PREDICTION OF OIL PALM YIELD FOR SMALLHOLDERS ESTATES IN TROPICAL REGION USING EXTRA TREES METHOD

NUZHAT KHAN

UNIVERSITI SAINS MALAYSIA

2023

PREDICTION OF OIL PALM YIELD FOR SMALLHOLDERS ESTATES IN TROPICAL REGION USING EXTRA TREES METHOD

by

NUZHAT KHAN

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

September 2023

ACKNOWLEDGEMENT

Alhamdulillah, I am greatly indebted to ALLAH (SWT) for His mercy and blessing for making this research work a success. I would like to express my heartfelt gratitude to my supervisor Dr. Mohammed Anuar Kamaruddin for his guidance, unwavering support, and motivation. His expertise and dedication have been invaluable in shaping my research and helping me achieve my academic goals. I am also very thankful to my co-supervisor Dr. Usman Ullah Sheikh for his excessive guidance, assistance, and constructive criticisms during this research. Without his continued support and interest, this thesis would not have been the same. Lastly, I acknowledge and appreciate the contribution of the Ministry of higher education Malaysia, long-term research grant scheme LRGS (Grant 203.PTEKIND.6777007) and the School of Industrial Technology Universiti Sains Malaysia for providing financial support and resources in completing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBRIVIATIONS

AdaBoost	Adaptive boosting
Ag-TECH	Agricultural technology
AI	Artificial intelligence
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural network
Auto-ML	Automated machine learning
BNN	Bayesian neural network
BR	Boosting regressor
CEC	Cation exchange capacity
CNN	Convolutional neural network
DNN	Deep neural network
DT	Decision tree
EDA	Exploratory data analysis
FFB	Fresh fruit bunches
GBR	Gradient Boosting regressor
GHG	Greenhouse gases
GIGO	Garbage in garbage out
IDE	Integrated Development Environment
IoT	Internet of things
IR4.0	4^{th} Industrial Revolution
K-NN	K nearest neighbor
KPI	Key performance indicator
LASSO	Least Absolute Shrinkage and Selection Operator
LOCF	Last observation carried forward
LR	Linear regression
MAE	Mean absolute error
MAPE	Mean absolute percentage error

MET	Malaysian meteorological department
MLR	Multiple linear regression
MPOB	Malaysian palm oil board
MSE	Mean square error
MTSD	Multivariate time series data
NDVI	Normalized difference vegetation index
NOCB	Next observation carried backward
OECD	Organization for Economic Cooperation and Development
PCA	Principal component analysis
R^2	R squared error
RF	Random forest
RFE	Recurrent features elimination
RMSE	Root mean square error
RMSLE	Root mean square logarithmic error
RT	Regression tree
SDG	Sustainable development goals
SVM	Support vector machine
VI	Vegetation indices
XGB	Gradient boosting tree

LIST OF SYMBOLS

Y	The actual value of oil palm FFB yield
$\hat{Y}$	The predicted value of oil palm yield
°C	Unit (used in temperature related parameters)
mm	Unit millimeter (used in rainfall, precipitation)
m/s	Unit meter per second (used in wind speed)
K	Kernel function of SVM
W	Weighted vectors
kWh	Unit Kilo watt hour (used in solar irradiance)
kPa	Unit Kilopascal (used in surface pressure)
σ	Standard deviation

LIST OF APPENDICES

RAMALAN HASIL KELAPA SAWIT UNTUK LADANG PEKEBUN KECIL DI RANTAU TROPIKA MENGGUNAKAN KAEDAH EXTRA TREES

ABSTRAK

Keselamatan makanan global dan penggunaan mampan sumber semula jadi sangat bergantung pada ketepatan masa dalam ramalan hasil tanaman. Kelapa sawit, sebagai tanaman paling menguntungkan untuk pengeluaran minyak di seluruh dunia, memerlukan ketepatan ramalan hasil untuk mengekalkan keseimbangan antara permintaan global dan penawarannya. Tesis ini mencadangkan pembelajaran mesin regresi untuk meramalkan hasil tandan segar kelapa sawit. Objektif utama penyelidikan ini adalah untuk mencipta model pembelajaran mesin yang dilatih dengan data sebenar untuk ramalan jangka panjang hasil kelapa sawit dengan ketepatan yang tinggi. Kajian tersebut menggunakan data yang diperoleh daripada beberapa sumber termasuk Lembaga Minyak Sawit Malaysia (MPOB), Jabatan Meteorologi Malaysia (MET) dan NASA. Cadangan metodologi dilaksanakan menggunakan data khusus yang direkodkan dari seluruh negeri Pahang Malaysia. Data ini terdiri daripada 18 pemboleh ubah termasuk hasil terdahulu, tanah dan pembolehubah cuaca, untuk meramalkan hasil masa hadapan dengan tepat. Analisis statistik memudahkan untuk menilai kualiti data dan mengekstrak maklumat pertanian. Hasil analisis korelasi mendedahkan kesalingbergantungan kompleks faktor-faktor yang mempengaruhi hasil. Penerokaan data telah diikuti dengan saluran paip prapemprosesan untuk menukar data mentah kepada maklumat yang bermakna. Saluran paip prapemprosesan data termasuk mengubah suai outlier, normalisasi, pemilihan ciri dan pemisahan data kepada set latihan, ujian dan pengesahan. Berdasarkan data yang disediakan, pemilihan model automatik digunakan untuk

