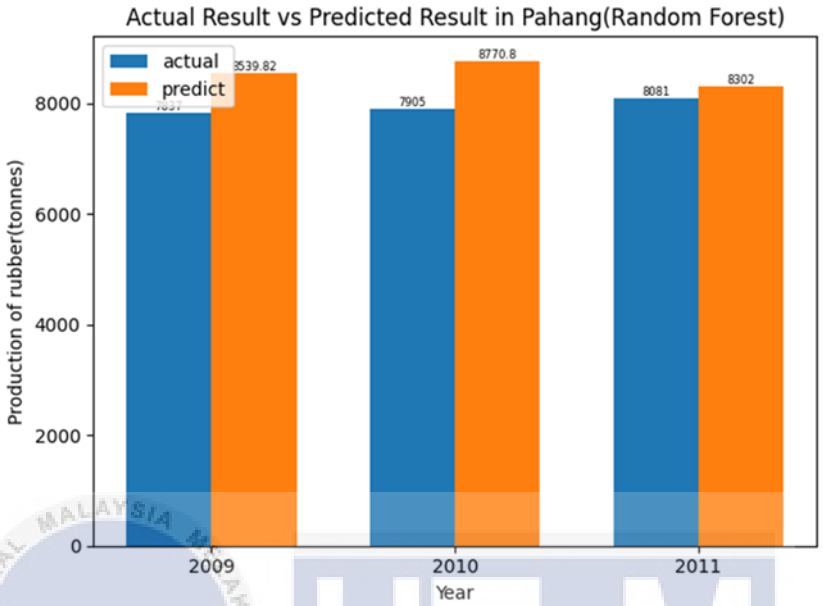

Figure 4.4: Prediction results in Pahang using Random Forest

Table 4.7: Results in Pahang using Random Forest

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Random Forest	Pahang	2009	7837	8539.83	0.0618	0.2259
		2010	7905	8770.8		
		2011	8081	8302		

As we can see that the prediction result in Pahang from year 2009 to 2011 in Figure 4.4, the value of predicted rubber production are higher than value of actual rubber production. The differences between both values in 2010 are relatively high. The difference between predicted and actual results is relatively small in 2011 while the

values between predicted rubber production and actual rubber production are similar. The MSE value obtained is 0.0618 while MAE value computed is 0.2259.

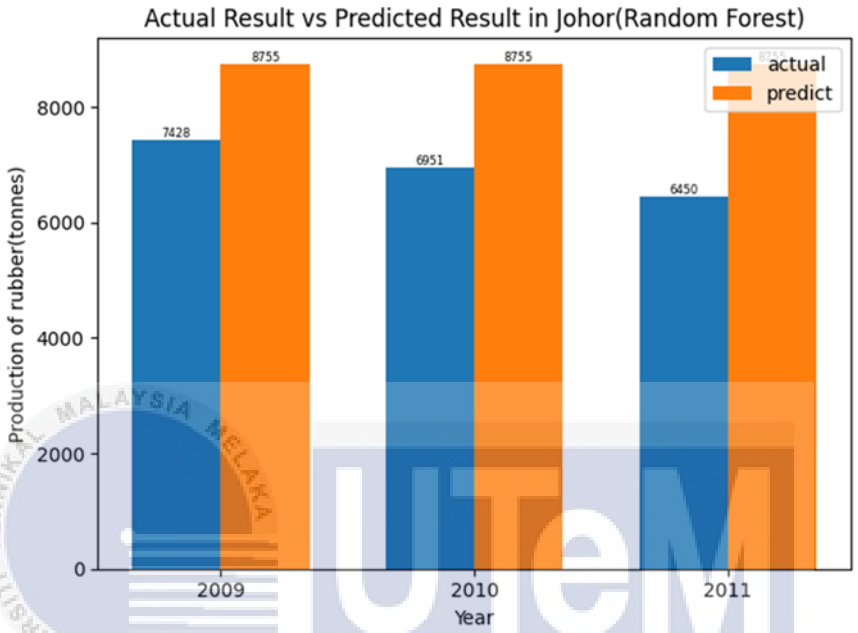

Figure 4.5: Prediction results in Johor using Random Forest

Table 4.8: Results in Johor using Random Forest

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Random Forest	Johor	2009	7428	8755	0.4004	0.6180
		2010	6951	8755		
		2011	6450	8755		

Figure 4.5, it represents the prediction results in Johor from year 2009 to 2011, the predicted results of rubber production are greater than actual results of rubber production. The values of predicted rubber production are constant and the differences

between predicted results and actual results are relatively high throughout these 3 years. The MSE value obtained is 0.4004 while MAE value computed is 0.6180.

4.3.2 Results from Decision Tree Algorithm

In this part, prediction results of rubber production are computed according to the state of each dataset using decision tree is illustrated.

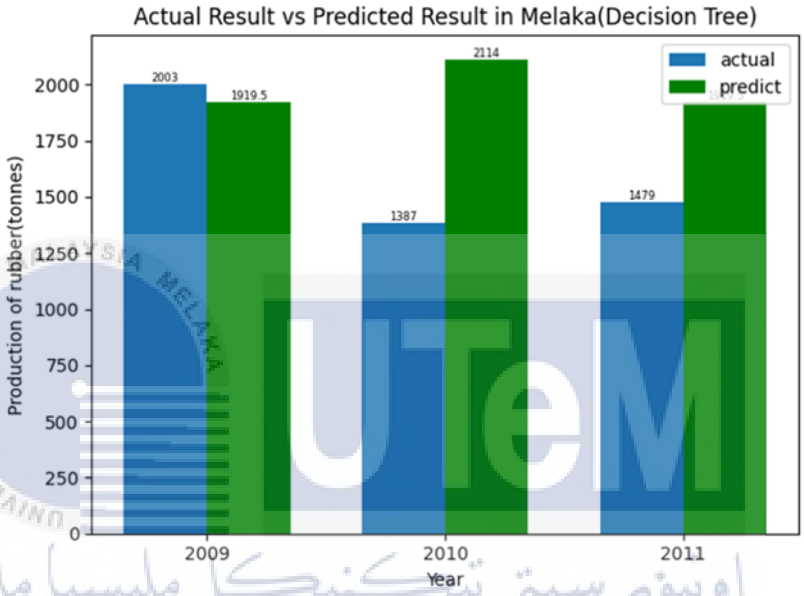

Figure 4.6: Prediction results in Melaka using Decision Tree

Table 4.9: Results in Melaka using Decision Tree

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Decision Tree	Melaka	2009	2003	1919.5	0.0794	0.2383
		2010	1387	2114		
		2011	1479	1919.5		

The blue colour bars in Figure 4.6 represent the actual results of rubber production while green colour bars represent the predicted results of rubber production in Melaka.

