HYBRID MODEL IN MACHINE LEARNING WITH ROBUST REGRESSION FOR HANDLING MULTICOLLINEARITY OUTLIER IN BIG DATA AND ITS APPLICATION TO AGRICULTURE

MUKHTAR

UNIVERSITI SAINS MALAYSIA

2023

HYBRID MODEL IN MACHINE LEARNING WITH ROBUST REGRESSION FOR HANDLING MULTICOLLINEARITY OUTLIER IN BIG DATA AND ITS APPLICATION TO AGRICULTURE

by

MUKHTAR

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

January 2023

ACKNOWLEDGEMENT

I am grateful to ALLAH Almighty for giving me this opportunity to study abroad at a higher level of Ph.D. and for guiding me on this path to make my journey smoother, and I would also like to express my sincere gratitude to my supervisor, Dr. Majid Khan bin Majahar Ali and Co. Supervisor Dr. Mohd Tahir Ismail for all their help, suggestions, comments and guidance on this full journey to completion of the thesis as without their guidance I will not be able to accomplish this task. I am also grateful to the Universitas Sultan Ageng Tirtayasa and Islamic Development Bank for providing me with financial support for studying abroad and the University Sains Malaysia for providing me with a chance to study. Special thanks to my colleagues and friends for their assisting along the journey. Lastly, this thesis is dedicated to my parents, my sister and members of my family for being always with me and for their patience and support.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

n	Number of sample size
p	Number of Variables
k	Number of variables left in the model
R^2	Coefficient of determination
SSE	Sum of Square Error
MSE	Mean Square Error
MAPE	Mean Average Percentage Error
IoT	Internet of Things

SUATU MODEL HIBRID DALAM PEMBELAJARAN MESIN DENGAN PENGANGGAR TEGUH DALAM DATA RAYA UNTUK MENYELESAIKAN MASALAH MULTIKOLINEARAN DAN APLIKASINYA DALAM PERTANIAN

ABSTRAK

Pertanian kelestarian dan keselamatan makanan memainkan peranan penting dalam ekonomi dunia. Tekanan terhadap pertanian lestari dan keselamatan makanan meningkat kerana pertambahan populasi. Data raya telah diperhatikan menggunakan teknologi Internet of Things (IoT) bagi keselamatan makanan yang mapan. Walau-bagaimanapun, menggunakan terlalu banyak pembolehubah dalam model regresi menimbulkan masalah, terutamanya jika terdapat pembolehubah yang tidak relevan. Algoritma regresi konvensional mempunyai bilangan had kecil dari segi pemilihan pembolehubah, multikolinearan, data terpencil, serta kepersisan model. Dalam kajian ini, sebanyak 29 pembolehubah bebas tunggal dan 435 pembolehubah interaksi bebas telah diketahui. Pembolehubah yang tidak relevan mempunyai implikasi untuk kekangan model yang mempunyai kecendurungan varians yang lebih tinggi. Di samping itu, model sedia ada boleh menjejaskan analisis statistik dengan ketara, yang boleh menyebabkan masalah *overfitting*. Untuk menangani isu- ini, kaedah 8 kriteria pemilihan (8SC) dicadangkan sebagai pemilihan pembolehubah penting. Mengukur pembolehubah penting untuk model pengiraan atau data yang diukur adalah tugas penting yang menyumbang kepada pemodelan. Pembolehubah penting adalah subset pembolehubah yang sesuai daripada pembolehubah asal. Penyelidikan ini mempunyai skop mengenai data besar dan pemilihan pembolehubah yang berkaitan dengan pertanian. Seterusnya, kaedah pembelajaran mesin seperti Ridge, Random Forest,

Support Vector Machines, Bagging, Boosting, Elastic Net dan Lasso menggunakan data raya diimplementasikan untuk menyelesaikan masalah multikolinearan. Batasan kajian ini adalah untuk menangani masalah seperti pembolehubah tidak relevan, multikolineariti dan outlier. Kaedah yang dicadangkan memilih 10, 20, dan 30 pembolehubah penting tertinggi. Data terpencil juga dirujuk sebagai anomali, keabnormalan, dan percanggahan. Data terpencil adalah masalah lazim dalam data raya atau data berdimensi tinggi dan boleh memberi kesan kepada model regresi. Ia boleh membawa kepada model salah spesifikasi, keputusan analisis yang tidak tepat dan menjadikan semua kaedah penilaian tidak bermakna. Beberapa kajian penyelidikan baru-baru ini telah menggunakan pembelajaran mesin dan pembelajaran statistik untuk mengkaji isu regresi untuk mencapai analisis yang berkesan pada pemodelan matematik. Model hibrid telah disokong oleh pembelajaran mesin dan pembelajaran statistik supaya gabungan pintar pelbagai kaedah boleh menjana output yang cekap dan boleh dipercayai. Dalam tesis ini, tujuh kaedah regresi berbeza digunakan - *Ridge*, *Random Forest*, Mesin Vektor Sokongan, *Bagging*, *Boosting*, Jaring Elastik, dan *Lasso* untuk mengkaji ketepatan ramalan menggunakan data pengeringan rumpai laut. Selepas itu, ia akan dianalisis melalui anggaran-M: Tukey Bi-Square, Hampel, dan Huber untuk menyelesaikan masalah data terpencil. Semua ukuran prestasi (MSE, MAPE, R-squared, dan SSE) menunjukkan bahawa keputusan yang lebih baik secara ketara diperolehi oleh hibrid Bagging-Hampel berbanding dengan yang lain. Ia boleh diandaikan bahawa kaedah Bagging – Hampel boleh bergantung pada penyiasatan ketepatan dalam data besar atau data berdimensi tinggi yang diperoleh daripada output pembelajaran mesin.