mengenal pasti model ramalan yang paling sesuai. Model terpilih kemudiannya dilatih, dan penilaian prestasi komprehensif telah dijalankan menggunakan pelbagai parameter termasuk pekali penentuan (R^2) sebagai penunjuk prestasi utama. Keputusan menunjukkan bahawa suhu yang melampau, kelajuan angin yang tinggi, air yang terhad, dan cuaca sejuk memberi kesan negatif kepada hasil kelapa sawit manakala hujan optimum, kelembapan tanah dan sinaran matahari adalah penyumbang positif. Daripada model yang dipertimbangkan, model Extra Trees menunjukkan prestasi terbaik dengan mencapai R^2 sebanyak 1.00 (100%) untuk data latihan, 0.91 (91%) untuk data ujian dan 0.88 (88%) untuk data pengesahan. Ciri yang paling penting untuk model Extra Trees ialah suhu minimum, kelembapan tanah di kawasan akar, dan sinaran suria. Model ini mengatasi model Pokok Keputusan dan model bukan berasaskan Pokok Keputusan yang lain termasuk Peningkatan Kecerunan, Hutan Rawak, AdaBoost dan Mesin Vektor Sokongan. Keputusan menunjukkan bahawa model berasaskan Pokok Keputusan memberikan ketepatan yang lebih baik dari segi ketepatan ramalan dan kuasa generalisasi untuk data agrometeorologi yang dipertimbangkan. Hasil kajian memajukan penyelidikan semasa dengan signifikan melalui ramalan hasil kelapa sawit dengan tepat menggunakan teknik pembelajaran mesin dengan mengambil kira senario dunia sebenar. Dapat disimpulkan bahawa menggunakan potensi kaedah keadaan-seni pembelajaran mesin untuk meramalkan hasil kelapa sawit boleh membantu peladang, penggubal dasar dan pihak berkepentingan lain dalam industri minyak sawit untuk membuat keputusan termaklum dan mengoptimumkan proses pengeluaran.

PREDICTION OF OIL PALM YIELD FOR SMALLHOLDERS ESTATES IN TROPICAL REGION USING EXTRA TREES METHOD

ABSTRACT

Global food security and sustainable use of natural resources heavily relies on the timely prediction of crop yields. The oil palm, being the most profitable crop for oil production worldwide, requires accurate yield predictions to maintain a balance between its global demand and supply. This thesis proposes machine learning regression approach to predict oil palm fresh fruit bunches yield. The main objective of this research is to develop a machine learning model trained on actual data to predict long-term oil palm yield with high precision. The study utilizes data obtained from multiple sources including Malaysia Palm Oil Board (MPOB), Meteorological Department Malaysia (MET) and NASA. The proposed methodology is implemented on site specific data recorded from an entire state Pahang Malaysia. The data is comprised of 18 variables including historical yield, soil, and weather variables, to accurately predict future yield. The statistical analysis facilitated to assess data quality and to extract the agricultural information. The outcomes of the correlation analysis reveal the complex interdependencies of yield influencing factors. The data exploration is followed by a preprocessing pipeline to convert raw data into meaningful information. The data preprocessing pipeline includes treating outliers, normalization, features selection and data splitting into training, testing, and validation sets. Based on the prepared data, the automated model selection is used to identify the most appropriate prediction model. The selected model is trained, and a comprehensive performance evaluation is carried out using multiple parameters including coefficient of determination (R^2) as key performance indicator. The results

show that temperature extremes, high wind speed, water limitations, and cold temperature negatively affect oil palm yield while optimum rainfall, soil moisture and sunlight are the positive contributors. From the considered models, the Extra Trees model performs best by achieving R^2 of 1.00 (100%) for the training data, 0.91 (91%) for the testing data, and 0.88 (88%) for the validation data. The most important features for the Extra Trees model are minimum temperature, root zone soil moisture, and solar irradiance. This model outperforms other tree and non-tree-based models which include Gradient Boosting, Random Forest, AdaBoost and Support Vector Machine. Results show that tree-based models provide better results in terms of prediction accuracy and generalization power for the considered agrometeorological data. The outcomes of the study significantly advance current research by accurately predicting oil palm yield using machine learning techniques considering real world scenario. It can be concluded that using potential of the state-of-the-art machine learning methods for predicting oil palm yield can help farmers, policymakers, and other stakeholders in the palm oil industry to make informed decisions and optimize production processes.

CHAPTER 1

INTRODUCTION

1.1 Research Background

Industry 4.0 (IR4.0) in agriculture refers to the integration of advanced technologies such as the Internet of Things (IoT), big data, artificial intelligence, and robotics into the agricultural industry [14]. These technologies allow farmers to optimize their operations by collecting and analyzing data on everything from weather patterns and soil conditions to crop yields and pest populations. Examples of IR4.0 applications in agriculture include; smart sensors and drones that can be used to monitor crop health, soil moisture, and weather conditions in real-time, precision agriculture, which uses global positioning system (GPS) and mapping technology to accurately apply fertilizers, pesticides, and other inputs to specific areas of a field [15], automatic irrigation systems that use real-time data to optimize water usage and, autonomous vehicles and robots that can perform tasks such as planting, harvesting, and soil preparation. AI-powered predictive analytics can be used to predict disease outbreaks, optimize planting and harvesting schedules and most importantly, forecast crop yields. IR4.0 agriculture can help increase crop yields, reduce costs, improve food security, and decrease the environmental impact of farming. The state-of-the-art technologies are expected to help farmers to make more data-driven decisions, which can lead to more efficient and sustainable farming practices [16]. Crop yield prediction is a critical yet interesting issue due to its significance for long-term yield intensification and optimum land use [17]. Many stakeholders in the agri-food chain, including agronomists, farmers, product exporters, and policymakers, benefit from crop yield forecasts [16]. Various crop-specific characteristics, environmental conditions, and management practices influencing crop production are some of the

confounders in developing a prediction model [18]. Recent research highlighted the need for environment-based crop yield forecasting as one of the ways to minimize the negative effects on crops due to climate variability and extreme weather conditions [19]. At the same time, yield forecasting is emphasized as an adaptation technology to climate change for global food security [11]. The conventional approaches to anticipate crop yield include; field surveys to observe the ground truth with human expertise [20], crop growth models governed by the environment, management strategies, and agronomic principles [21], remote sensing techniques that capture the current status of crops to estimate the final yield [22], statistical models mapping environment and yield in linear statistical equations [23], as well as the combinations of these approaches [24]. However, one of the main limitations of the aforementioned methods is regarding their incompetence to capture fluctuating abiotic environmental factors [25].