In 2009, the predicted value of rubber production is lower than actual value of rubber production. However, the predicted values of rubber production are higher than actual values of rubber production in year 2010 and 2011. The differences between both values are relatively high especially the results in 2010. The value of MSE computed is 0.0794 while value of MAE is 0.2383.

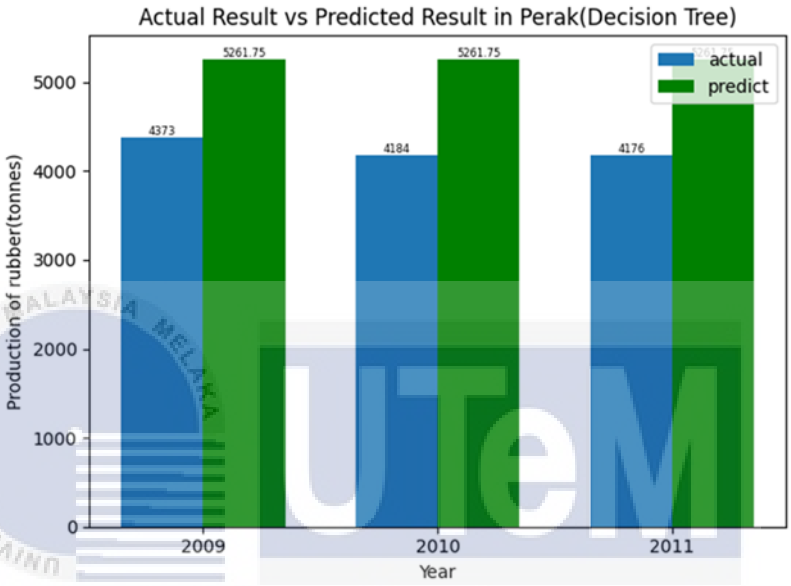

Figure 4.7: Prediction results in Perak using Decision Tree

Table 4.10: Results in Perak using Decision Tree

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Decision Tree	Perak	2009	4373	5261.75	0.1123	0.3339
		2010	4184	5261.75		
		2011	4176	5261.75		

Figure 4.7 illustrates the prediction results in Perak from year 2009 to 2011. The predicted results of rubber production are greater than actual results of rubber production. The values of predicted rubber production are consistent and the

differences between both results are relatively high throughout these 3 years. The value of MSE computed is 0.1124 while the value of MAE computed is 0.3339.

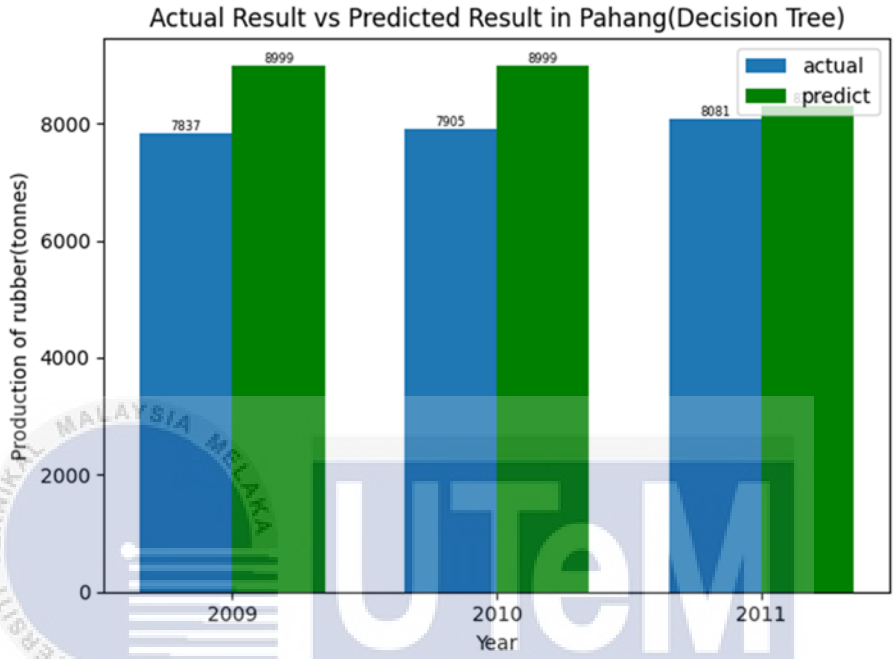

Figure 4.8: Prediction results in Pahang using Decision Tree

Table 4.11: Results in Pahang using Decision Tree

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Decision Tree	Pahang	2009	7837	8999	0.1241	0.3127
		2010	7905	8999		
		2011	8081	8302		

Figure 4.8 displays the prediction results in Pahang from year 2009 to 2011. The predicted results of rubber production are more than actual results of rubber production. The values of predicted rubber production are the same in year 2009 and

2010 and there are significant deviation between values of predicted results and actual results. In 2011, the predicted result of rubber production is close to the actual results thus the difference between both results is relatively low. The value of MSE is 0.1241 while the value of MAE obtained is 0.3127.