HYBRID MODEL IN MACHINE LEARNING WITH ROBUST REGRESSION FOR HANDLING MULTICOLLINEARITY OUTLIER IN BIG DATA AND ITS APPLICATION TO AGRICULTURE

ABSTRACT

Sustainable agriculture and food security are critical to the global economy. Pressures on sustainable agriculture and food security are increasing due to population growth. Big data has been used Internet of Things (IoT) technology for observing sustainable agriculture and food security. However, using an excessive number of variables in a regression model might be problematic, especially if there are irrelevant variables. Standard regression algorithms have a number of limitation in terms of variable selection, multicollinearity, outlier, as well as model accuracy. In this research, 29 independent single variables and 435 independent interaction variables were identified. Irrelevant variables have implications for model constraints with a higher variance biases. Furthermore, the current models might substantially impact on statistical analysis, potentially leading to an overfitting issue. To address this problem, the 8 Selection Criteria (8SC) is implemented as the selection of important variables. Measuring relevant variables for a computational model or collected data is an important step in the modelling process. Important variables are a subset of the appropriate variables from the original variables. This research has a scope on big data and variable selection related to agriculture. Hence, machine learning as variable selections such as Ridge, Random Forest, Support Vector Machines, Bagging, Boosting, Elastic Net and Lasso using big data are implemented to solve multicollinearity issue. The limitation of this research were to address the problems

such as irrelevant variables, multicollinearity and outliers. The proposed method selects the 10, 20, and 30 highest important singificant variables. The outliers are also referred to as anomalies, abnormalities, and discrepancies. The outliers are a common problem in big data or high-dimensional data and it can affect regression models. Outliers might lead to incorrectly stated models, inaccurate analytical results, and render all evaluation methods meaningless. Several current research studies have used machine learning and statistical learning to evaluate regression issues to achieve effective analysis on mathematical modelling. The hybrid model has been supported by machine learning and statistical learning so that an intelligent combination of numerous methods can generate efficient and reliable output. In this study, seven different regression algorithms were applied such as Ridge, Random Forest, Support Vector Machine, Bagging, Boosting, Elastic Net, and Lasso to evaluate the accuracy of prediction using seaweed drying. Thereafter, it will be investigated through M-estimates such as Tukey Bi-Square, Hampel, and Huber that is hybrid model of Machine Learning and Robust Regressions. The performance measures, MSE, MAPE, R-squared, and SSE, showed that significantly better results were obtained by Bagging-Hampel compared to the others. It can be assumed that the Bagging-Hampel method can rely on accuracy investigations in big data or high-dimensional data obtained from machine learning and robust regression outputs

CHAPTER 1

INTRODUCTION

1.1 Introduction

Machine learning is an essential topic in high technology for now. Machine learning concerns the application of data science tools between patterns and relationships which are before unknown. Machine learning is closely associated with big data. Machine learning has two approaches. The first approach tries to produce an overall summary of the dataset to identify and explain the substantial variables, and the second approach tries to know the detection of pattern (Cioffi et al., 2020).

The analysis tasks in Machine learning are divided into several categories: summarization, prediction, regression, and classification. Machine learning is a branch of artificial intelligence based on the idea that computers can learn from data, recognise patterns, and make judgments with little or no human intervention. Machine learning is a scientific discipline concerned with the design and development of algorithms that allows them to develop responses based on empirical data (Boutaba et al., 2020). Machine learning techniques and algorithms have indeed become powerful tools for analysing complex and big data, supporting scientists in numerous scientific and technological breakthroughs (Louppe, 2014). Machine learning requires cross-disciplinary proficiency in several areas such as theory of probability, statistics, data mining, and computer science theory (Schmidt et al., 2020). Machine learning is designed to obtain knowledge from available data (Roh et al., 2020). Machine learning has two groups such as supervised and unsupervised learning. Supervised learning is also divided into two sub-groups such as classification and regression. The dependent variable in the classification is discrete and continuous in regression. Among other

definitions, Machine learning applies in more and more scientific fields including: agriculture (Sharma et al., 2020), aquaculture (Cortés et al., 2017; Zhou et al., 2018), and food security (Fragni et al., 2018; Maione & Barbosa, 2019).