Recent advancements in precision agriculture have introduced machine learning models for crop yield prediction [26]. Machine learning combines the strengths of the previous methods, such as remote sensing and growth simulation models with data-driven modeling to produce reliable forecasts [27]. Machine learning algorithms use outputs of conventional methods as features and try to approximate a function that connects predictors/features (environmental factors) to the target (crop yield) [28]. Numerous machine learning and deep learning models have been proposed for environment-based yield prediction of various crops [29]. However, machine learning is underutilized for predictive analysis in oil palm agriculture industry [30].

Despite all technology gaps, the oil palm industry is growing rapidly to fulfill the increasing global demand. Conversely, this crop is affecting tropical forests,

biodiversity, and associated ecosystems negatively [31, 32]. One of the major challenges related to oil palm crop is its unimpeded expansion which has violated a perceived moral obligation of sustainability [33]. Therefore, the oil palm sector is under increasing environmental, economic, and political pressures for endangering the ecological future [34]. The long-term viability and resiliency of the oil palm industry is determined by the capability of estate managers to make strategic decision and procedural changes [35] . In this regard, the most suitable solution rather than opening new lands, is acclimating the latest technology to elevate the yield by reducing the gap between actual yield and potential yield [36-38]. However, some factors, including fluctuating weather, may influence the outcomes significantly [39]. Therefore, data-intensive frameworks created in the context of the agro-environmental domain for oil palm yield forecasting are required.

Consequently, evidence-based decision-making can be achieved by associating machine learning with real data. So far, limited research has been conducted for oil palm yield prediction because it is a complex process that involves many factors such as genetics, soil conditions, weather, and management practices. Additionally, oil palm is a relatively new crop in terms of widespread cultivation, so there may not have been as much interest or funding for research in this area compared to more established crops. However, as the demand for palm oil continues to grow, more research is likely to be conducted in the future to improve yield prediction and overall production efficiency. To this point, some degree of research conducted for oil palm yield prediction includes small scale oil palm yield predicted using a Bayesian network and artificial neural network (ANN) [40]. Similarly, yield trends are statistically investigated by involving climate change to predict country-level oil palm yield [41]. A recent study predicted oil palm yield on block level using normalized difference

vegetation indices (NDVI) from satellite images. The proposed design is sensitive to canopy density, focused on soil and only suitable for small area [42]. Besides, it is not useful in case of mixed canopy because of being unable to distinguish among oil palm trees and other trees [43]. While existing statistical models uncovered linear patterns but failed to interpret nonlinear dependencies in the data [44]. Data greedy ANN, on the other hand, is unexplainable and unaccountable owing to the "black box" effect [45]. A thorough investigation of existing studies indicates some downsides such as: (1) data scarcity, (2) absence of machine learning application for predictive modeling of oil palm yield on state level and (3) lack of a generic system to persuade reusability.

To deal with the shortcomings of existing methods, a robust machine learning methodology is presented in this research. The methodology is proposed to develop a spatially transitional yield forecasting model according to the meteorological variability of the site. This research complies with the need for a modular prediction workflow that can be used to: (1) better understand the convenience of multivariate time series data, (2) improve data quality through a set of pre-processing techniques, (3) select significant feature subset, (4) select appropriate machine learning model by comparing several suitable models automatically, and (5) predict oil palm yield using historical observations. Since yield is a quantity, regression approach particularly designed to predict continuous outcomes such as weight, price, temperature, and numbers is considered. This work combines conventional machine learning regression approach with automated machine learning (auto-ML) to establish a precise yet flexible prediction method for oil palm fresh fruit bunch (FFB) yield.

It is widely believed that a reliable machine learning prediction model can be achieved using comprehensive data pre-processing and precise training. This is the

novelty of the study in which the detailed exploratory data analysis is envisaged to enhance the machine learning model's training process by defining the pre-processing steps for data preparation. Other than the models' prediction power, this study will contribute to the comprehensive understanding on the learning process of machine learning models and the impacts of different environmental parameters on models' performance. Besides, yield responses to its influencing factors will also be investigated with the help of bivariate data analysis during data mining. The proposed technique has not been reported for crop yield prediction before. The results from this study will contribute towards a better understanding on the relationships between oil palm yield variations and environmental factors.

1.2 Problem Statement

Despite many studies concerning machine learning, limited research has been done on its application to predict oil palm yield [46].

i. In particular, there is a notable lack of studies that have assessed the effectiveness of machine learning algorithms for predicting oil palm yield [42]. This is rather surprising, as the regression is very popular for yield prediction of those crops that need early forecasts to ensure global food security [47-49]. Thus, it is of great interest to learn if regression is a more viable substitute to the conventional methods of oil palm yield prediction under varying weather and soil moisture conditions. Furthermore, it is unclear what would be the performance of machine learning regression models in oil palm yield prediction, and under what data pre-processing pipeline the models would perform optimally. To investigate the application of machine learning

regression algorithms, agricultural data analysis should be performed from statistical perspective [50, 51].

ii. To present a comprehensive analysis, it is required to consider both agricultural and machine learning aspects simultaneously [52]. It is difficult to apply machine learning in realistic agrometeorological situations because the regression must cope with erratic nature of the oil palm yield and the changing environment (soil, weather) [53, 54]. In addition, the performance of regressors involves interdependent yield influencing factors and subsequently the training process in the dynamics of the yield and environmental conditions [54].

iii. Due to the radically different variable types involved, plus the highly non-linear nature and non-convexities of the real-world agrometeorological data, linear methods have never been a preferable choice to solve the yield prediction problem [55]. It has been concluded in literature that the exact problem formulation for such prediction models is very complex and rarely leads to a reliable forecast [52]. Furthermore, the conventional approaches such as artificial neural networks require large dataset for training that is not available in case of oil palm yield as it is recorded only once a month [56]. On the other hand, oil palm yield prediction is related to regression and non-linearity of the real-world data; hence the automated machine learning method for model selection is more appropriate because according to the ‘no free lunch theorem’ there is no reliable single model for all real-world complex situations [57].

