FORECASTING USING ARTIFICIAL INTELLIGENCE OR MACHINE LEARNING ALGORITHM

BY

MR. KASIDIT SINGTHONG

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING (LOGISTICS AND SUPPLY CHAIN SYSTEMS ENGINEERING)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2020

FORECASTING USING ARTIFICIAL INTELLIGENCE OR MACHINE LEARNING ALGORITHM

BY

MR. KASIDIT SINGTHONG

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING (LOGISTICS AND SUPPLY CHAIN SYSTEMS ENGINEERING)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2020

THAMMASAT UNIVERSITY
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

INDEPENDENT STUDY

BY

MR. KASIDIT SINGTHONG

ENTITLED

FORECASTING USING ARTIFICIAL INTELLIGENCE OR MACHINE LEARNING ALGORITHM

was approved as partial fulfillment of the requirements for the degree of Master of Engineering (Logistics and Supply Chain Systems Engineering)

on July 2, 2021

Member and Advisor

(Associate Professor Jirachai Buddhakulsomsiri, Ph.D.)

Member Warut P.

(Assistant Professor Warut Pannakkong, Ph.D.)

Director

(Professor Pruettha Nanakorn, D.Eng.)

Independent Study Title	FORECASTING USING ARTIFICIAL INTELLIGENCE OR MANCHINE LEARNING ALGORITHM
Author	Mr. Kasidit Singthong
Degree	Master of Engineering (Logistics and Supply Chain Systems Engineering)
Faculty/University	Sirindhorn International Institute of Technology/ Thammasat University
Advisor	Associate Professor Jirachai Buddhakulsomsiri, Ph.D.
Academic Years	2020

ABSTRACT

In canned fruit industry, an important factor for maintaining product quality is a capability of the producers to properly prepare mixtures of other ingredients that match quality characteristics of its raw material (RM), i.e. fresh fruits.

This study involves a statistical model and machine learning models to predict the degree Brix of raw material in a canned pineapple production process at a major manufacturer in Thailand. Prediction models include multiple linear regression, support vector machine, artificial neural network, deep belief network, and ensemble model. Input data are pineapple's color, harvest month, and indicator variables representing geographic areas of the pineapple sources. The objective is to support the laboratory technicians with accurate degree Brix prediction, which can help reducing the packing medium preparation cost and time in order to avoid production delay. After the models are trained, fine-tuned, and tested, the final models can predict the degree Brix within approximately 9% error, which is considered to be satisfactory for this process. The models are then used to generate prediction values for incoming batches of the pineapple for 11 months of production data (one season).

Finally, a decision support system (DSS) containing a database of prediction values and key input variables is developed for a user. DSS is developed to show the best predicted value of °Brix of RM based on its performance. Moreover, a reliability level is applied to the DSS to enhance the ability of the program. The reliability level is classified into four levels. The very high reliability level is resulted from the inputs of the independent variables that exactly matched with the data sets used to construct the prediction models, which is high accuracy. The high reliability level was resulted from the input in which only one attribute does not match with the existing data. The medium reliability level was presented when there is a new combination of attribute in the input data when compare with the existing data (i.e., new combination of district and collector). The low reliability level was resulted when the color or month of input matches only with one attribute of the existing data.

Keywords: Prediction, Artificial Intelligence, Machine Learning, Support vector machine, Artificial Neural Network, Deep belief network, Ensemble method, Decision Support System

ACKNOWLEDGEMENTS

I would like to express the appreciation to my advisor, Assoc. Prof. Dr. Jirachai Buddhakulsomsiri, who teaches, advises, and escorts me to the right direction. Not just only in term of academics, but also life lessons as well. Without his guidance, this study would not be possible. Also, I would like to express my gratitude to Asst. Prof. Dr. Warut Pannakkong, who did not hesitate to give me knowledges and advices.

Secondly, I would like to express my thankfulness to my friends and family who stay by my side from the beginning.

Lastly, I apologize to all other unnamed people who helped me in many ways to complete this study.

Mr. Kasidit Singthong

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

Symbols/Abbreviations	Terms
°Brix	Degree of Brix
RM	Raw material
PM	Packing medium
TSS	Total soluble solids
SSC	Soluble solids content
ML	Machine learning
AI	Artificial intelligence
SVM	Support vector machine
ANN	Artificial neural network
DBN	Deep belief network
FFNN	Feed-forward neural network
DNN	Deep neural network
DSS	Decision support system
DF	Degree of freedom
ANOVA	Analysis of variance
Adj SS	Adjusted sums of squares
Adj MS	Adjusted mean squares
MAE	Mean absolute error
MAPE	Mean absolute percentage error

CHAPTER 1
INTRODUCTION

1.1 Introduction

Thailand is one of the world’s major agricultural exporters, which is primarily driven by food processing industry (USDA, 2019). In Thailand there are more than 10,000 food processing factories operating in 2018, and the food industry has continued to grow (BOI, 2019; USDA, 2019).The main international markets for Thai processed food are Japan, the US, and the European Union (EU). According to Thailand: Food Processing Ingredients from Foreign Agricultural Service (FAS), Thailand exported processed foods with a value of 14,197 million USD in 2019, an increase of 6.78 percent from the previous year. A variety of raw materials for Thailand food processing industry are such as seafoods, meat, and fruits (USDA, 2019). Among all processed fruits, the most important exported product from Thailand is canned pineapple. Based on the statistics record from Food and Agriculture Organization of the United Nations (FAO, 2021), Thailand is the world’s fourth largest pineapple producer behind Costa Rica, Philippines, and Brazil. However, Thailand has been the world’s largest exporter of canned pineapple since 1984 because the domestic consumption is relatively low compared to other major producers.

With today market competition, product quality is important in achieving competitive advantage in the long run (Djekic et al., 2019; Escobar et al., 2018; San-Payo et al., 2019). An important factor for maintaining product quality is the capability of the producers to properly prepare mixtures of other ingredients that match quality characteristics of its raw material (RM), i.e. fresh fruits. For canned fruit industry, important sensory characteristics are such as texture, flavor, color, juiciness, and ripeness. Particularly, in Thailand, pineapple farming is mostly not industrial agriculture, but rather household farming as there are over 48,000 households of pineapple farmers. Majority of the farmers (approximately 83.8%) have relatively small sized farms, ranging from 0.32 to 6.4 Ha, with an average farm size of 1.9 Ha. Farming practices may vary from one geographical area to another. This makes the yield in terms of amount and quality highly depends on the farmers knowledge and experience, as well as climate, unlink industrialized farming under control environment. As a result, there is a high natural variability in the pineapple sensory characteristics such that producing canned pineapple that have consistent quality requires highly skilled laoratory technicians.