Agriculture plays a vital role in sustaining all human activities. Significant challenges such as overpopulation competition for resources pose a threat to the planet's food security. To tackle the ever-increasing complex problems in agriculture offers essential tools to overcome agricultural sustainability challenges (Naik & Suresh, 2018). Data analytics hold the key to ensuring future food security, food safety, and ecological sustainability. Disruptive information and communication technologies such as machine learning, precision farming, and big data can address several issues such as techniques and crop modelling for agriculture sustainability (Sharma et al., 2020).

The current study presents the application of machine learning for sustainable agriculture. Sustainable agriculture and food systems need to be sufficient and nutritious while minimizing environmental impact and enabling producers to earn a decent living (Eyhorn et al., 2019). Agriculture plays a vital role in the world economy. Pressure on the agriculture system will increase as the population increases. Rapid population growth in developing countries means that more food will be needed to fulfil the demands of rising populations (Scholten et al., 2020).

Precision agriculture emerges as a new scientific field. Agriculture technology is a data-driven strategy to increase and improve agricultural output while minimizing environmental impact (Bochtis et al., 2018).

The data generated in precision agriculture is provided by a sensor that enable a better understanding of the operational environment, which leads to more precise and faster decision-making. Agriculture has adopted big data technologies, which are expected to become more common in the next years. Agriculture uses big data technologies to improve accuracy, decision-making tools, and food security zones. The big data repositories include essential information (knowledge) that can be used to analyse scientific data or to give data for multidisciplinary decisions such as 'artificial intelligence of farming' to increase food security and agricultural productivity (Christensen et al., 2018).

Precision farming is a solution to overcome the farmer's and industry challenges in generating necessary information. Malavade and Akulwar (2016) found that precision farming devices can support farmers to get the relevant information and make an accurate decision. These capabilities are essentially initiated by a large number of datasets which consist of different variables and a dependent variable, including their relationship (Hashem et al., 2015).

Seaweed has become an important aquaculture industry to be strengthened and has been highlighted as one of the high-value commodities under Malaysia National Agro Policy the year 2011 – 2025. Malaysia has the potential to be a major seaweed player in region, provided the country has fully developed and utilized the existing strengths. These include the availability of infrastructure, manpower, product quality, transfer of technology, industry support, and marketing (Ali et. al, 2017). Seaweed is one of important agriculture crop in Malaysia that really concern by the government. Seaweed has been given much more concern because of high nutrition value and its growth in short term that is only 45 days per cycle besides economic potential to eradicate poverty in Semporna, Sabah Malaysia.

1.2 Problem Statement

Although there are machine learning methods that implemented in agriculture but there are some limitations on existing methods especially on seaweed post-harvest technology. Too many variables that include especially in seaweed post-harvest process. Seaweed dried using v - Groove hybrid solar drier (vGHSD) that installs in Semporna, Sabah Malaysia.

The vGHSD solar driers are solar collector, forced fans, drying chamber, flushing system components. The solar are located before entering solar collector, Temperature (℃) in front of down v-Groove (Solar Collector), Temperature (℃) in front of up v-Groove (Solar Collector), Temperature (℃) Solar Collector, Temperature (℃) behind inside chamber, Temperature (℃) Infront of (Inside Chamber), and Temperature (℃) from solar collector to chamber. Since we have a lot of sensors, determination of important parameter are crucial. A very rare study had been contacted using issue $p > n$ (number of parameter > observations).

Nowadays, many disciplines have to deal with big dataset that additionally involve a high number of variables. The outlook for big data and machine learning in agriculture especially seaweed drying is very promising. In recent years agricultural and environment data have been increased in exponential rates by the widely use of automated data collection tools and systems. The yield data from precision farming applications have been one of the recent contributions in this increase. Because of this growth, machine learning techniques will be helpful to discover useful or meaningful information in agricultural big data. Machine learning as one of the most popular among many data mining techniques could be used in agriculture data analysis (Cebeci & Yildiz, 2015).

There are many variables involved in the real-life problem, so it is difficult to choose an efficient model out of all possible models relating to analytical factors. Interaction terms affecting the model also need to be addressed because of its vital role in the actual dataset (Javaid et al., 2020). In the same vein with (Chen et al., 2017) found that many models only emphasize a single term without considering variable interactions.

At the moment, there are three common issues in regression methods especially for big data such as irrelevant variables, multicollinearity, and outlier. This thesis provides a conceptual framework to solve their problems. The objectives of conceptual framework were to simultaneously address the issues using Hybrid Model both Machine Learning and Robust Regression.

The first problem statement focuses on irrelevant variables. The data generated at an exponential rate has resulted in Big Data. This data has many characteristics and consists of variables and contains valuable information. However, using too many variables in regression models becomes a problem, especially if there are irrelevant variables. Irrelevant variables can lead to noise and negatively influence the regression model. Irrelevant variables have implication for model constraint that have higher variance and bias. In addition, the existing model can significantly affect the statistical analysis, which can lead to overfitting.