1.3 Hypothesis

The main hypotheses of this research are as follows:

i. Data from multiple sources can be utilised for oil palm yield prediction.
ii. The small holders can benefit from the investigation to mitigate climate change effects.
iii. It is hypothesized that the machine learning regression approach can be applied to complex agrometeorological time series data for predictive analysis of oil palm yield.
iv. The historical observations of abiotic yield influencing factors if used as input features for model training can enhance learning process of machine learning model by providing real trends and patterns.
v. The correlation among input features and oil palm yield reflects the crop's responses to environmental variations.

1.4 Research Objectives

This research aimed to achieve the following objectives:

i. To pre-process, filter and correlate the data and select suitable machine learning model.
ii. To train machine learning models for oil palm yield prediction.
iii. To evaluate and compare the performance of different prediction models.

1.5 Scope of Research

The scope of the research in this thesis is as follows:

i. The environmental variables used for the model training are limited to historical records of oil palm yield, weather, and soil data of the study site.

ii. Since most of the variables in data contained temporal trends, study was focused on temporal aspects and spatial variations were not considered.

iii. The study mainly focused on performance comparison of tree-based regression models due to their suitability in terms of flexibility, explainability and handling multi-collinearity.

iv. All models and methods applied in the study were exclusively machine learning based rather than deep learning due to data set size limitation.

v. Extreme weather conditions and yield values which are beyond normal range play role of outliers in the dataset.

vi. Trees (number, age, health) in study area remained constant throughout study period.

vii. Experiment is performed for oil palm plantations on the study site Pahang state Malaysia.

viii. Study area was assumed under perfect field management (weeding, pruning, harvesting, irrigation, fertilizers, and pesticides application etc.).

1.6 Research Significance

The renewed interest for crop yield prediction using machine learning methods due to the global food security trends provide an opportunity to rethink the approach to address existing oil palm yield prediction problem. The problem requires intervention from the state-of-the-art as it is likely to remain significant for short and long-term policies in the palm oil producing countries. In this regard, machine learning

incorporating the intelligent and data driven algorithms can contribute for a timely and accurate yield prediction to ensure smooth harvesting, fruit storage, milling operations and import-export agreements etc. The automated machine learning (Auto-ML) has been utilized for the very first time to select the best regression model for oil palm yield prediction according to the available data. The potential application of the supervised machine learning in this thesis provides agricultural researchers an alternative to statistical crop models. The research via a case study of Pahang (a state in Malaysia) can help policy makers to ascertain the strengths of machine learning techniques for better crop management in regions with inadequate or stagnated oil palm production. In addition, it helps the planters to investigate the impacts of different environmental factors on oil palm before plantation. Since the machine learning algorithms can learn complex patterns in actual data and obtain useful hidden information to improve the overall performance in terms of both field management (in field practices) and crop management (fruit handling). Furthermore, this work supports the United Nation's Sustainable Development Goal (SDG) under Goal Number 2 achieving global food security and Goal Number 7 clean energy by substituting fossil fuels with biofuels.

1.7 Research Methodology

To perform the proposed research, following methodology has been carried out:

i. A comprehensive literature review on oil palm is performed focusing mainly on the factors that cause variations in oil palm fresh fruit bunch yield. The supervised machine learning based prediction commonly known as regression is selected as most feasible option to analyze the

effect of environmental factors on oil palm production. The literature review aims to highlight the strengths and limitations of the various aspects regarding oil palm yield patterns, its modelling and integration of agricultural data within context of machine learning for further research. Moreover, data pre-processing steps, machine learning regression algorithms and multiple machine learning models for oil palm yield prediction are examined and compared.

ii. A critical and strategic literature review of machine learning applications in oil palm agricultural industry is performed. The review focuses on observing the purpose, methods, input data, constraints of machine learning and agricultural components in the realistic scenario. Several performance indicators for machine learning models based on statistical evaluation metrics are highlighted. Besides, the objective of the review is to look for a gap in the existing literature.

iii. Application of machine learning regression is proposed to timely predict site specific oil palm yield based on weather and soil conditions. Machine learning models are trained in Python using historical records of oil palm yield, meteorological data, and soil moisture conditions of the study site. The suitable prediction model is selected using Auto-ML based on coefficient of determination (R^2). The ultimate goal of the model is to minimize the prediction error and maximize the accuracy which refers to the model's ability to make correct predictions. The optimal prediction results obtained by model are benchmarked with other machine learning models for validation.

iv. In-depth data analysis is performed in this study to ascertain the resiliency of oil palm crop during varying weather and soil moisture conditions and to ensure the selection of the quantifiable yield influencing factors as input parameters for predictive analysis. For benchmarking, multiple machine learning models trained on same data with identical pre-processing pipeline, have been compared. The performance of considered machine learning models is examined in a 10-fold learning process by six statistical evaluation metrics in addition to learning curve, validation curve, and residual analysis.

v. Lastly, the trained models are analyzed to examine their performance on unseen data. The analysis is carried out by assuming temporal variations of parameters in training and testing set, climatological parameters, and crop responses. A large number of validation strategies and evaluation metrics are included from existing knowledge on regression analysis. Moreover, the impact of each individual feature is observed on the prediction performance of the trained model.