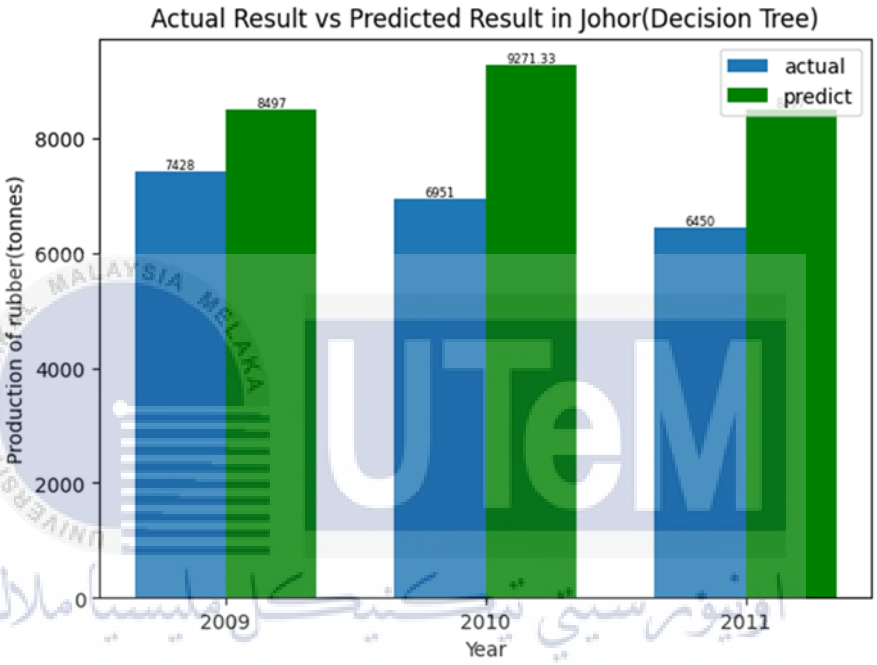

Figure 4.9: Prediction results in Johor using Decision Tree

Table 4.12: Results in Johor using Decision Tree

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Decision Tree	Johor	2009	7428	8497	0.4155	0.6180
		2010	6951	9271.33		
		2011	6450	8497		

As we can see from year 2009 to 2011, Johor is able to record the predicted results of rubber production are greater than actual results of rubber production as shown in Figure 4.9. There are tremendous deviation between values of actual rubber

production and predicted rubber production. The difference of values between both results is relatively high in 2010. The MSE value computed is 0.4155 while MAE value obtained is 0.6180.

4.3.3 Results from Linear Regression Algorithm

In this section, prediction results of rubber production are generated according to the state of each dataset using linear regression is explained.

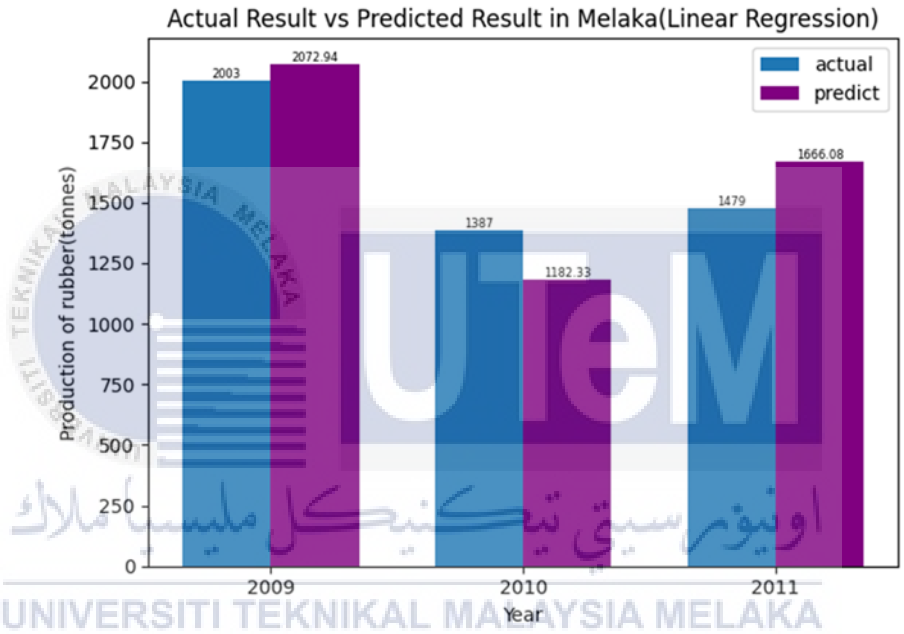

Figure 4.10: Prediction results in Melaka using Linear Regression

Table 4.13: Results in Melaka using Linear Regression

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Linear Regression	Melaka	2009	2003	2072.94	0.0089	0.0880
		2010	1387	1182.33		
		2011	1479	1666.08		

The blue colour bars represent the actual values of rubber production while purple colour bars represent the predicted values of rubber production in Melaka. The

differences between the actual results and predicted results throughout these 3 years are relatively low especially the results in 2009 as shown in Figure 4.10. In year 2009 and 2011, the predicted values of rubber production are slightly greater than actual values of rubber production. In contrast, the predicted value of rubber production is less than actual value of rubber production in 2010. The MSE value computed is 0.0089 while MAE value is 0.0880.

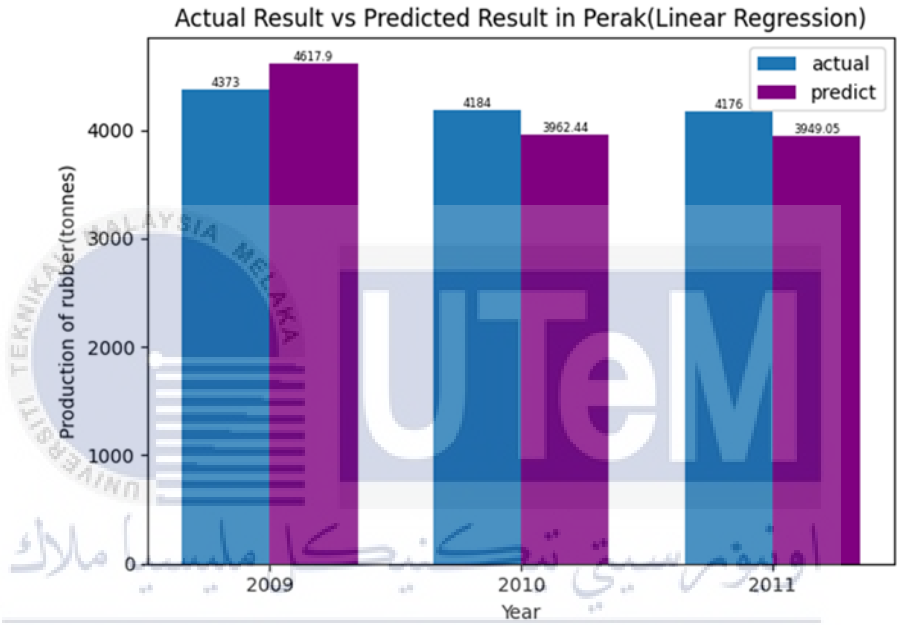

Figure 4.11: Prediction results in Perak using Linear Regression

Table 4.14: Results in Perak using Linear Regression

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Linear Regression	Perak	2009	4373	4617.9	0.0058	0.0758
		2010	4184	3962.44		
		2011	4176	3949.05		

The illustration of the prediction results in Perak is shown as Figure 4.11. The value predicted in 2009 is higher than the actual value of rubber production while the values predicted in year 2010 and 2011 lower than the actual value of rubber production. The differences between the values of actual and predicted rubber production are relatively small throughout these 3 years. The MSE value is 0.0058 and the MAE value computed is 0.0758.