This study involves predicting an important quality characteristic of the RM of a canned pineapple production process at one of the largest canned fruit producers in Thailand. In the production line, packing medium (PM) preparation is a critical process. This is because the cut-out strength of canned pineapple must reach a certain level to maintain the quality standard. The cut-out strength depends on the amount of PM filled and two other factors. The first one is total soluble solid (TSS), which is an amount of sugar content in the solution, measured in terms of degree of Brix (°Brix). The °Brix depends on the ripeness and succulence of the pineapples. The °Brix of the PM needs to match with the °Brix of the incoming raw material (pineapples) that contain high natural variability. The other factor is the ratio of RM to PM. Currently, the PM preparation is carried out by laboratory technicians, who measure the °Brix and prepare the right volume of PM. This makes the PM preparation process costly and time-consuming. To support the laboratory technicians in reducing the preparation cost and time, predicting the RM °Brix is a desirable task, since accurate prediction can guide the technicians to prepare the PM more efficiently.

A set of prediction models for the °Brix is constructed from the data collected at the RM receiving process. First, multiple linear regression model is fitted to investigate the relationship between the °Brix (dependent variable) and independent variables, including color, harvest month, and some indicator variables representing the RM's source. Machine learning (ML) algorithms have been utilized as very effective prediction tools in many fields because of their analytical capabilities (Wei et al., 2019). In this study, three widely used ML algorithms are implemented to gain high prediction performance.

ML is one of the most powerful techniques that can build accurate prediction models because it can extract complex relationship in large data without presuming the characteristic (e.g., linear or nonlinear) of the relationship among the variables. ML is not only flexible in handling various data characteristics, but it also can learn hidden patterns to improve its performance through training process (Seyedzadeh et al., 2018).

In addition, ML algorithms can deal with large number of observations (records) and number of variables or attributes (Voyant et al., 2017). With such capabilities, many ML models are applied to various fields such as financial analysis (Shin et al., 2005; Zhang et al., 1999), stock market (Lin et al., 2013), energy consumptions (Ascione et al., 2017; Ferlito et al., 2015), image recognition (Boughorbel et al., 2005; Wu et al., 2015), customer segmentation (Kashwan et al., 2013), quality control (Escobar et al., 2018; Lease, 2011; San-Payo et al., 2019; Tsen et al., 1996), and food industry (Amraei et al., 2017; Sanaeifar et al., 2016; Tsakanikas et al., 2020).

The ML algorithms implemented in this study include artificial neural network (ANN), support vector machine (SVM), and deep belief networks (DBN). ANN is one of the most powerful techniques to construct the prediction model from training data (Seyedzadeh et al., 2018). ANN can uncover complicated relationships in the training data, it is more prone to an overfitting problem. (Jin et al., 2005; Lawrence et al., 1997). Overfitting problem arises when the ANN training process is performed too excessively. An overfitted model would perform very well only in the model training process but perform worse for other unseen data in real usage environment (Akande et al., 2014). According to Piotrowski et al. (2013), the number of input variables can cause the overfitting problem.

SVM is another widely used technique that has an ability to handle the data that contain many input variables without causing much overfitting problem (Cai et al., 2004; Lin et al., 2013; Singla et al., 2011). SVM can deal with any complex problem through the use of kernel functions (Singla et al., 2011). Kernel functions can handle both linear and non-linear problem by transforming an input data to a suitable form to create an appropriate hyperplane, which has a high dimensional space (Cai et al., 2004; Moraes et al., 2013). However, choosing an appropriate kernel function to obtain desirable results requires modeler’s experience (Cawley et al., 2010). In addition, SVM features a complexity constant (C), a hyperparameter that reduces the errors that occur in the training process. With a proper value of C, SVM constructs the hyperplane that can classify almost all of the training dataset including even some outliers or noises in the training dataset. Such hyperplane, which highly relies on training dataset may lead to an outstanding performance for training dataset, but the model might not be generalized. Nevertheless, for a modeler who is inexperienced with the overfitting problem, SVM is considered a better model than ANN in this regard. Since SVM is a classification-based ML technique, its performance might be worse than the performance of ANN.

DBN is firstly introduced by Hinton et al. in 2006 (Hinton et al., 2006). DBN is a feedforward neural network that has the same architecture as ANN, but with a different training process. DBN has two main phases for building a prediction model, pre-training and fine-tuning phase. In the pre-training phase, the restricted Boltzmann machine (RBM) is applied with a greedy algorithm to obtain a good set of initial weights. After that, the weights in the whole network are adjusted in the fine-tuning phase by a backpropagation algorithm to obtain an optimal set of weights. While ANN only has one training phase. Major disadvantage of DBN is that it requires a large amount of computational resources and a high level of programming skill to construct and optimize its performance (Karhunen et al., 2015).

In addition, to obtain high prediction accuracy from the three ML models, grid search (GS) technique is applied to find appropriate values of hyperparameters for each ML algorithm. An advantage of GS is that it is simple to implement, and it can provide an optimal (or near optimal) hyperparameter setting. However, a major drawback is that GS requires much computational resource to perform since it tests all possible combinations of hyperparameter values in their respective given ranges. After performing GS, the three final ML models and the regression model are used to generate the predicted °Brix values for every batch of RM in the historical data. To improve the prediction performance, ensemble method is applied to the result of the prediction models to combine the result and generate a better result. In this study, simple average and weight average based on error are utilized to integrate the predicted value of each prediction model to obtain a new predicted value (Winkler & Makridakis,1983).

Moreover, decision support system (DSS) as the form of a database application, containing the predicted °Brix and the corresponding input variables values, is then developed to provide support to the laboratory technicians at the canned pineapple factory plant. The DSS can look up the predicted value of °Brix for the new incoming RM after the user enters the key input variable values. The DSS can also be used for RM procurement planning purpose, such as choosing geographic area(s) or supply source(s) to order RM in a given month.

Contributions of the study are as follows. Demonstrate the use of effective prediction models in an industrial application, along with how to perform hyperparameter settings for these models to improve the prediction accuracy. In addition, a use case of developing the prediction results and input variables into a database application that not only can give accurate prediction for the new incoming RM, but also help the RM procurement planning process.

1.2 Problem Statement

The raw material in the canned pineapple production includes fresh pineapple and other ingredients, such as sugar, citric acid, etc. In Thailand, canned pineapple producers purchase fresh pineapples from many farmers. Because of some geographical factors and cultivation practice, quality characteristics of fresh pineapples (i.e., flavor, size, color) from different farms usually vary. Production factory has to take a random sample to check the quality of incoming raw materials at the receiving operation. In the production process, the fresh pineapples are sorted and weighed, then transferred on conveyor through cleaning and cutting operations. In the cutting operation, pineapple peel and pineapple core are separated and sent to pineapple juice production. The main part of fresh pineapple goes through another cleaning operation and trimmed to remove the remaining peel before being sliced into the final shape (fancy slice). One whole pineapple usually gives eight to ten slices. The pineapple slices are sorted according to size and color, then packed into cans. In the next operations, the can is filled with packing medium, then removed the air and sealed before entering the sterilization process. The finished product is the canned pineapple containing fancy slices (i.e., whole slice) of pineapple rings, which is the highest grade in the market.