1.8 Research Contribution

This research contributes to the development of a more accurate, flexible, and efficient method for oil palm yield forecasting using machine learning, with potential benefits for small holding farmers and the palm oil industry as a whole. The thesis has made several contributions to the field of agricultural environmental data analysis and machine learning application for prediction of crop yield in general and oil palm yield in particular. The major contributions of this thesis are described as follows:

1) A comprehensive data analysis and pre-processing is performed.

i. This work utilizes actual agricultural and environmental data, which provides realistic, accurate and reliable information unlike previous study using hypothetical weather parameters [40] or not considering environmental factors [58].

ii. In this work, agricultural information is extracted through trends, patterns, and correlations in actual data that was obtained from multiple sources. This allowed for a better understanding of the factors that influence oil palm yield, which is essential for improving crop management practices.

iii. This work performed oil palm yield gap analysis using Spearman's multivariate correlation method to identify main environmental factors causing yield reduction.

iv. A two-fold strategy for variable selection from several yield influencing factors within context of machine learning and afterwards important features are selected. This method helped to select the yield influencing factors reflecting temporal variations as input features by reducing redundancy.

v. The introduction of a comprehensive pre-processing pipeline to obtain high accuracy by ensuring the quality of the input data for the predictive models.

2) The thesis suggests application of machine learning regression for oil palm yield prediction.

i. The integration of machine learning and auto-ML in agricultural application for crop yield forecasting is proposed. This approach provides an automated and efficient way to develop suitable predictive

model. This helps to identify the most appropriate, accurate and reliable model(s) for predicting oil palm yield, which can have important practical implications for crop management and decision making.

ii. The work has identified most suitable configuration of prediction models by hyperparameters tuning.

iii. The design of a generic and reusable workflow provides an outline for future research and practical applications.

3) Performance comparison and comprehensive analysis of oil palm yield prediction models is performed.

i. The performance of multiple models in terms of goodness of fit R^2 along with multiple measures of prediction errors using k-fold cross validation is investigated and compared unlike previous works [40].

ii. The best model obtained 100% and 91% accuracy on training data and test data respectively which is superior to previous works [40, 59].

iii. The performance of trained model is evaluated by getting predictions on unseen data. It provided more realistic picture of the potential performance of proposed model on new data. The model achieved significantly low errors and high R^2 on unseen validation data.

iv. Analysis of learning process of prediction model with help of residuals, and prediction error in addition to the learning curve and validation curve gives insight into multiple aspects of the data and the model.

1.9 Organization of Thesis

This thesis is composed of five chapters, each of which contains specific information regarding the research. Chapter 1 presents background of the research.

This chapter begins with the significance of crop yield prediction in precision agriculture. It also provides a brief information about the conventional yield forecasting methods which are field surveys, remote sensing, yield modeling and combination of abovementioned ones. It also then discuses about the machine learning applications for crop yield prediction in oil palm agricultural industry. Moreover, this chapter includes clearly stated problem statement, research objectives, research significance, research scopes and thesis contributions.

Chapter 2 provides the review of related literature on the research. In addition, the chapter explains about existing methods to predict crop yield, previously adopted techniques in agriculture in general and literature emphasized on oil palm yield prediction. It also provides clear picture of current situation of oil palm industry in study site Pahang state Malaysia. Several available machine learning methods adopted for oil palm yield prediction are discussed that then lead to further explanation on prediction strategies. This chapter also discusses the recent trends on oil palm yield prediction modeling. Additional areas covered in this chapter are the description about input data in form of yield influencing factors, their roles in predictive modeling, data exploration methods, data pre-processing techniques for machine learning application, machine learning types, and tree-based vs non-tree-based regression algorithms.

Chapter 3 illustrates the used materials and adopted methodology for this research. The first part presents the list of variables used followed by description of the research experimental flow chart. The subsequent sections clarify the exploratory data analysis procedure, step by step data pre-processing in a pipeline based on previous data exploration and model training process for oil palm yield prediction, followed by evaluation of the model using six statistical evaluation metrics along with

residual analysis, prediction error analysis, learning curve and validation curve. The last section of this chapter includes the performance comparison of proposed model with existing similar and dissimilar machine learning regression models.

Chapter 4 emphasizes on the discussion of results attained from the research. The first part of the results consists of the data exploration for quality assessment and information extraction, subsequent section contains performance evaluation of the proposed model. This is followed by final part of the chapter that presents the multi-criteria performance comparison of the models. The final part of the results and discussion includes the prediction errors-based comparison of several machine learning models.

Chapter 5 concludes the research in accordance with its proposed objectives. Moreover, future research recommendations related to conducted work are also provided in subsequent section of this chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides fundamental ideas on overall advancements of oil palm, precision agriculture, and machine learning-related concepts including big data, key types of machine learning models, their prediction mechanism, performance evaluation metrics, data pre-processing techniques, and suitable methods adopted to carry out the research. These are followed by existing studies involved in relevant experiments. Due to the multidisciplinary nature of the research, experiments are performed after an extensive literature review from several disciplines. The review is mainly conducted from 1) literature containing works on broader machine learning applications in the oil palm industry. This part of the literature assisted in successfully identifying the research gap in oil palm yield prediction using machine learning methods. 2) Afterward, deep insight into literature containing crop yield prediction supported with materials and methods selection with a special emphasis on machine learning models. 3) Then, a closer look into the literature on existing works specifically related to oil palm yield prediction gave a great insight into current developments in the oil palm yield prediction problem. 4) In addition, specific literature on applications of machine learning for predictive analysis in fields other than agriculture is also referred to for gaining knowledge of existing prediction methods. Detailed pieces of information obtained from the abovementioned literature review are included in this chapter together with identified research gap. In the end, conclusive remarks on reviewed literature are given.