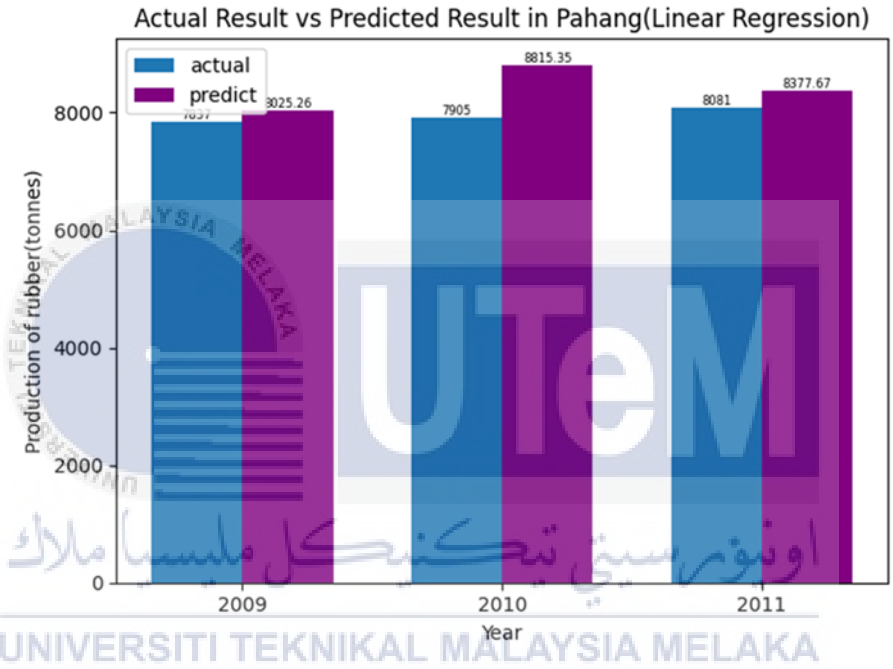

Figure 4.12: Prediction results in Pahang using Linear Regression

Table 4.15: Results in Pahang using Linear Regression

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Linear Regression	Pahang	2009	7837	8025.26	0.0455	0.1761
		2010	7905	8815.35		
		2011	8081	8377.67		

Figure 4.12 displays the prediction results in Pahang from year 2009 to 2011. The values predicted are higher than the actual values of rubber production. In 2009 and 2011, the differences between the actual and predicted values of rubber production are relatively low. However, there is significant deviation between actual and predicted values of rubber production in 2010. The MSE value obtained is 0.0455 while the MAE value computed is 0.1761.

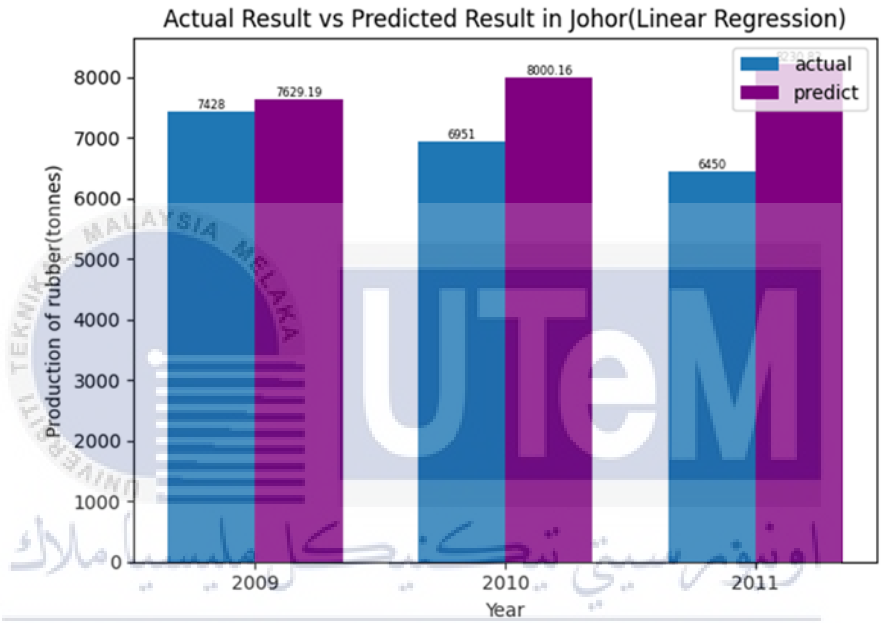

Figure 4.13: Prediction results in Johor using Linear Regression

Table 4.16: Results in Johor using Linear Regression

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Linear Regression	Johor	2009	7428	7629.19	0.1672	0.3446
		2010	6951	8000.16		
		2011	6450	8230.82		

All values are predicted higher than actual values of rubber production in Johor from year 2009 to 2011. In 2009, the difference between actual and predicted results of rubber production is relatively low as shown in Figure 4.13. However, there are tremendous deviation between actual and predicted results in 2010 and 2011. The MSE and MAE values are computed which are 0.1672 and 0.3446 respectively.

4.3.4 Results from Neural Network

The prediction results of rubber production are computed according to the state of each dataset using neural network is described.