An important operation is the packing medium preparation. The samples of pineapple taken at the receiving operation is analyzed to measure their chemical characteristics to specify how the packing medium should be prepared. The most important characteristic is the °Brix (a measurement of the sucrose content) of the raw material. The packing medium that is prepared in advance has to match the °Brix of the raw material so that the finish product has the °Brix that is conformed to the specification. In the current practice at the factory, the packing medium preparation is carried out during the processing times of all operations before the packing medium is filled to the can. This may not give enough time for the laboratory technicians to prepare the packing medium, which result in production delay. Therefore, accurate prediction of the °Brix is a critical support to the laboratory technicians in order to reduce the time for packing medium preparation.

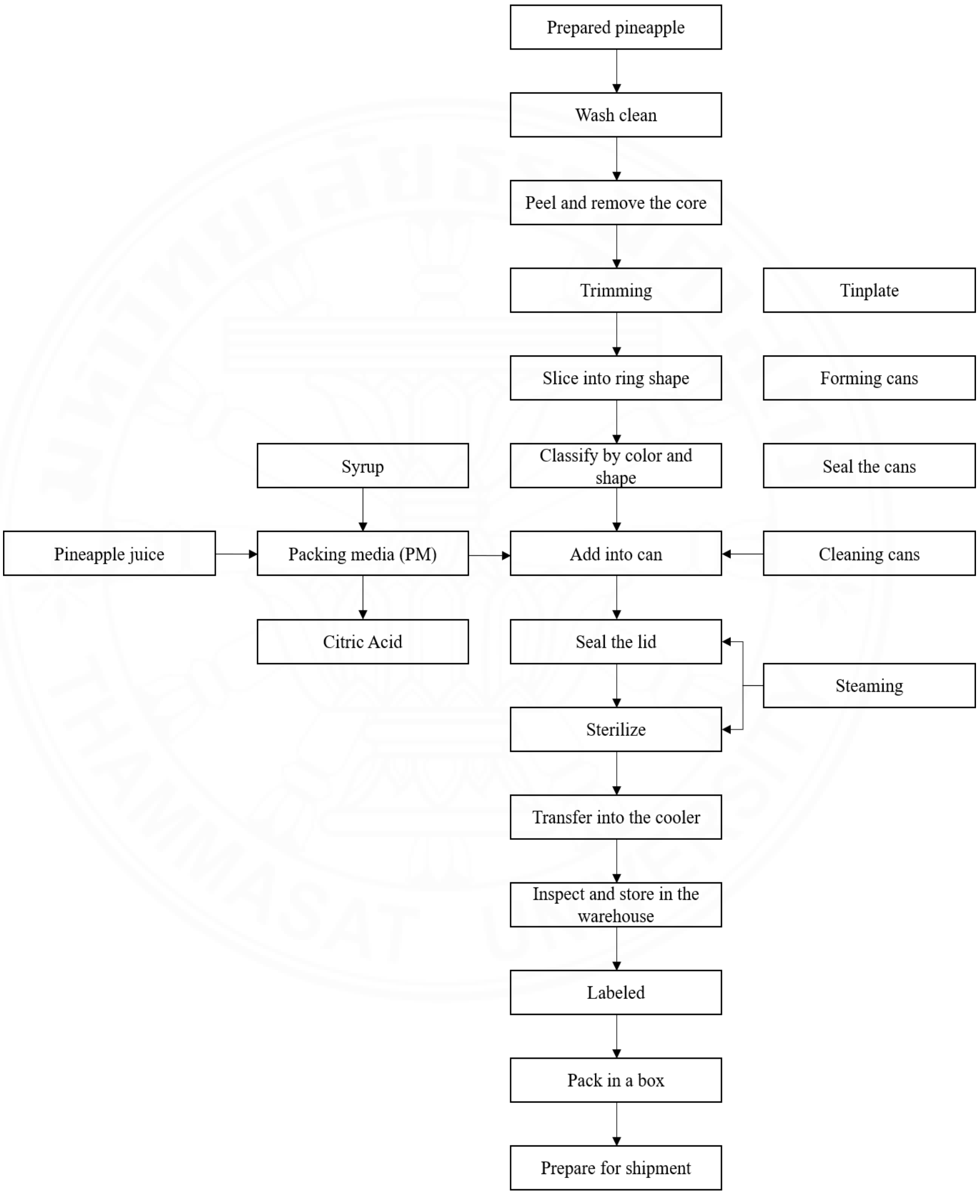


Figure 1.1 Step of making canned pineapple.

To perform the prediction, this independent study demonstrates an implementation of the regression model to identify the relationship and the pattern of the input data consists of color of pineapple, harvest month, geographical, and other collector identifications. Also, apply ML algorithms to construct the prediction model and to make the prediction models more effective, a grid search is performed to appropriately set the hyperparameters of the ML algorithms. In addition, the ensemble method is applied to the prediction model to improve the prediction performance. Moreover, to provide the DSS program for user to simplify the process of prediction for the users.

1.3 Objectives

- To analyze the characteristic of the raw material
- To construct multiple linear regression
- To construct machine learning prediction model
- To apply ensemble method with prediction model
- To apply prediction model with decision support system (DSS)

CHAPTER 2
REVIEW OF LITERATURE

Several research studies used ANN, SVM, and DBN for prediction problem that relate to food or agricultural products are reviewed and compared to this study. Some relevant studies that implement ANN as prediction model for various agricultural and food applications are as follows. Chia et al. (2012) used ANN to predict soluble solids content (SSC) of pineapple from four pineapple sample datasets collected from different days. In the study, ANN is trained by using one of the four datasets and the other three datasets are used as validation sets. The results show that the ANN can predict SSC content with the root mean square error from 0.71 to 1.01 °Brix. Kerdpiboon et al. (2006) proposed ANN to construct the prediction model for dried carrots to predict the shrinkage. The variables that are used in the study are moisture content and dimension of the cell wall structure. Moreover, the multiple regression model is applied for the comparison, it was found that ANN is better. Torkashvand et al. (2017) compared the result from ANN with multiple linear regression model for the prediction of kiwifruit firmness. The result shows that ANN has a better performance than another model. Zenoozian et al. (2007) applied ANN to construct the prediction model for the color intensity, the shrinkage, and shape factor (deformation) by considering moisture ratio, pumpkins thickness, air velocity, temperature, and time duration. The prediction accuracy of the model is satisfactory. Sucipto et al. (2018) constructed the prediction model by using ANN for sugar content and moisture content in the sugarcane based on the bioelectrical properties of the sugarcane. The model yields a satisfactory performance with the accuracy of 97.23 percent. Lashgari et al. (2017) applied a classification model based on acoustic features of apples (i.e., frequency and amplitude) to classify the firmness of golden and red apple during storage. The objective is to make an efficient quality assessment of the stocks for the trading purpose. The results indicated that a classification performance of approximately 85 percent of the F1 score, which represents the model’s accuracy. However, the authors stated that the classification accuracy still need to be improved.