2.2 Precision Agriculture in IR4.0

Over two centuries ago, the majority of the global population was engaged in agriculture, with over 90% of people working in this field [60]. However, today, the picture is vastly different. Specifically, in 38 major developed countries that are part of the Organization for Economic Cooperation and Development (OECD), more than 80% of the population is involved in the service industry. Because of this number of people engaged in agriculture has decreased to just 2-3% [61]. Not only has the population involved in agriculture decreased, but in many developed countries, the average age of individuals in farming households is also increasing. For example, in the Republic of Korea, more than 50% of farm households have a population over 60 years old, and over 40% are over 65 [62]. This shift in population from agriculture to industry and services, particularly manufacturing, means that currently only 5% of the global population is engaged in agriculture, yet it accounts for more than 60% of the world's business [63]. In light of these facts, developed countries are looking for solutions to agricultural problems through mechanization and artificial intelligence (AI). The 4th Industrial Revolution (IR 4.0) presents an opportunity to accelerate the scale and commercialization of agriculture [61]. As a result, the future of agriculture is expected to evolve into high-tech industries where systems are integrated with AI and big data. These systems will converge into a single unit, combining farm machinery, soil seeding, farm management, irrigation and production forecasting. Using the core technologies of the 4IR, such as robotics, big data, and AI, agriculture will enter a new era of "super fusion" [64]. As this era evolves, economic, social, and ethical values will be instilled in various industries and expressed through business models. There are three ways in which the IR4.0 will have a major impact on the agricultural sector. First, precise optimization will solve many current problems in

agriculture [65]. In terms of global food production, enough food is produced for the entire population, yet 30-50% of produced food is discarded, while many people still suffer from starvation [66]. About 80% of the water on the planet is used for agriculture, yet only 20% of viable crops are grown, and the remaining unused surplus is discarded [61]. In the UK, the use of nitrogen has led to the development of blue disease [67]. Each of these problems can be solved through precision agriculture, a method to develop the conventional agriculture with help of modern technology [68, 69].

In today's world, big data is analyzed using AI and machine learning that is useful in the food and agriculture industry for sustaining the supply chain. Use of smart technology can significantly improve the entire process of food production and agriculture [70]. Through various research efforts, state-of-the-art methods have been developed to assist in food processing and agriculture [71]. AI has played a crucial role in ensuring sustainable food production during the IR 4.0 through precision agriculture [72]. The demand for AI in the agricultural technology (Ag-TECH) industry has increased, impacting sustainable food production to feed the future. This has specific implications for real-time monitoring of the farming process, the politics behind sustainable food production, and investment, which is a critical factor in the current situation [73, 74]. Innovations in agriculture can help to increase production while preserving natural resources [75]. As such, promoting agricultural innovation is crucial for the long-term success of the food industry [76]. It is widely acknowledged that ongoing innovation is needed in order to achieve sustainable intensification of agriculture, and that the adoption of new farming practices plays a critical role in this [77]. Researchers have been giving increasing attention to precision farming as a means to achieve this. Precision farming provides a comprehensive system approach

to boost profitability, optimize yield, enhance quality, and lower expenses by managing the spatial and temporal crop and soil variability within a field [78, 79]. With the help of modern tools and smart technologies, such as internet of things (IoT), remote sensing, big data analysis, robotics, optimization, and predictions, precision farming can help farmers to achieve their goals [68, 80, 81].

2.3 The Oil Palm

Oil palm is a valuable crop that has many uses, and it produces two types of oil from its fresh fruit bunches (FFB) [82]. Crude palm oil is extracted from the pulp of the fruit and palm kernel oil is obtained from the seeds inside the fruit [83]. Oil palm has a high per-hectare oil production, with yields 5 times higher than rapeseed, 8 times higher than sunflower, and 10 times higher than soybeans [84]. Due to its profitability and the increasing global demand for palm oil, the area dedicated to oil palm plantations is rapidly expanding [85]. However, this expansion is posing a potential challenge to sustainable land use, biodiversity, and the associated ecosystem [86], as shown in Figure 2.1.

Several negative impacts of expanding oil palm plantations, such as environmental impacts (represented in rectangles) in the form of greenhouse gas (GHG) emissions, climate change, and peatland degradation, and social impacts (presented in circles) such as conflicts have been identified so far [87].