However, using too many variables in the model algorithms can be problematic, particularly if there are irrelevant variables. This can lead to overfitting, in which noise or irrelevant variables may exert undue influence on the model decisions (Peralta et al., 2015). Additionally, there may be redundancies variables. In addition, Omara et al. (2018) found that in many empirical analyses a crucial problem

is the presence in the set of variables not significantly contributing to explain the analysed phenomenon, but capable to create a random noise which prevents from distinguishing the main effects and relevant predicts.

Variable selection is an important topic in machine learning, especially for big data. Variable selection is a knowledge discovery tool which provides an understanding of the problem through the analysis of the most relevant variables. For example, in regression, it is very useful to select and maintain variables with predictable capabilities.

The aims of variable selection usually are to improve predictive model capabilities and to provide a more comprehensive understanding of the prediction model by reducing ineffective and unnecessary variables.

The second problem statement focuses on multicollinearity. Multicollinearity is a phenomenon that can occur when running a multiple regression model. In this age of big data, multicollinearity can also be present in the big data. The problem of multicollinearity in the dataset can strongly distort ordinary least-squares estimates and lead to unreliable results. The multicollinearity problem is generally recognized in regression model. The presence of serious multicollinearity would reduce the accuracy of the parameter estimate in regression models.

The main problem associated with multicollinearity includes unstable and biased standard errors leading to very unstable for assessing of the predictors, which could result in unrealistic and unstable interpretations.

Machine learning only provides the rank of significant variables (Zien et al., 2017), which means that techniques did not have rules in selecting the suitable range of significant variables (Gómez et al., 2017). The highest significant variables are the ranking of the independent variables that contribute to the dependent variable. The highest significant variables are a suitable of the variable from original variables (Rohart et al., 2017).

Machine learning approaches become an interesting issue in big data era. Machine learning aims to solve the problem of multicollinearity issue (Jaya et al., 2020). Multicollinearity is two or more independent variables with high correlation. Multicollinearity, which exist when two or more explanatory variables in a regression model are highly correlated, is a frequently encountered problem in multiple regression analysis. It is a problem that is often encountered in multiple regression analysis. The high multicollinearity will decrease the precision of parameter evaluate in the models of regression (Tamura et al., 2017).

Machine learning methods are applied as a variables selection to select the most significant features with their perspective. Machine learning techniques provide algorithm to decrease the variance and decrease the sample error to handle multicollinearity problem. Numerous machine learning are used in various field of scientific disciplines. Therefore, machine learning is a regression method to solve multicollinearity.

Numerous machine learnings have been suggested to handle big data problems (Zhou et al., 2017). In this study, machine learning is used for variable selection. Limitation of machine learning cannot provide how many significant and unsignificant variables (Drobnič et al., 2020). The highest significant variables are the ranking of

the independent variables that contribute to the dependent variable (Rohart et al., 2017).

The third problem statement focuses on outlier. The outliers are common in big data and can create severe regression problems. They can lead to model misspecification, inaccurate analysis results and make all evaluation methods meaningless.

The outliers can occur due to various reasons such as human error, mechanical error, and instrument error. Furtheremore, an incorect estimate model can lead to rising errorneous results and significantly influence the mean and standard deviation and lead to either over or underestimated values.

An *outlier* is a data point that is significantly different from surrounding points (Ayadi et al., 2017). Outlier detection identifies data objects which are drastically different from the rest of the data set. Outliers can happen due to different reasons such as human error, fraudulent behaviour, mechanical faults, and instrument error. Regardless of the source outliers, their detection can reveal system faults, fraud, and interesting patterns in the data. The detected outliers can assist the expert in narrowing down the scope of analysis and understanding the root cause(s) of such anomalous behaviours. Machine learning outliers may lead to unreliable or even wrongly identified models. Moreover, incorrect estimation of model parameters may also give rise to erroneous conclusions. A simple example to illustrate the effect of unwanted outliers on the results of data analysis is in statistical analysis, where the presence of outliers in the data can significantly affect the estimation of the mean and standard deviation of a sample data, which can lead to either over-or under-estimated values

(Perez & Tah, 2020). Outlier detection has been studied and applied in different fields to detect anomalous behaviours (Abghari et al., 2018).

Therefore, an important property of robust regression is method with respect to outliers in big data. Robust regressions are required where the estimated values are not much influenced by much smaller or much larger observations. Hence, robust regression is a regression method which is designed to adrress outliers.

1.3 Objectives of the Study

i. To optimize significant variables using subset variable selection of 10, 20, and 30 for the highest significant variables.

ii. To analyse the forecasting and accuracy of the model when interaction effects are included using proposed Standard Operation Prosedure (SOP).

iii. To compare the accuracy of the proposed model with a hybrid of machine learning and robust statistical model.