A review of some recent studies that utilized SVM in livestock and agricultural practices is as follows. Alonso et al. (2013) studied a prediction application that used carcass weight as an important factor to calculate the price of beef cattle. The researchers proposed to use SVM to predict the weight before the slaughter process of animals. Amraei et al. (2017) used an SVM for live broiler chicken weight prediction. The result showed that using SVM to predict body weight was almost the same as performing manual measurement. Sanaeifar et al. (2016) used SVN to predict banana quality. The Inputs are color features in different color spaces, where TSS, firmness, pH, and acidity were determined as output. The result indicated that support vector regression has lower estimation error and require less computational time than ANN. Mukarev et al. (2012) predicted the °Brix of 1,053 samples of peach by using SVM. The peaches were harvested from different places and time across the US from 2002 to 2004. Moreover, they compared the accuracy between SVM and a partial least square regression. The results is that SVM performed better in term of accuracy. Meng et al. (2015) demonstrated the use of SVM to predict the °Brix in a syrup of sugarcane boiling process. Easy-to-measure variables are selected as inputs. Hyperparameters of the SVM are tuned by using a particle swarm optimization with cross-validation, to improve the accuracy and generalization capacity of the model. The results suggested that SVM can predict °Brix with satisfied value of maximal relative error (MRE).

A review of some studies that featured DBN is given next. Aulia et al. (2020) applied DBN to predict macronutrient, i.e., carbohydrate, protein, fat, and other essential nutrients for infant food. To obtain the characteristic of the food, a near-infrared spectroscopy is used before applying the DBN model for prediction. Längkvist et al. (2013) performed various types of models for classification problem of meat spoilage. The results showed that DBN could give faster and better classification performance than other approaches. Mohan et al. (2017) proposed the DBN to estimate crop production rate based on various factors, i.e., type of soil, season, availability of water, and risk factor. The results showed high potential for estimating the crop productivity based on a verification from real time data.

In the literature, there are many previous research studies that applied ML models to prediction problems in agricultural and food products. However, to the best of my knowledge, there is no previous research that developed the ML techniques to predict the RM °Brix to support the PM preparation process in a canned fruit production.

CHAPTER 3
METHODOLOGY

3.1 Data collection

The first phase of this study involves data collection about the batches of incoming raw material (fresh pineapples) are recorded during the RM receiving operation of the production, 2,202 batches of pineapples. A sample of pineapples is taken from each batch and analyzed to measure the °Brix. At the same time the pineapples are being preprocessed, i.e., size sorting, washing off physical contaminations, before entering production. The data obtained from each sample are recorded as color, °Brix, harvest month, geographic factors (i.e., province and district area), collector identification, farm location. The description of each attributes is shown in the following section.

3.1.1 Degree of Brix (°Brix)

The total soluble solid content is measured as °Brix, percentage of the total soluble solid content, which might consist of many types of soluble solid (i.e., various acid, saline, etc.), in the particular solution. In the canned pineapple production, the °Brix is used to determine the level of sweetness of the raw materials, work-in-process, and finished product. The refractive index is applied to measure the concentration of the sugar content by comparing to the standard sugar (i.e., sucrose). The °Brix of the raw material has an important role in determining the amount of PM, which must be prepared in advance, that makes the sweetness of the final product meet the specification. Accurate predicted value of the °Brix of RM can help to reduce the time to prepare PM to avoid any delay in the production process.

3.1.2 Color

Color is a physical characteristic that indicates the ripeness of the RM. In this application, color is rated on a five-point scale. A scale of one means that the pineapple pulp is the most yellow, and a scale of five the pineapple pulp is the palest. Each pineapple may contain different levels of color in its pulp, where more yellow section may contain high sweetness level (high value of °Brix), and in the pale section may contain low sweetness level (low value of °Brix). From multiple boxplot, there seems to be some difference in the °Brix as shown in Figure 3.1. A preliminary statistical analysis using ANOVA also confirms significant difference in the average °Brix for different levels of color (see Table 3.1).

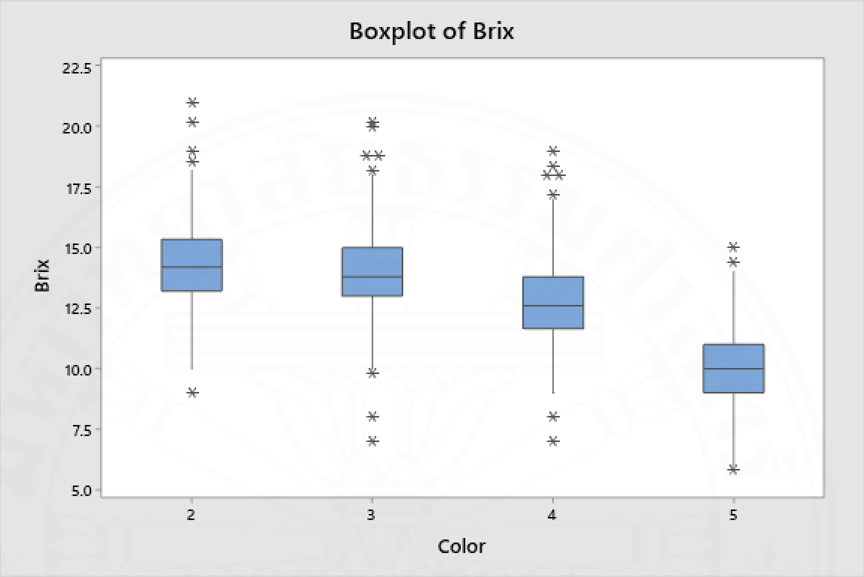


Figure 3.1 Average °Brix with respect to color.

Table 3.1 Analysis of variance table of color.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Color	3	6948	2315.87	943.74	0.000
Error	2174	5335	2.45		
Total	2177	12282			

3.1.3 Harvest month

In a season, harvest months are different in terms of environment factors, e.g., temperature, humidity, rainfall, etc. From multiple boxplot of the °Brix, there seems to be some differences in the °Brix as shown in Figure 3.2. Similar analysis using ANOVA also confirms significant differences in the average °Brix for different harvest month (see Table 3.2).