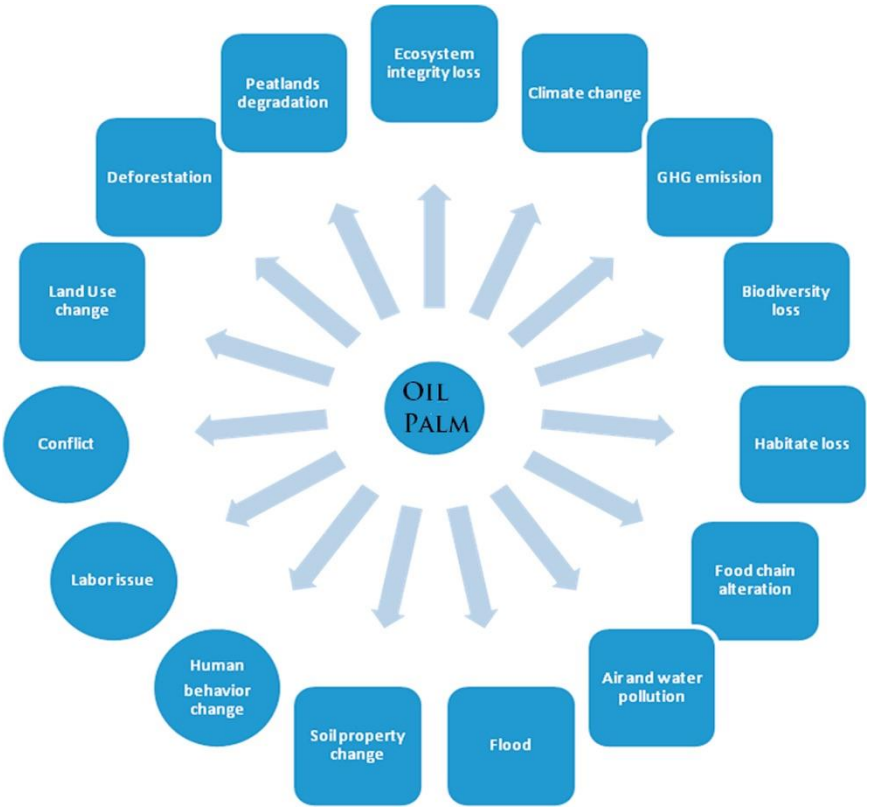

Figure 2.1 Negative impacts of oil palm expansion

2.3.1 Growth Cycle and Life Cycle of Oil Palm Trees

The lifecycle of an oil palm (Elaeis guineensis) is characterized by distinct stages that span several years. It begins with the germination of oil palm seeds, usually taking place in nurseries. When the oil palms mature, they enter the reproductive phase, typically around 4 to 5 years of age, marked by the emergence of their first inflorescences. Pollination occurs through wind and insects, leading to the development of fruit bunches. These bunches take around 5 to 6 months to mature, turning from green to yellow-orange. Harvesting begins when the fruit reaches its optimum oil content as shown in Figure 2.2 sourced [88].

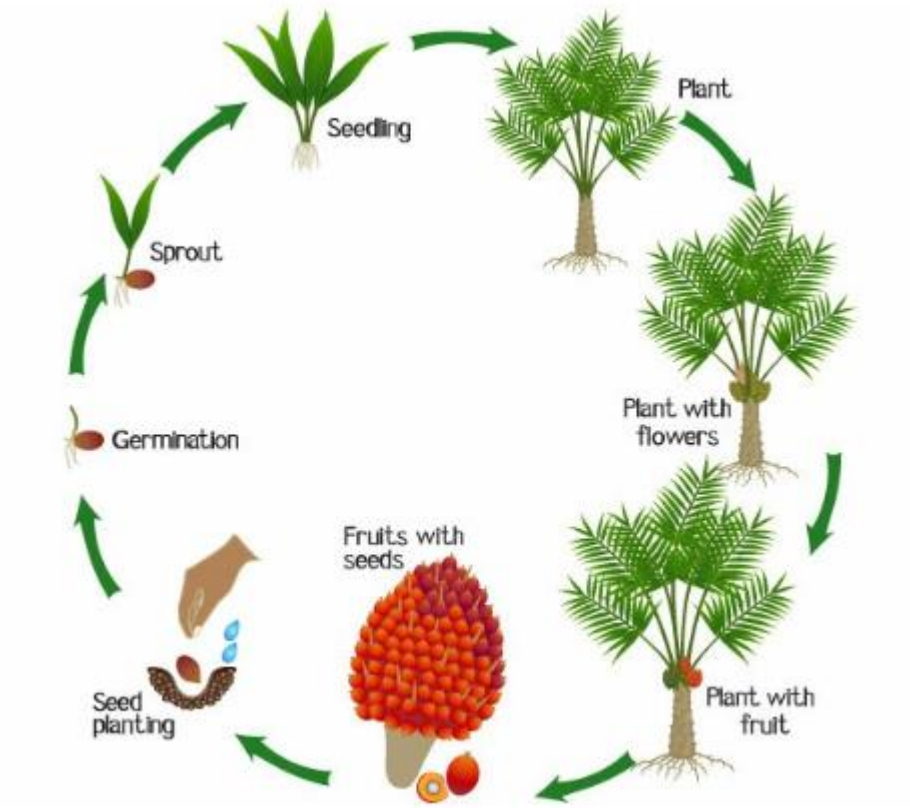

Figure 2.2 Growth cycle of oil palm

Entire lifespan of oil palm trees after seedling phase which is the initial stage where oil palm seeds germinate and develop into seedlings in nurseries is divided into four stages. First stage is the juvenile phase which lasts for about 3 to 4 years, during this phase the tree grows in height and establishes a strong root system [89]. Then, around the age of 4 to 5 years, the oil palm enters the second phase which is immature phase. At this point, it starts producing its first inflorescences and enters the reproductive stage. The first fruit bunches begin to develop, marking the beginning of the tree's fruit-bearing phase [90]. The third phase is mature phase when the oil palm trees reach peak production stage. This typically occurs between 8 to 20 years, depending on various factors. During this phase, the tree consistently produces fruit bunches, and the yield is highest. The mature phase is critical for economic returns in oil palm plantations [91]. However, during fourth phase the productivity of oil palm

gradually declines due to factors such as reduced strength and susceptibility to diseases. Eventually, the tree becomes less economically viable, leading to the need for replanting to maintain the plantation's productivity and sustainability. Meanwhile, after 25 years of planting oil palm trees become too tall to harvest economically [92]. Lifecycle of oil palm with respect to fruit production is presented in Figure 2.3.