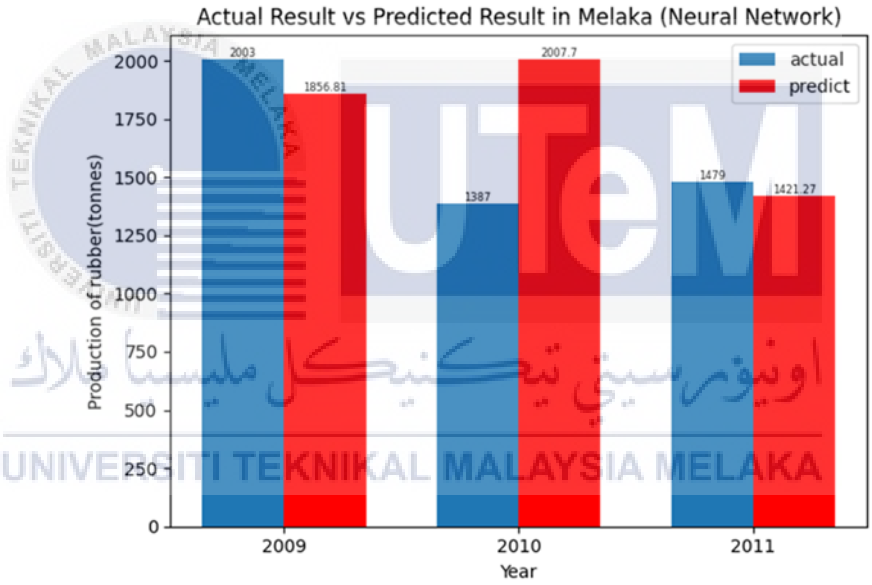

Figure 4.14: Prediction results in Melaka using Neural Network

Table 4.17: Results in Melaka using Neural Network

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Neural Network	Melaka	2009	2003	1856.81	0.0446	0.1571
		2010	1387	2007.7		
		2011	1479	1421.27		

As can be seen from Figure 4.14, the blue colour bars represent the actual values of rubber production in Melaka while the red colour bars represent the predicted values of rubber production. In 2009 and 2011, the values predicted are less than the actual values of rubber production. In contrast, the value predicted is significantly greater than the actual value of rubber production. The differences between actual and predicted values in 2009 and 2011 are relatively low while huge deviation between actual and predicted values is shown in 2010. The MSE and MAE values are computed which are 0.0446 and 0.1571 respectively.

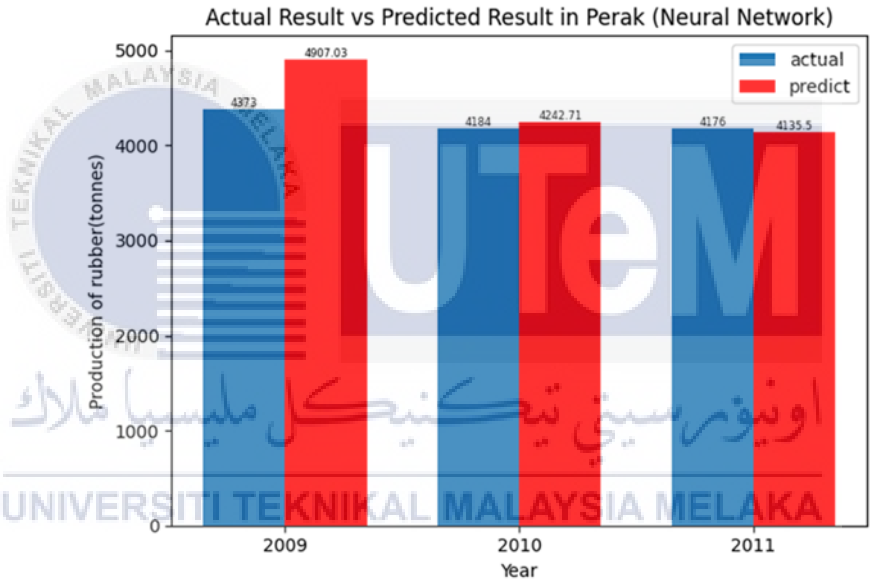

Figure 4.15: Prediction results in Perak using Neural Network

Table 4.18: Results in Perak using Neural Network

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Neural Network	Perak	2009	4373	4907.03	0.0104	0.0693
		2010	4184	4242.71		
		2011	4176	4135.5		

Figure 4.15 shows the prediction results in Perak for year 2009 and 2010. The predicted result are higher than the actual values of rubber production. In 2011, the value predicted is slightly lower than actual value of rubber production. There is significant deviation between actual and predicted values in 2009 while the differences between actual and predicted values are relatively low as shown in 2010 and 2011. The MSE and MAE values are computed which are 0.0104 and 0.0693 respectively.

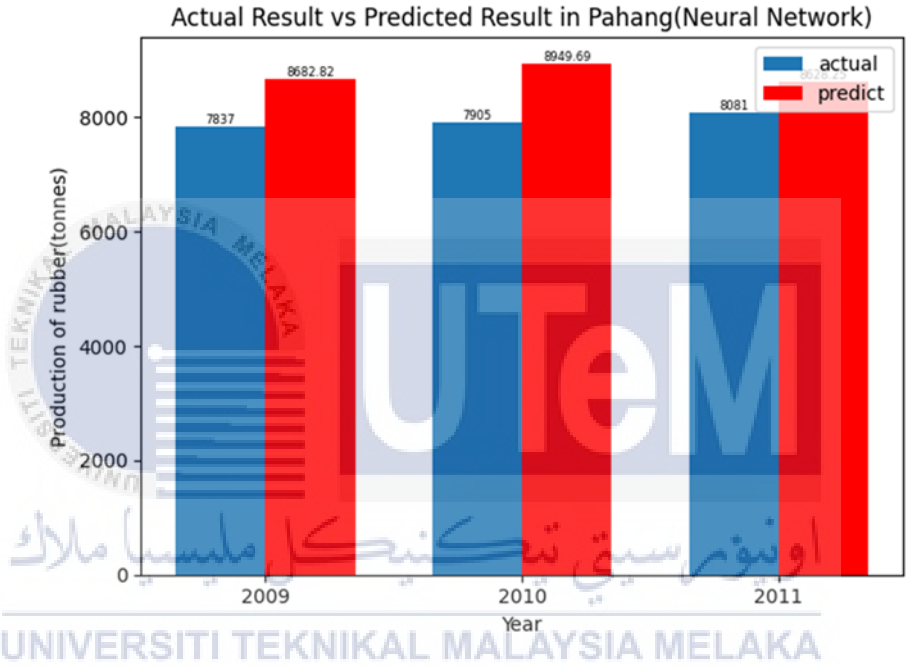

Figure 4.16: Prediction results in Pahang using Neural Network

Table 4.19: Results in Pahang using Neural Network

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Neural Network	Pahang	2009	7837	8682.82	0.1007	0.3077
		2010	7905	8949.69		
		2011	8081	8628.25		

The prediction results in Pahang is shown as Figure 4.16 and all the values predicted are higher than the actual values of rubber production. There are significant deviation between actual and predicted results of rubber production. The MSE and MAE values are computed which are 0.1007 and 0.3077 respectively.