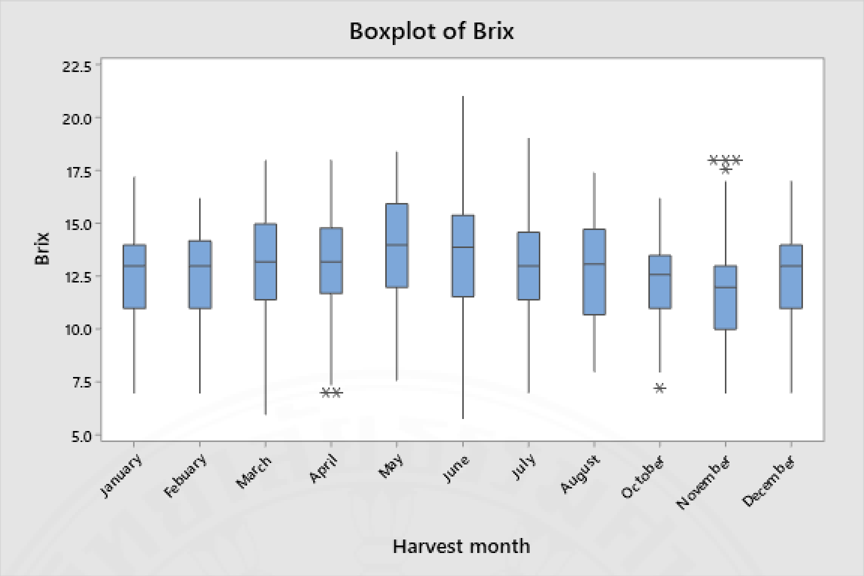


Figure 3.2 Average °Brix with respect to harvest month.

Table 3.2 Analysis of variance table of harvest month.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Month	10	722.9	72.290	13.54	0.000
Error	2191	11698.4	5.339		
Total	2201	12421.3			

3.1.4 Geographic variables

Geographic variables that indicate districts and provinces where RM comes from are recorded. These variables represent different weather conditions at different geographical areas. Actual cultivating conditions in different geographic areas, such as temperature during harvesting, rainfall, type of soil, etc. are not directly considered to simplify the data collection process. This is based on a hypothesis that these variables sufficiently representing geographical factors for the purpose of the prediction models.

The canned pineapple producer has made farming contracts with 24 harvest collectors, which could be agriculturists or intermediary traders, who collect RM from one or more farming areas. Multiple boxplots suggest some differences in the °Brix with respect to district, province, and collectors, as shown in Figure 3.3-3.5. Similar analysis using ANOVA also confirms significant differences in the average °Brix among district, province, and collectors (see Table 3.3-3.5).

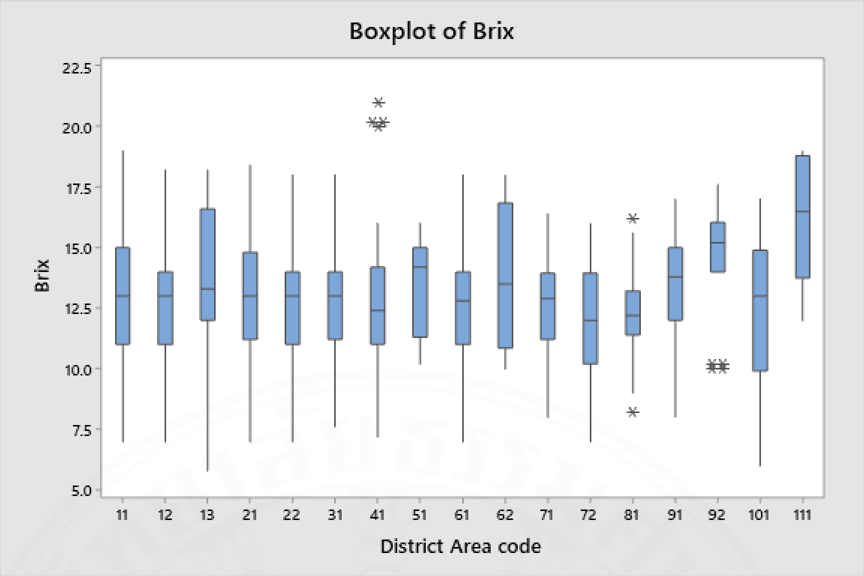


Figure 3.3 Average °Brix with respect to district area.

Table 3.3 Analysis of variance table of District area.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
District	16	335.1	20.941	3.79	0.000
Error	2161	11947.4	5.529		
Total	2177	12282.4			

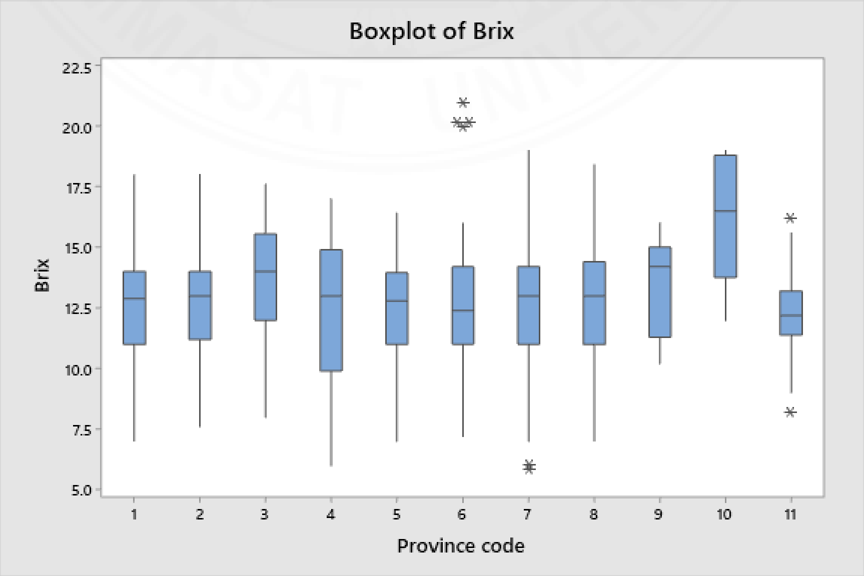


Figure 3.4 Average °Brix with respect to province.

Table 3.4 Analysis of variance table of Province.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Province	10	223.4	22.336	4.01	0.000
Error	2167	12059.1	5.565		
Total	2177	12282.4			

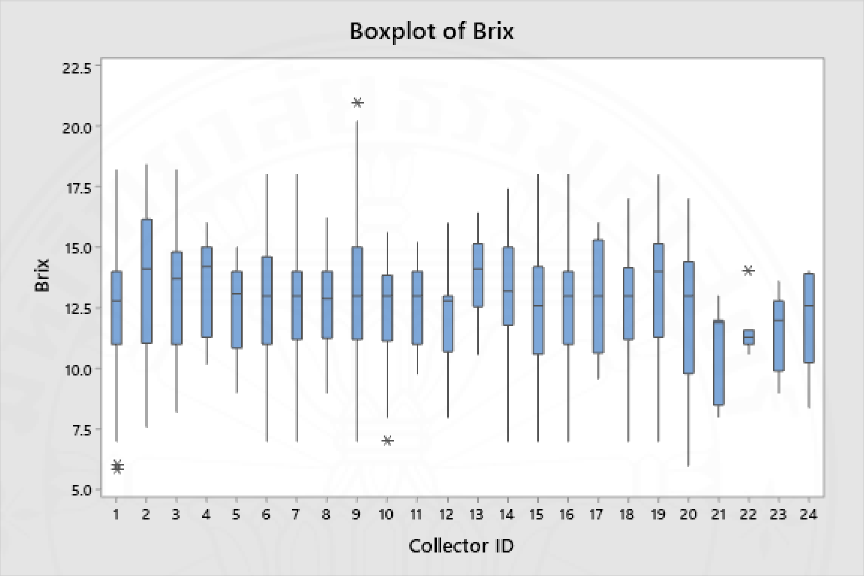


Figure 3.5 Average °Brix with respect to collector.