1.4 Scope and Limitations of the Study

This research has a scope on big data related to agriculture. In modern scientific research, datasets with hundreds or thousands of features are common. Big data sets face significant challenge to machine learning, statistics learning and other fields. In recent years, data pertaining to agriculture have been increased in exponential rates due to using automated data collection methods and systems widely. The harvest information from smart farming application have been one of the recent contributions in this improving. Due to this enormous growth, machine learning algorithms will be useful to find relevant features for agriculture big data. Increasing the number of

features affects the time complexity of algorithm used to analyse the data. However, only a small number of them are considered to be actually significant to the dependent feature. The irrelevant features threaten the model accuracy. Numerous studies have been performed on this problem. Hence, the big data issue must be reduced to a low-dimensional one.

In the big data analysis, variable selection is one of the most essential role and one reasonable solution. Variable selection will eliminate irrelevant variables and improve the regression accuracy. It reduces the original variable to be smaller, aiming to maintain and save significant variables and remove redundant variables. It is important in big data for increasing regression performance, which has been extensively studied and is frequently utilised in a variety of fields.

The limitation of this research were to address the problems based on both multicollinearity and outliers. For forecasting, model selection has its importance in big data era. This thesis focuses on the model selection issue in the agriculture field. In previous research, the regressions are utilized by the different researchers for the efficient forecasting. These is also another kind of regression methods are applied for forecasting. Multicollinearity is considered a big issue in regression analysis. But not every model can deal with this problem (Gujrati, 2004). For handling multicollinearity, this thesis applies the proposed methods such as Ridge, Random Forest, Support Vector Machines, Bagging, Boosting, Lasso, and Elastic Net. These models are variable selection. Similarly, the presence of outliers is considered as big issue in data analysis. Therefore, M robust regression methods, M Bi Square Tukey, M Hampel, and M Huber, are available to handle this problem. Some researchers have already been working on interaction effects (Ali et al., 2017). The SOP models, this study develops an efficient model. The current study focused on developing hybrid models

using machine learning including Ridge, Random Forest, Support Vector Machines, Bagging, Boosting, Lasso, and Elastic Net, and M robust regression. In this work, all possible models are involved. Due to the inclusion of all possible models, only the 10 highest important parameters, the 20 highest important parameters, and the 30 highest important parameters are used in big data analysis. By including the interaction effect, the study now consists of a total of 63 parameters. The developed SOP shows that an efficient model can deal with the problem of multicollinearity and outliers. In the context of big data issues, the developed SOP can also provide efficient forecasting results.

1.5 Significance of the Study

The agriculture sector is regarded as one of Malaysia's most essential economic parts. Agriculture is an industry sector that is benefitting enormously from the development of sensor technology (Internet of Things), data mining, and machine learning techniques in the latest years. The proposed model able to increase the efficiency of the drier which will contribute directly to increase for seaweed farmers or industries.

The research project helps to identify the key factor related to dried seaweed quality and productivity. The factor affecting the removal of the moisture ratio and the factors affecting the collector efficiency are observed. The key significance of the research is the inclusion of the interaction factors in the analysis. It can also highlight the importance of dealing with all possible models. The four phases provided for this project are in a position to obtain an efficient chosen model. The efficient model is now ready to predict factors related to the collector efficiency and the moisture ratio removal of seaweed using the solar drier. The developed SOP that has been developed

can be used in any model. The issue of modelling in big data can be addressed through the development of SOPs by considering the important parameters in field of agriculture.

1.6 Thesis Framework

This thesis is organized as follows:

The thesis is organised as follows. Chapter 1 covers the background to the problem and the research gap. Chapter 2 discusses the types of regression analysis that researchers have already used. The theory regarding the methods applied in this research is discussed in Chapter 3. Chapter 4 makes provision for the application of the proposed big data set procedure. Comparisons are observed with other existing estimators, and results are noted. Chapter 5 comprises the summary and conclusions based on the results of Chapter 4. It also discusses the future directions and limitations of the research project.

CHAPTER 2

LITERATURE REVIEW

Regressions are consistently interesting to discuss. However, there are still unresolved concerns in the regression phase. The issues in regression are irrelevant variables, multicollinearity, and outlier. The Chapter 2 is divided into three problem statements. The first problem statement discusses variable selection in order to deal with irrelevant variables through the use of Machine Learning and Sparse regression. The second problem statement focuses on the issue of multicollinearity. The third problem statement covers robust regression to solve the issue of outlier, and eight selection criteria.

The first problem statement is divided into two parts, such as sections 2.1 and 2.2. Section 2.1 presents some literature concerning the investigation of Machine Learning as variable selection to deal with irrelevant variables. Section 2.2 illustrates literature review sparse regressions as variable selection. These techniques implement a means to control regression coefficients. These algorithms can decrease the variance and reduce our sample error to solve this issue.

The second problem statement focuses on the issue of multicollinearity. The issue of multicollinearity has been stated in Section 2.3. The multicollinearity problem is generally recognized in regression model. The presence of serious multicollinearity would reduce the accuracy of the parameter estimate in regression models. The main problem associated with multicollinearity includes unstable and biased standard errors leading to very unstable for assessing of the predictors, which could result in unrealistic and untenable interpretations.