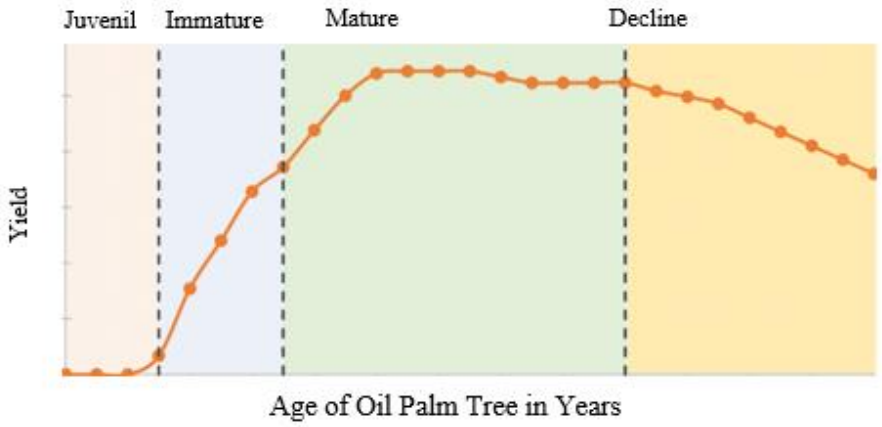

Figure 2.3 Lifecycle of oil palm

2.3.2 Oil Palm Yield

Yield generally refers to the quantity of production obtained from a specific crop [93]. However, in agriculture, there are different concepts of yield based on which it can be divided into four categories: 1) potential yield, 2) water-limited yield, 3) nutrient-limited yield, and 4) actual yield [94]. Potential yield is the maximum yield that can be achieved from a specific seed on a specific site. Water-limited and nutrient-limited yield are the yields that are reduced by water and nutrient limitations. Actual yield is the yield recorded on a specific site. This is the actual in-field yield that is reduced by various biotic and abiotic factors [95, 96].

2.3.3 Yield Gap

The crop yield gap is the difference between the actual yield of a crop and its potential yield [97], which is the maximum yield that can be achieved under ideal conditions. The ideal conditions refer to the absence of unfavourable weather, barren, disease outbreaks and poor field management [98]. Factors such as weather, pests, and soil quality can cause a yield gap by reducing the actual yield of a crop [99, 100]. Identifying and addressing these factors is crucial for increasing crop yields and enhancing food security [101, 102]. Yield gap analysis is a common method to identify the causes of yield reduction[103]. Studies that used data mining and correlation techniques to analyze crop yield gaps [104, 105] have found that this method provides insight into variations in yield gaps from the perspective of independent variables in order to identify factors behind yield reduction [106, 107].

2.3.4 Potential Strategies to Improve Oil Palm Yield

In order to meet the global demand for palm oil, two strategies can be employed: 1) increasing production by planting more oil palm trees or 2) implementing measures to achieve high yield from existing plantations [108] , as shown in Figure 2.4, if the yield demand is 24 tons and two land parcels are producing 8 tons of yield each, instead of using more land (method 1), yield in existing parcels can be enhanced (method 2) with the assistance of modern technology for data-driven and well-calculated decision-making to mitigate the harmful effects of yield-reducing factors [109]. This is a more sustainable approach in terms of land use and resource allocation compared to the former method.

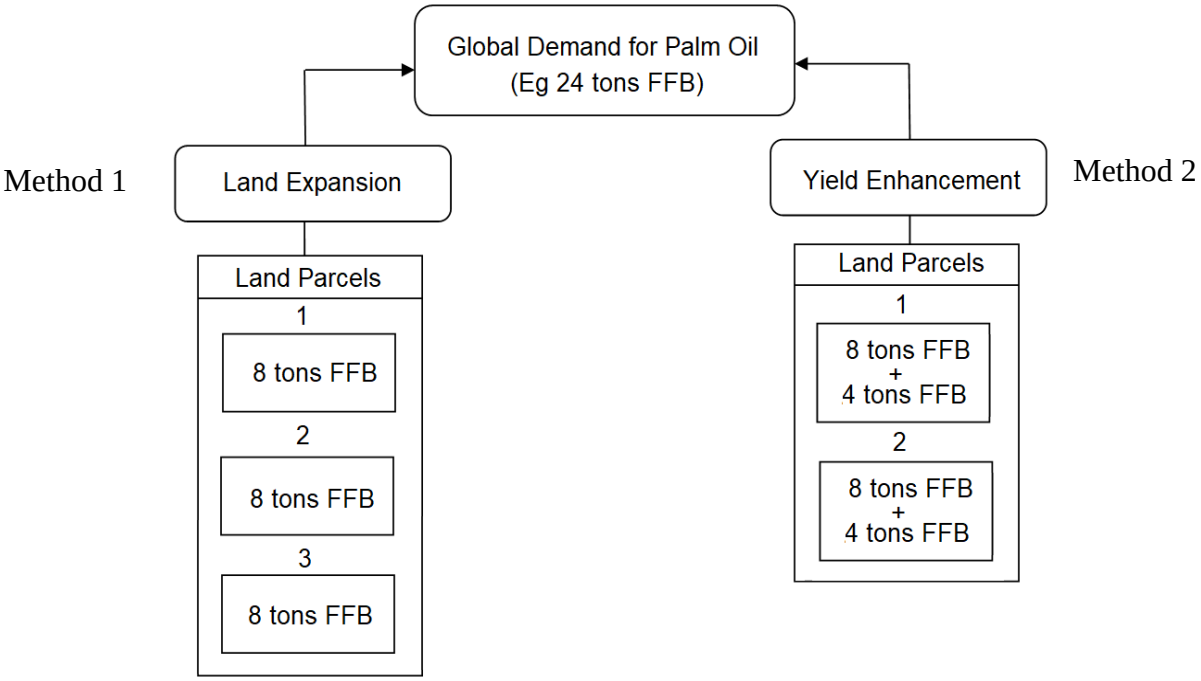

Figure 2.4 Example diagram of two potential strategies to satisfy the increasing demand for palm oil

2.4 Crop Yield Prediction

Timely crop yield prediction has been an integral component of food supply chain management [110]. Several techniques, methods and tools are introduced to make crop yield prediction easy and reliable [111]. Mainly, crop yields are predicted with conventional methods based on human expertise and some mathematical models are also design specifically for this purpose [112]. The key techniques widely applied for crop yield prediction are discussed in subsequent sections.

2.4.1 Yield Prediction with Conventional Methods

Conventional methods for crop yield prediction involve counting the number of trees in a given area and using this information to estimate the potential yield of the crop based on human expertise [69]. This method is considered more appropriate for tree crops, such as fruit trees and nut trees [113]. Moreover, more precise conventional method consider fruit counting for predicting crop yield [114]. This method involves