Table 3.5 Analysis of variance table of Collector.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Collector	23	286.6	12.460	2.24	0.001
Error	2154	11995.9	5.569		
Total	2177	12282.4			

3.2 Data preparation

After data collection, they are preprocessed into the types of data that fit the framework of the prediction models. The data preparation is performed by creating binary variables to represent each categorical variable. For example, 12 binary variables are used to indicate each harvest month.

In the model training process, the data are randomly partitioned into three datasets, which are training dataset, validate dataset, and test dataset. The training dataset, which contains 70% of all data (1,541 batches), is used for the prediction model to learn patterns of the data. The validation set contains 20% of the data (441 batches) that the prediction model has never seen. This dataset is used to prevent model overfitting and to evaluate the model for the hyperparameters tuning. The test set is the last 10% of the data (220 batches), also unseen, for evaluating the performance of the prediction model in terms of the model generalization.

3.3 Model evaluation

In this study, the regression model and three ML algorithm models with the different setting of hyperparameter are evaluated using mean absolute error (MAE) as shown in Eq. 1 and mean absolute percentage error (MAPE) as shown in Eq. 2, respectively.

$$\mathrm{MAE} = \frac{\sum_{i}^{N} |y_i - x_i|}{N} = \frac{\sum_{i}^{N} |e_i|}{N} \tag{1}$$

where y_i and x_i are the predicted value and the real value of sample i, respectively, e_i denotes the error of prediction for i, and N denotes the total samples

$$\mathrm{MAPE} = \frac{\sum_{i}^{N} \left|\frac{y_i - x_i}{x_i}\right| \times 100}{N} = \frac{\sum_{i}^{N} \left|\frac{e_i}{x_i}\right| \times 100}{N} \tag{2}$$

where y_i and x_i are the predicted value and the real value of sample i, respectively, e_i denotes the error of prediction for i, and N denotes the total samples.

3.4 Regression model

Regression is commonly used to identify important predictors, forecast a trend, and estimate factor effects. It is a powerful statistical tool for modeling the relationships between multiple variables. One variable is treated as a response variable to be predicted and other variables are its predictor. The relationship is expressed as a linear function in terms of the regression model parameters as shown in the following equation.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_i X_{ki} + \varepsilon_i \quad ; i = 1,2,\ldots,n \tag{3}$$

where Y_i denotes observed value of °Brix, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ denote the coefficients of each independent variable, $X_1, X_2, \ldots, X_k$ denote the independent variables, and ε_i denotes the error term.

3.5 Machine learning

Machine Learning algorithm (ML) is algorithms that automatically learn through the data without human intervention.

3.5.1 Support vector machine

Support vector machine (SVM) has been created in 1995, which is a supervised learning model to apply with classification or regression problem by associated with learning algorithm (V. N. Vapnik, 1995). Support Vector Machine is a famous technique for classification and also can handle the regression problems (da Costa, N. L., et al., 2018; MIN, J., & LEE, Y., 2005). Most of their performance seems outstanding other models, and not only for the performance. Most of machine learning such as ANN can overfit when the number of training cycle is too much, while SVM does not has this problem. Main mechanic of SVM is hyperplane (linear classifier) that separates the data. Many hyperplanes can be used as a classifier. Still, the model must find the optimal hyperplane that separates and has the largest margin or maximum margin between each class's dataset, which is called the optimal hyperplane. Simultaneously, SVM memorizes the data that located near the boundary of each classification group and called them as Support Vectors. When SVM is used to analyzed to a non-linear data, a kernel function is applied to transform the data into linear data. However, the type of kernel function must be matched with the type of the non-linear data to convert the data into linear data effectively (Yeh, C. Y., Huang, C. W., & Lee, S. J., 2011)

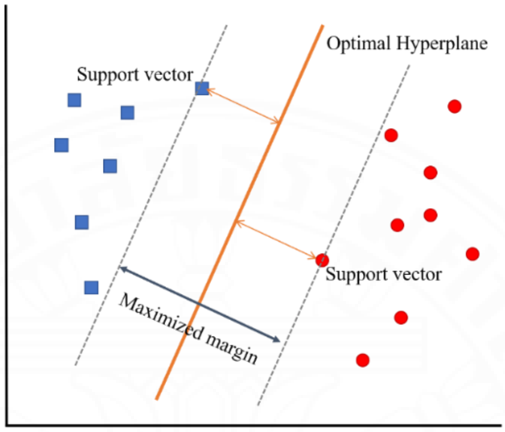


Figure 3.6 Support vector machine (SVM).

3.5.2 Artificial neural network

Artificial neural network (ANN) is the one of the most popular ML algorithms over decades. ANN simulates the human neural system's mechanism in generating an output from multiple inputs. The human neuron system consists many neuron cells. Each neuron cell has three main components: Dendrites, cell body, and axon. Dendrites will receive the signals from the other neuron cells, then the signals are sent to the cell body. In the cell body, the signals' level is adjusted. Then, the adjusted signals are combined and sent to another neuron cell through the axon as the output.

ANN follows the mechanisms of neural system in human as follow: ANN acquires knowledge through training, and knowledge is stored within the node interconnection, also known as interconnection weight. ANN has a similar learning capability to human, which ANN is learned by training. The training process of ANN is to adjust the interconnection weight. For the node interconnection is comparable to how the information is kept human's cell. The structure of ANN is concluding an input layer, hidden layer(s), and an output layer that are connected in a network shape (Dave Anderson and George McNeill., 1992). In this study, a feed-forward neural network

(FFNN) is applied to construct the prediction model. The predicted value is obtained at the output layer.

At the beginning of the training process, the initial interconnection weights are randomly set up. Then, the backpropagation algorithm is applied to iteratively reduce the prediction error value by adjusting the interconnection weights to make the predicted value close to the actual value. In each training cycle, the error is computed and backpropagated to determine the changing in the interconnection weights. However, a learning rate is applied to control how much to change the interconnection weights according to the results from the backpropagation algorithm.

An adjusted interconnection weight is transferred back to the network and repeat the algorithm of ANN, which repeats itself according to the number of training cycle which is one of the hyperparameters. The number of training cycle should be set carefully, if the value is too low means the model is not training enough. Also, if the value if too high means there is possibility that the model is over-fit. ANN approach has been used in prediction of moisture content of agricultural products prediction, prediction of viscosity, iodine, and other factors in biodiesel, and prediction of viscosity of fruit juice with ambience factors (Barradas Filho, et.al., 2015; P. Rai, et.al., 2005; Topuz, A., 2010).