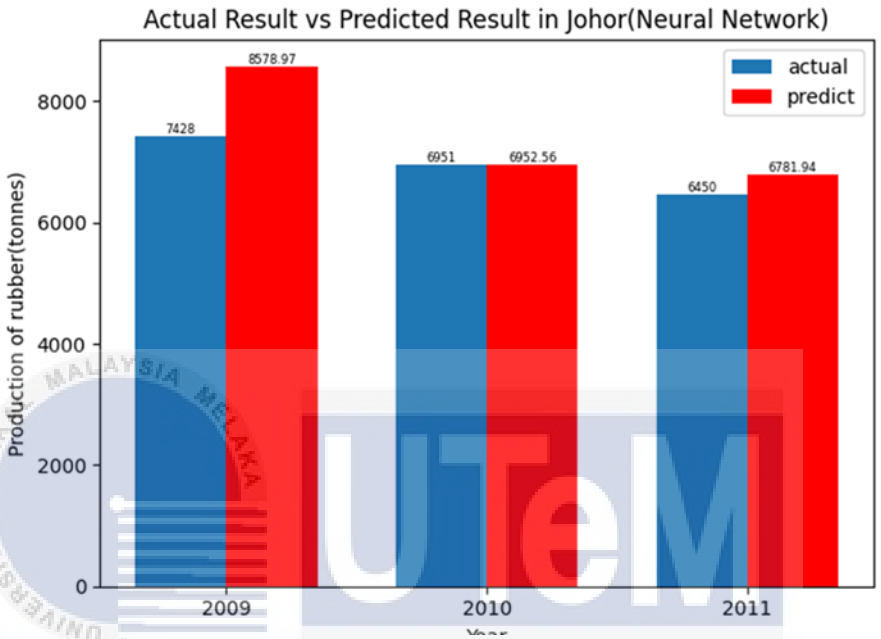

Figure 4.17: Prediction results in Johor using Neural Network

Table 4.20: Results in Johor using Neural Network

Type of model	State	Year	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Neural Network	Johor	2009	7428	8578.97	0.0556	0.1688
		2010	6951	6952.56		
		2011	6450	6781.94		

Meanwhile, in Johor all the values predicted are slightly higher than the actual values of rubber production as stated in Figure 4.17. The differences between actual and predicted results in 2010 and 2011 are relatively low while significant deviation

between actual and predicted result is shown in 2010. The MSE value is 0.0556 while the MAE value obtained is 0.1688.

4.4 Discussion

There are 4 different types of machine learning models which are being used in this proposed system. Performance of machine learning algorithms are being compared to determine which of them provides the accurate results. The machine learning models as Random Forest, Decision Tree, Linear Regression and Neural Network are used in this work. All the datasets contains several features such as climate factors, planted area of rubber tree and total number of rubber production. These datasets are collected yearly from year 2000 to 2011. There are 4 available datasets can be obtained from official website of Department of Statistic Malaysia and website of Open Government Data Malaysia and the source of datasets are from Melaka, Perak, Pahang and Johor. All the predicted results, value of Mean Squared Error (MSE) and Mean Absolute Error (MAE) are computed and tabulated as table below.

Table 4.21: Comparison between prediction models

Type of model	State	Actual Result of Rubber Production (tonnes)	Predicted Result of Rubber Production (tonnes)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	Average MSE value	Average MAE value
Random Forest	Melaka	2003	1919.5	0.3946	0.4782	0.2448	0.4169
		1387	2020.25				
		1479	3272.6				
	Perak	4373	5227.67	0.1222	0.3454		
		4184	5435.63				
		4176	5227.67				
	Pahang	7837	8539.83	0.0618	0.2259		
		7905	8770.8				
		8081	8302				
	Johor	7428	8755	0.4004	0.6180		
		6951	8755				
		6450	8755				

Decision Tree	Melaka	2003	1919.5	0.0794	0.2383	0.1828	0.3757
		1387	2114				
		1479	1919.5				
	Perak	4373	5261.75	0.1123	0.3339		
		4184	5261.75				
		4176	5261.75				
	Pahang	7837	8999	0.1241	0.3127		
		7905	8999				
		8081	8302				
	Johor	7428	8497	0.4155	0.6180		
		6951	9271.33				
		6450	8497				
Linear Regression	Melaka	2003	2072.94	0.0089	0.0880	0.0569	0.1711
		1387	1182.33				
		1479	1666.08				
	Perak	4373	4617.9	0.0058	0.0758		
		4184	3962.44				
		4176	3949.05				
	Pahang	7837	8025.26	0.0455	0.1761		
		7905	8815.35				
		8081	8377.67				
	Johor	7428	7629.19	0.1672	0.3446		
		6951	8000.16				
		6450	8230.82				
Neural Network	Melaka	2003	1856.81	0.0446	0.1571	0.0528	0.1757
		1387	2007.7				
		1479	1421.27				
	Perak	4373	4907.03	0.0104	0.0693		
		4184	4242.71				
		4176	4135.5				
	Pahang	7837	8682.82	0.1007	0.3077		
		7905	8949.69				
		8081	8628.25				
	Johor	7428	8578.97	0.0556	0.1688		
		6951	6952.56				
		6450	6781.94				

The actual results are obtained from the dependent variables from testing subset and the results are represented as the actual total number of rubber production. The results are predicted from independent variables in testing subset and the predicted results are predicted according to the total number of rubber production. The

measurement unit of rubber production is in tonnes. The evaluation of each model are calculated in order to compare the prediction performance of each algorithm.

From Table 4.21, the maximum MSE and MAE values computed in Random Forest which are 0.4004 and 0.6180 respectively which are obtained from prediction of rubber production in Johor. The minimum values of MSE and MAE are 0.0618 and 0.2259 are recorded from prediction of rubber production in Pahang. The average MSE value is 0.2448 while the MAE value is 0.4169. In Decision Tree, the maximum MSE and MAE values in Johor rubber production are 0.4155 and 0.6180 respectively. However, the minimum values of MSE and MAE are computed from Melaka, which are 0.0794 and 0.2383. The average values of MSE and MAE are 0.1828 and 0.3757 respectively. In Linear Regression, the maximum MSE and MAE values are 0.1672 and 0.3446 in Johor. The minimum values of MSE and MAE are obtained from Perak, which are 0.0058 and 0.0758. The average values of MSE and MAE are 0.0569 and 0.1711. In Neural Network, the maximum MSE and MAE values are computed from prediction of rubber production in Pahang and the values are 0.1007 and 0.3077 respectively. However, the minimum values of MSE and MAE are 0.0104 and 0.0693 are obtained from prediction of rubber production in Perak. The average values of MSE and MAE are 0.0528 and 0.1757.