The third problem statement covers robust regression to solve the issue of outlier. The use of robust regression analysis has been discussed in Section 2.4. The literature concerning outlier issue has been illustrated in section 2.5. Section 2.6 reviews the Eight Selection Criteria (8SC). Finally, In section 2.7 presents summary of the literature review.

2.1 Machine Learning

Machine learning is a subfield of science concerned with the formal research of learning systems. This is a multidisciplinary area that borrows and creates upon concepts from mathematics, computer science, statistics, cognitive science, engineering, and optimisation theory (Ghahramani, 2004). Machine learning algorithms attempt to find pattern in data. In their basic forms, that frequently means identifying a predictive relationship between factors (Bengio; 2012).

Machine learning is applied to train computers or machines to manage the data more effectively (Coelho, 2012). These are several machine learning approaches to build predictive models. Machine learning is a field in data analytic that focuses on mathematical algorithms development to forecast future events (Menagie, 2018).

Computer/systems in machine learning can learn from the past data, and practice. Finding patterns in big data is the purpose of machine learning algorithms, without being specifically programmed. Computers are capable to analyse big data in order to discover patterns and rules in a manner that are too difficult for humans to do. The fundamental concept behind machine learning is that computers or machines can automatically learn from experience. The computer analyses big data in order to uncover patterns and rules hidden in the data.

Machine learning is a subfield of artificial intelligence that aims to create systems that are capable of learning from their previous experience. Machine learning methods are often data-driven, inductive and general in nature. Machine learning approaches are established and implemented to generate predictions or decisions in a broad range type of task for example: handwriting, spam filtering, or activity detection (SAS, 2016). Figure 2.1 below presents machine learning branch.