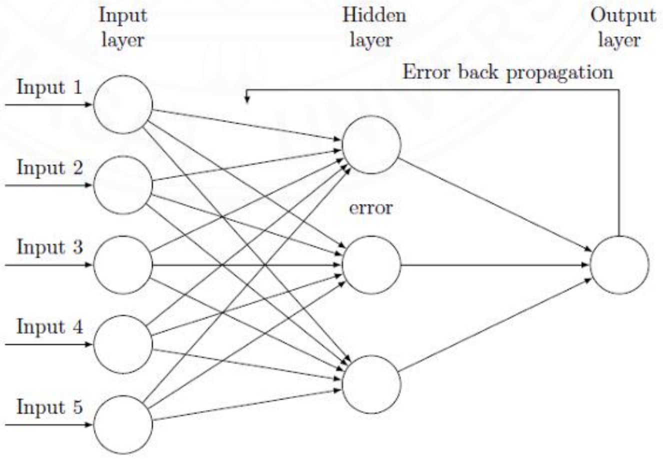


Figure 3.7 Artificial neural network architecture (Source: maxversace.com).

3.5.3 Deep belief network

Deep belief network (DBN) is introduced by Hinton in 2006 with a motivation to make a faster learning process (Hinton, G. E., et al., 2006). The structure of DBN is multilayer of Restricted boltzmann machine (RBM) in FFNN composition. RBM is an algorithm that can learn the pattern of data. The RBM structure is composed by two layers, including an input layer and a hidden layer of neurons (Hinton and Sejnowski, 1986). With the ability of initial values (i.e., bias and interconnection weights) generating, DBN uses RBM in the pre-training phase. that applies a greedy algorithm to initialize the bias and interconnection weights, and then they are sent to the fine-tuning phase. (Kuremoto, T., et al., 2014).

The fine-tuning phase in DBN utilizes the backpropagation algorithm as in ANN, which usually takes a long time to process, especially when the initial values are inappropriate. With the model that applied RBM to find the initial value, the backpropagation algorithm's processing time is shorter. In addition, the performance is usually better since the initial values are already optimized. This mechanism gets rid of the ANN's drawback, which is randomization of the initial values. After the fine-tuning phase is finished, the adjusted interconnection weight is applied to the output layer for the prediction. (Abraham, A., 2005).

DBN can be applied as prediction model in various fields such as prediction reservoir landslide displacement and prediction for deformed coal. (Li, H., et al., 2020; Wang, X., et al., 2020)

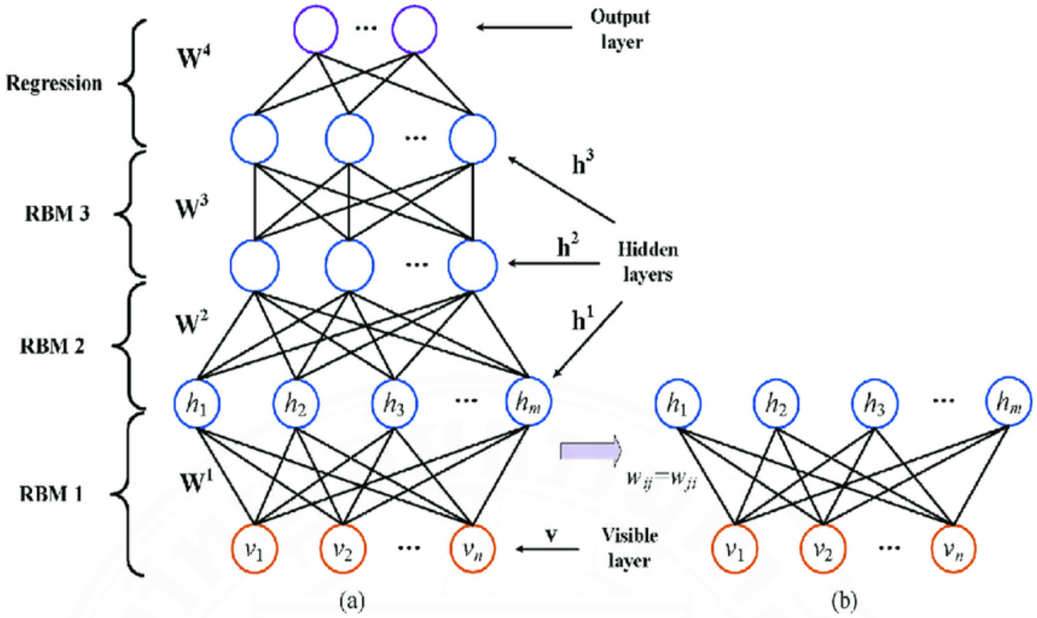


Figure 3.8 (a)Deep Belief Network architecture

(b) Restricted Boltzmann machines (RBM)

(Source: Shao, Haidong, et.at. (2018).

3.5.4 Grid search (GS) for hyperparameter tuning

Grid search is the basic technique for finding appropriate hyperparameter setting of ML models . GS requires a lower bound, an upper bound, and the number of steps for each hyperparameter. Then, all possible combinations of hyperparameters are generated and tested to determine the setting that maximizes the ML models performance (Bergstra, J., & Bengio, Y., 2012 ;Lin, S. W., et.al., 2008). In this study, GS is implemented on ANN, SVM, and DBN. The lists of hyperparameters and their ranges for ANN, SVM, and DBN are in Tables 3.6-3.8, respectively.

Table 3.6 ANN hyperparameters list.

Hyperparameter	Description	Lower bound	Upper bound	Steps
HN	Number of hidden nodes	1	20	20
TC	Number of training cycles	10	1000	21
LR	Learning rate	0.0001	0.1	21

The hyperparameters that are considered in ANN are shown in Table 3.5 above. The number of hidden nodes is used to characterize complex relationships among the inputs, the more the hidden nodes, the higher the complexity. Learning rate is the rate of change of interconnection weights of input variables that connect with hidden nodes, the lower learning rate, the slower the learning process and larger training cycle, the more complexity of the input that model can analyze. However, the complexity of inputs is increasing the computational time.

Table 3.7 SVM hyperparameters list.

Hyperparameter	Description	Lower bound	Upper bound	Steps
C	Complexity constant	-1	100	102
Conv	Convergence epsilon	0.001	0.1	11

For SVM, the two common hyperparameters include complexity constant (C) and convergence epsilon (Conv). C is the hyperparameter that controls the preciseness of the classification. For the large value of C, the model would give more concern on the misclassifying than maximizing margin between the hyperplane and support vectors. Conversely, a small value of C is focusing on maximizing the margin but does not concern much about the misclassifying. Conv is a stopping criterion according to Karush-Kuhn-Tucker (KKT) constraint of the training algorithm for adjusting the hyperplane iteratively according to the errors in the last iteration. The algorithm stops when the improvement of the error in the current iteration is smaller than the value of Conv. Then, four kernel functions are tested including dot, anova, epachnenikov, and radial.