The value of MAE is considered for selecting the most accurate machine learning algorithm in crop yield prediction system. According to article “Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in assessing the average model performance” [38], the authors mentioned that MAE is preferred for comparisons of average error of model performance. The reason is MSE or RMSE calculates the average error depends on sum of squared errors. This means that as the

total error is concentrated within a smaller number of increasingly large individual errors, the total square error will increase [38]. The sum of squared error is then divided by n and the value of MSE or RMSE will rise along with sum of squared error because the variance associated with the frequency distribution of error magnitudes also grows. However, MAE only calculate the total absolute error and divided by n. The value of MAE only varies within the set of error magnitudes [38]. Apparently, MSE or RMSE is not a proper metric in measuring the average error to compare the performance of model. By referring the average MAE values from table 4.21, Linear Regression is selected because it contributes the least value of MAE among the machine learning algorithms.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Introduction

In this chapter, the results and discussion is summarised as the conclusion of this work. The future improvement of this work also is been suggested and further discussed to improve the quality of this work.

5.2 Conclusion

Prediction crop yield system using machine learning is developed and successfully analyse using several machine learning algorithms such as Random Forest, Decision Tree, Linear Regression and Neural Network. Four datasets from different states are applied and these datasets are collected from Melaka, Perak, Pahang and Johor from year 2000 to 2011. In this work, the training subset are selected from 2000 to 2008 and testing subset are selected from 2009 to 2011. Data pre-processing is important

because this process will standardise the variables of dataset within certain range. Then, fine tuning the prediction model is done to figure out which hyperparameters are most suitable for specific machine learning algorithm. This step is taken to ensure that the training data fit better in machine learning algorithm and tends to generate the least error rate. Thus, each machine learning algorithm has undergone the process of choosing the hyperparameters while predicting the rubber production. Then, the prediction results of rubber production are calculated according to states using various prediction models. The actual and predicted results of rubber production from year 2009 to 2011 are then compared and illustrated. The performance of prediction models is evaluated by computing the Mean Square Error (MSE) and Mean Absolute Error (MAE). Each machine learning algorithms are compared to select the most accurate prediction model which able to produce an outstanding performance. As conclusion, Linear Regression contributes the least value of MAE and computes the most accurate prediction results among the machine learning algorithms.

Choosing a suitable machine learning algorithm is crucial step before developing prediction system. An accurate prediction crop yield system can help farmers in financial decisions, marketing and insurance [39]. Furthermore, policy and decision making can be made earlier when the prediction system shows there is crop production shortage in the future.

5.3 Future Work

The dataset in this project only consists of limited number of features such as temperature data, rainfall data, humidity data and rubber tree planted area to predict the rubber production. This system can be improved by adding more features which is believed to give an improvement of the production of rubber. Some climatic factors

are can be considered can be added such as vapour pressure, wet day frequency, sunlight and cloud cover. Moreover, soil parameters also can be measured as agriculture feature. For instance, the pH value of soil, soil salinity and soil moisture also could be included in soil parameters. Additional sensors and instruments are necessary to be implemented in rubber tree estate in order to obtain real-time data. Hence, this future work is not only capable to predict the rubber production, but also is able to suggest the fertiliser composition by taking account the soil nutrients level.

REFERENCES

1. "PRESS RELEASE SELECTED AGRICULTURAL INDICATORS, MALAYSIA, 2019," *DEPARTMENT OF STATISTICS MALAYSIA*. Nov. 29, 2019.
2. A. A. Khin, R. L. L. Bin, S. B. Kai, K. L. L. Teng, and F. Y. Chiun, "Challenges of the Export for Natural Rubber Latex in the ASEAN Market," *IOP Conference Series: Materials Science and Engineering*, vol. 548, no. 1, Aug. 2019.
3. T. B. Sapkota, M. L. Jat, R. K. Jat, P. Kapoor, and C. Stirling, "Yield Estimation of Food and Non-Food Crops in Smallholder Production Systems," *Methods for Measuring Greenhouse Gas Balances and Evaluating Mitigation Options in Smallholder Agriculture*, Springer International Publishing, pp. 163–174, 2016.
4. V. Sellam and E. Poovammal, "Prediction of Crop Yield Using Regression Analysis," *Indian Journal of Science and Technology*, vol. 9, no. 38, 2016.
5. M. Champaneri, D. Chachpara, C. Chandvidkar, and M. Rathod, "Crop Yield Prediction Using Machine Learning," *International Journal of Science and Research (IJSR)*, 2020.

6. “Methodology for Estimation of Crop Area and Crop Yield Under Mixed and Continuous Cropping Publication Prepared in the Framework of the Global Strategy to Improve Agricultural and Rural Statistics,” *Food and Algriculture Organisation of the United Nations*. Mar. 2017.
7. P. Charoen-Ung and P. Mittrapiyanuruk, “Sugarcane Yield Grade Prediction Using Random Forest With Forward Feature Selection and Hyper-parameter Tuning,” *Advances in Intelligent Systems and Computing*, vol. 769, pp. 33–42, 2019.
8. B. Fulkerson, D. Michie, D. J. Spiegelhalter, and C. C. Taylor, “Machine Learning, Neural and Statistical Classification,” *Technometrics*, vol. 37, no. 4, p. 459, Nov. 1995.
9. P. C. Austin and E. W. Steyerberg, “The Number of Subjects Per Variable Required in Linear Regression Analyses,” *Journal of Clinical Epidemiology*, vol. 68, no. 6, pp. 627–636, Jun. 2015.
10. M. Belgiu and L. Drăgu, “Random Forest in Remote Sensing: A Review of Applications and Future Directions,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114. Elsevier B.V., pp. 24–31, Apr. 01, 2016.
11. M. Batra and R. Agrawal, “Comparative Analysis of Decision Tree Algorithms,” *Advances in Intelligent Systems and Computing*, vol. 652, pp. 31–36, 2018.
12. R. Trifonov, R. Yoshinov, G. Pavlova, and G. Tsochev, “Artificial Neural Network Intelligent Method for Prediction,” *AIP Conference Proceedings*, Sep. 2017, vol. 1872.
13. M. H. M. Hazir, R. A. Kadir, E. Gloor, and D. Galbraith, “Effect of Agroclimatic Variability on Land Suitability for Cultivating Rubber (Hevea

brasiliensis) and Growth Performance Assessment in the Tropical Rainforest Climate of Peninsular Malaysia," *Climate Risk Management*, vol. 27, Jan. 2020.