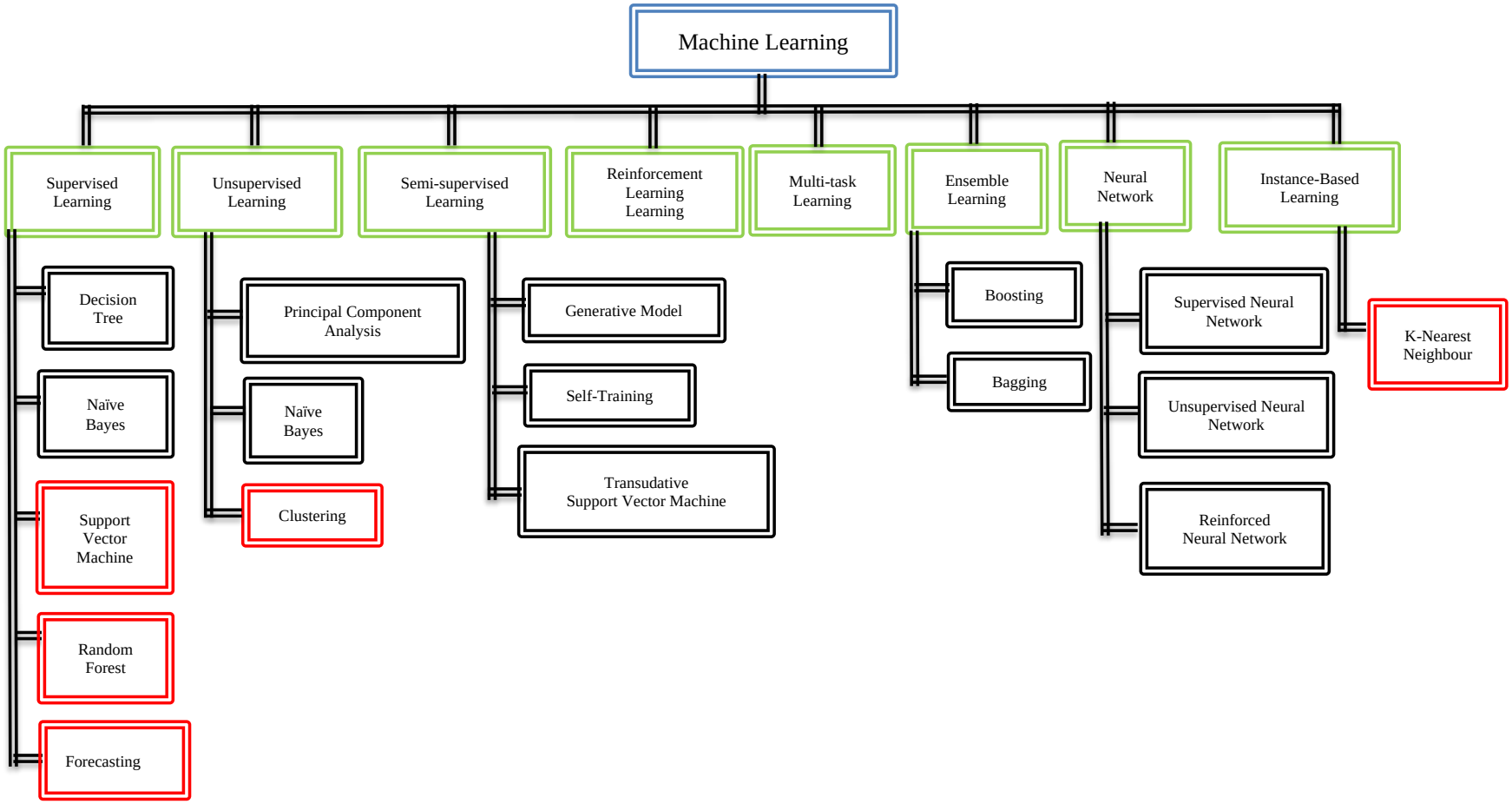

Figure 2.1 Machine Learning

Three major categories in machine learning problem exist: classification, clustering, and regression. The classification involves assignment a set of potential classes to an observation, for example an email message: spam or not. As a result, Classification produces discrete n-array output. Clustering separates a set of observations into groups in order to maximise the similarity of observation within each group and as different as possible from one group to the next, for example: pattern recognition (Giridhar et al., 2019).

Specifically, this is the main issue under consideration in this research. Regression defines as the estimation of the relationship of a response from one or more predictors. Both the response and one or more predictors have continuous value ranges in general.

Table 2.1 Study about Machine Learning

Author(s) (year)	Sample size			Field of study		Simulation		Data dimension		SOP	Remark	Regression	Classification
	Small	Medium	Big	Agriculture	Non agriculture	Monte Carlo	Other	$p<n$	$p > n$				
(Nagpal & Singh, 2018)	√	√			√		√		√		Random Forest		√
(Agusta & Adiwijaya, 2019)			√		√		√		√		Random Forest		√
(Izquierdo & Zurita, 2020)	√	√	√		√		√	√			Random Forest	√	√
(Kulkarni & Sinha, 2013)	√	√	√		√		√	√			Random Forest		√
(Čeh et al., 2018)			√		√		√	√			Random Forest	√	
(Couronné et al., 2018)			√		√		√	√			Random Forest		√
(Schonlau & Zou, 2020)	√				√		√	√			Random Forest	√	
(Han & Kim, 2019)	√				√		√	√			Random Forest	√	√
(Liu et al., 2021)	√				√		√	√			Support Vector Machines	√	
(Adnan et al., 2022)			√		√		√	√			Support Vector Machines	√	

Table 2.1 (Continued)

Author(s) (year)	Sample size			Field of study		Simulation		Data dimension		SOP	Remark	Regression	Classification
	Small	Medium	Big	Agriculture	Non agriculture	Monte Carlo	Other	$p<n$	$p > n$				
(Cao et al., 2019)			√		√		√	√			Support Vector Machines		√
(Deiss et al., 2020)		√			√		√	√			Support Vector Machines	√	√
(Yu et al., 2017)	√				√		√	√			Support Vector Machines	√	
(Kranjčić et al., 2019)											Bagging		
(Taser, 2021)	√				√		√	√			Bagging		√
(Athanasopoulos et al., 2018)	√				√		√	√			Bagging	√	
(Kotsiantis, 2014)											Bagging		
(Alelyani, 2021)		√			√		√	√			Bagging	√	
(Tuysuzoglu & Birant, 2020)	√	√	√		√		√	√	√		Bagging		√
(Lin et al., 2021)	√				√		√	√			Bagging	√	

Table 2.1 (Continued)

Author(s) (year)	Sample size			Field of study		Simulation		Data dimension		SOP	Remark	Regression	Classification
	Small	Medium	Big	Agriculture	Non agriculture	Monte Carlo	Other	$p<n$	$p > n$				
(Yang et al., 2016)	√	√			√		√	√			Boosting	√	√
(Basha et al., 2018)	√				√			√			Boosting	√	
(Momparler et al., 2016)			√		√			√			Boosting		√
(Zheng, 2012)			√		√		√	√			Boosting		√
(Gezy & Poggi, 2008)	√				√			√			Boosting	√	
(Freund et al., 2017)			√		√			√	√		Boosting	√	√
(Bühlmann & Hothorn, 2007)													

Nagpal and Singh (2018) proposed a variable selection method appropriate for high dimensional data. This research presented a proposed Random Forest Statistical Test technique for abbreviated RFST. This approach is designed to solve the issue of high dimensional data. Random Forest is utilized in this proposed approach as RFST considers both the influence of each predictor separately and their multivariate interaction with other predictors. The RFST approach eliminates irrelevant variables using the highest importance features obtained from Random Forest. The RFST then allocates the statistically different variables to different groups. Finally, a representative from each group is selected to eliminate redundant or unnecessary features which findings in the selection of features are smaller than high dimensional data. These are two phases to the feature selection. During the initial phase, The random forest is applied to determine the importance of every feature, as a result, the features with high significant could be given more priority than less significant. In the last phase, features clustering is achieved around significant features to produce the reduced feature set. The stated findings indicated that the proposed method is beneficial to find the smaller set of factors with better accuracy in classification.

According to Agusta and Adiwijaya (2019), the proposed method is MBRF (Modified Balanced Random Forest) that is classification approach to solve imbalanced data issue. It is to modify the procedure of a Balanced Random Forest by implementing a clustering algorithm according to under-sampling for each data bootstrap-decision-tree in the Random forest method. Clustering algorithms used to optimize performance are Agglomerative Clustering, K-mean, Spectral Clustering, and Ward Hierarchical Clustering. The MBRF produced superior results as well as decreased running time in comparison to others.

Liu et al. (2021) proposed hybrid model based on Support Vector Regression (SVR) and Genetic Algorithm (GA) to handle missing data in effect of equipment health diagnostics. First, GA is utilised to improve SVR, and GA-SVR could be presented as a novelty method. The GA-SVR approach is trained through the variable to which the missing data belongs. The single-variable prediction approach can be achieved. In order to reconstruct the training set, the correlation analysis is applied, and the GA-SVR is trained by using the data of factor related to missing data to obtain the multivariate prediction method. Then, the dynamic weigh is shown to combine both the single and multiple-variable prediction technique according to certain principles, the combined prediction algorithms are used to fill in the missing data. The GA-SVM is used to assess equipment failure using the filled data as input. Finally, the findings are provided to validate the effectiveness of the GA-SVR.

Adnan et al. (2022) focused on developing robust streamflow to predict the accuracy model through balancing the capabilities of exploitation and exploration to determine the optimum parameter of the machine learning methods. To do that, the hybrid model is integrated methods between the Simulated Annealing (SA) method and the Mayfly Optimization Algorithm (MOA) as SAMOA to select the best parameters of Support Vector Machines (SVM) to solve the weakness of MOA algorithm. The validations model are R-square, RMSE, and MSE. Test finding indicated the SVM-SAMOA is better than the SVM-MOA and SVM.

Athanasopoulos et al. (2018) explored Bagging, Bootstrap Aggregating, technique as modelling and predicting for tourism demand. The objective of this research is to improve the accuracy of regression forecasting to consider the process of feature selection that are beneficial in industrial applications. The procedures of features selection is GETS (General-to-Specific) method relating to inferential

statistics and stepwise algorithm according to MPA (Measure of Predictive Accuracy). The findings according to the arrivals of tourist from the data of six source markets to Australia recommends that Bagging is beneficial to improve the forecasting precision of the models under consideration.

Alelyani (2021) discussed ensemble learning according to the Bagging algorithm to enhance the stability of variable selection in medical data through the reduction of data variance. This research performed an experiment that employed four microarray datasets suffering from ultra-high dimensionality, and the sample size is a relatively small. On each of the datasets, this research used five variable selection methods to choose a variety of variables. Selection stability has been significantly improved using the proposed technique. The proposed method maintains the accuracy of classification. The improvement in stability is between from 20 to 50 percent.

Tuysuzoglu and Birant (2020) proposed a modified Bagging, eBagging (enhanced Bagging), that applies a new procedures (error-based-bootstrapping). The eBagging builds training sets to solve this issue. The eBagging have been compared with other classification methods such as KNN, SVM, C4.5, and Naïve Bayes. The findings reveal that eBagging is better than others in terms of classification accuracy and training error reduction.

Zheng (2012) proposed QBoost (Quantile Boost) approach that forecasts quantile of the interested dependent variable for binary classification and regression. QBoost regression conducts gradient descent in the space of functional to minimize the function of objective applied by Quantile Features (Qreg). In the classification situation, the class label is determined by a hidden feature, and the quantiles of the class label are determined by fitting the corresponding quantiles of the hidden feature.

Comprehensive experiments and elaborate analysis ascent to achieve the QBoost technique. The findings indicated that QBoost is better than others. Furthermore, QBoost can solve problem in high-dimensional data and is more robust to disturb from features.

2.2 Sparse Regression

Regressions continue to get significant appreciation and attention. However, there is still multicollinearity issue in regressions. Multicollinearity is defined as a high correlation between two or more independent variables. It is a prevalent issue that is often encountered in regression algorithms. The accuracy of parameter assessment in regression algorithms will be reduced by multicollinearity (Tamura et al., 2017).

Sparse regressions are applied as a variable selection to select the most significant variables with their perspective. They provide methods for controlling the regression coefficient, which is able to decrease the variance and decrease the sample error to solve the multicollinearity issue. They are applied in various fields of scientific disciplines (Ahrens et al., 2020). So, sparse regression is a regression analysis designed to handle multicollinearity. In this thesis, we will use three types of regularization regressions such as Lasso, Ridge, and Elastic Net. Table 2.3 below presents the regularization regression Analysis.

Al-Jawarneh and Ismail (2021) proposed EMD – ELN (the Empirical Mode Decomposition – Elastic Net), it is a method for determining the decomposition components of factors that have significant effect on the response (dependent) under multicollinearity. The EMD-ELN technique is quite successful to identify the decomposition components with the existence or absence of multicollinearity in the factors. The EMD - ELN approach improved accuracy in regression modelling.