14. A. Shah, A. Dubey, V. Hemnani, D. Gala, and D. R. Kalbande, "Smart Farming System: Crop Yield Prediction Using Regression Techniques," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 19, Springer, pp. 49–56, 2018.
15. L. Girish, S. Gangadhar, T. R. Bharath, K. S. Balaiji, K. T. Abhishek, "Crop Yield and Rainfall Prediction in Tumakuru District using Machine Learning," *National Conference on Technology for Rural Development (NCTFRD-18)*, 2018.
16. T. Osman, S. Shahjahan Psyche, M. Rafik Kamal, F. Tamanna, F. Haque, and R. M.Rahman, "Predicting Early Crop Production by Analysing Prior Environment Factors," *Advances in Intelligent Systems and Computing*, vol. 538, 2017.
17. N. Gandhi, O. Petkar, L. J.Armstrong, and A. Kumar Tripathy, "Rice Crop Yield Prediction in India using Support Vector Machines," *13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2016.
18. Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber, "Accurate Prediction of Sugarcane Yield Using a Random Forest Algorithm," *Agronomy for Sustainable Development*, vol. 36, no. 2, Jun. 2016.
19. A. T. M. Shakil Ahamed *et al.*, "Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh," *16th International Conference on Software*

Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Jun. 2015.

20. S. Pavani, S. B. P. Augusta, "Heuristic Prediction of Crop Yield using Machine Learning Technique," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1S3, pp. 135–138, Dec. 2019.
21. S. Sahu, M. Chawla, and N. Khare, "An Efficient Analysis Of Crop Yield Prediction Using Hadoop Framework Based On Random Forest Approach," *International Conference on Computing, Communication and Automation (ICCCA2017)* , 2017.
22. M. Tahmid Shakoor, K. Rahman, S. Nasrin Ratya, and A. Chakrabarty, "Agricultural Production Output Prediction Using Supervised Machine Learning Techniques," *2017 1st International Conference on Next Generation Computing Applications (NextComp)*, Jul. 2017.
23. T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," *Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018*, pp. 8–13, Jan. 2019.
24. S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, no. 6, pp. 1148–1153, 2016.
25. Y. Y. Song and Y. Lu, "Decision Tree Methods: Applications for Classification and Prediction," *Shanghai Archives of Psychiatry*, vol. 27, no. 2, pp. 130–135, Apr. 2015.
26. L. Breiman, *Random Forests*, vol. 45. 2001.

27. J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," in *Proceedings - 2nd World Congress on Computing and Communication Technologies, WCCCT 2017*, Oct. 2017.

28. Kavitha. S, Varuna. S, and Ramya. R, "A Comparative Analysis on Linear Regression and Support Vector Regression," *Online International Conference on Green Engineering and Technologies (IC-GET*, 2016.

29. L. Qin *et al.*, "Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index," *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, Apr. 2020.

30. S. Liu and E. Dobriban, "Ridge Regression: Structure, Cross-Validation, and Sketching," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.02373

31. N. Gauraha, "Introduction to the LASSO: A Convex Optimization Approach for High-dimensional Problems," *Resonance*, Apr. 2018.

32. I. N. da Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni, and S. F. dos Reis Alves, "Artificial Neural Network Architectures and Training Processes," *Artificial Neural Networks*, Springer International Publishing, 2017, pp. 21–28.

33. H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer Perceptron: Architecture Optimization and Training," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, p. 26, 2016.

34. S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. Springer International Publishing Switzerland, 2015.

35. A. Botchkarev, “Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio.” [Online]. Available: https://ssrn.com/abstract=3177507 [2018].

36. G. S. Handelman *et al.*, “Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods,” *American Journal of Roentgenology*, vol. 212, no. 1. American Roentgen Ray Society, pp. 38–43, Jan. 01, 2019.

37. I. el Naqa and M. J. Murphy, “What Is Machine Learning?” *Machine Learning in Radiation Oncology*, Springer International Publishing, pp. 3–11, 2015.

38. C. J. Willmott and K. Matsuura, “Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance,” *CLIMATE RESEARCH*, vol. 30, pp. 79–82, Dec. 2005.

39. Y. Zhao, A. B. Potgieter, M. Zhang, B. Wu, and G. L. Hammer, “Predicting Wheat Yield at the Field Scale by Combining High-resolution Sentinel-2 Satellite Imagery and Crop Modelling,” *Remote Sensing*, vol. 12, no. 6, Mar. 2020.

APPENDICES

Appendix A: Initial Results

Results in Melaka before hyperparameter optimisation

Appendix B: Hyperparameter Optimisation

```
import pandas as pd

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestRegressor

from sklearn.model_selection import  GridSearchCV



crop_data =
pd.read_csv(r'C:\Users\USER\Documents\CROPS\crop1melaka.csv')


X = crop_data.iloc[:, :-1].values

y = crop_data.iloc[:, [-1]].values

train_X,  test_X,  train_y,  test_y  =  train_test_split(X,  y,
test_size=0.2, random_state=1,shuffle=False)

stand = StandardScaler()

X= stand.fit_transform(X)


y = stand.fit_transform(y)


gsc = GridSearchCV(

    estimator=RandomForestRegressor(),

    param_grid={

        'bootstrap': [True, False],
```

```
        'max_depth': [2,3,4,5,6,10,20,30,40],
        'max_features': [1,2,3,4,5],
        'min_samples_leaf': [2, 3, 4,5,6,7,8,9],
        'min_samples_split': [2,4,5,7,9],
        'n_estimators': range(1,10),
    },
    cv=5, scoring='neg_mean_squared_error', verbose=0, n_jobs=-1)

grid_result = gsc.fit(train_X, train_y.ravel())
best_params = grid_result.best_params_
print('Best parameters found:\n', best_params)
```