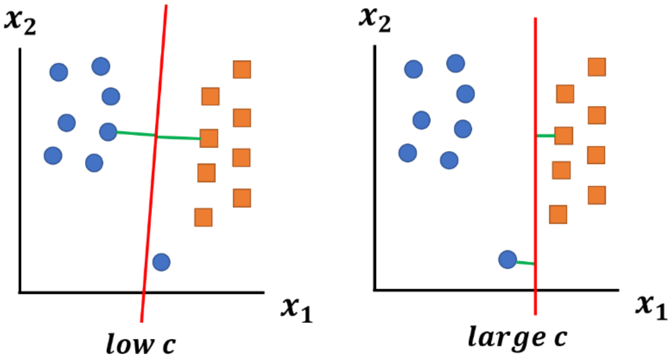


Figure 3.9 The mechanism of the value of C.

Table 3.8 DBN hyperparameters list.

Hyperparameter	Description	Lower bound	Upper bound	Steps
HN	No. of hidden nodes	1	20	3
n_i	No. of pre-training cycles	100	1,000	3
N	No. of training cycles	500	10,000	3
LR_{RBM}	Pre-training learning rate	0.01	0.1	3
LR	Learning rate	0.0001	0.01	3
N_b	Batch size	30	150	3

As the structure of DBN, firstly, the pre-training phase, the hyperparameters consist of number of hidden nodes, number of pre-training cycles, and pre-training learning rate. In fine-tuning phase, number of training cycles, learning rate. An additional hyperparameter for both training phases is batch size. Description of DBN hyperparameters is the same as in ANN except batch size, which is number of samples to be passed through the model in each time that the interconnection weights are adjusted. With the smaller number of batch size, the longer computational time for each training cycle is required.

3.6 Ensemble method

Ensemble method is the method that combine multiple models which have been a well-known method that applied with machine learning since 1990s. The ensemble method integrates predicted value from each model to get a synthetic predicted value. In addition, the ensemble of model must be more accurate than the individual model (Winkler, R., & Makridakis, S.,1983). However, it has been noticed that the ensemble method does not always yield a better performance. As state in no free lunch theorem, which state that no algorithm always obtain the most accurate in all situations (Wolpert and Macready, 1997). Therefore, most researchers are focusing on constructing an appropriate ensemble method that fit well with the particular data for each study. In this study, the focused ensemble methods are simple average and weighted average based on error.

3.6.1 Simple average method

Simple average method is the simplest form for ensemble method, the average predicted values are calculated by taking the average from each prediction model by using Eq.5.

$$y_i = \frac{\sum_i^k d_i}{k} \tag{4}$$

where d_i denotes as prediction result of method i, k denotes as the total models that used for ensemble method, and y denotes as result of ensemble method

3.5.2 Weight average based on error

Weight average is a modification of simple average method, which the prediction result of each method is multiplied with its own weight. In this study, weight is defined based on the mean absolute error from each prediction model, which can be calculated by using Eq. 3. The weight of each model is expressed by Eq. 5 and the weight and the predicted value of the ensemble method by using weight average based on error can be calculated by the following Eq. 6-7.

$$\sum_{i=1}^{k} W_i = 1, \qquad 0 \leq W_i \leq 1, \qquad i = 1,2,3, \ldots k \tag{5}$$

$$W_i = \frac{e_i^{-1}}{\sum_i^k e_i^{-1}} \tag{6}$$

$$y_j = \sum_i^k W_i d_{ij} \tag{7}$$

where W_i denotes as the weight of the prediction model i, d_i denotes as predicted value from model *i*, e_i denotes as error from model *i*, *k* denotes as total model that use for ensemble method and *y* denotes as predicted value of ensemble method.

CHAPTER 4
RESULT

4.1 Regression result

Five regression models are constructed to screen out insignificant independent variables by stepwise regression using a significant level of 0.05. The considered variables for each model are as shown in Table 4.1.

Table 4.1 List of variables in the regression models.

Model	Month	Color	District area	Collector	Farm location
Base model	✓	✓			
Model A	✓	✓	✓		
Model C	✓	✓		✓	
Model AC	✓	✓	✓	✓	
Model F	✓	✓			✓

1. Base model is the model that includes harvesting month and color variable.
2. Model A is the base model extension that also considers district area (regional planting zone). This is to test whether district areas, which may have different plantation conditions (climate, soil type, water source, etc.), have significant association with the °Brix.
3. Model C is another extension that includes information regarding RM collector. This variable represents a cluster of farms in the vicinity of the collector yard. In Thailand, many small farms in an area usually sell their harvests to a nearby collector yard, which gathers RM in that area and transport to the buyer (factory). This variable is tested to see whether different collectors (i.e. clusters of farms) have some relationship with the °Brix.
4. Model AC contains both district area and collector variables in addition to the base model.
5. Model F is the extended model that contains the farm location variable.

Five models from stepwise regression are used to predict the °Brix. The best model with the lowest MAE and MAPE is model C. Prediction performance on the validation and test datasets are provided in Table 4.2.

Table 4.2 Summary of the performance of the five regression models.

Model	MAE		MAPE	
	Validation	Test	Validation	Test
Base model	1.0986	1.1279	9.21%	9.24%
Model A	1.0823	1.1284	9.14%	9.22%
Model C	1.0816	1.1100	9.15%	9.09%
Model AC	1.0960	1.1316	9.29%	9.27%
Model F	1.2049	1.1627	10.37%	9.53%

The regression equation and ANOVA from Model C is shown in equation (8) and Table 4.3, respectively. Figure 4.1-4.2 illustrate that the regression model assumptions are satisfied.

$$\begin{aligned} y = 11.057 &+ 2.657x_1 - 0.5919\,x_1^2 + 0.562\,x_2 + 0.642\,x_3 + 1.303\,x_4 \\ &+ 0.900\,x_5 + 0.427\,x_6 - 0.430\,x_7 - 0.773\,x_8 \\ &+ 0.2389\,x_9 + 0.308\,x_{10} + 0.895\,x_{11} + 0.281\,x_{12} \\ &+ 0.484\,x_{13} + 0.776\,x_{14} + 1.251\,x_{15} + 0.822\,x_{16} \qquad (8) \end{aligned}$$

where y = °Brix, x_1 = color (a numerical variable), and the rest are indicator variables:

x_2 = 1 if harvest month is March, or 0 otherwise,

x_3= 1 if harvest month is April, or 0 otherwise,

x_4= 1 if harvest month is May, or 0 otherwise,

x_5 = 1 if harvest month is June, or 0 otherwise,

x_6= 1if harvest month is July, or 0 otherwise,

x_7= 1 if harvest month is October, or 0 otherwise,

x_8= 1 if harvest month is November, or 0 otherwise,

x_9= 1 if RM is from collector identification 1, or 0 otherwise,

x_{10} = 1 if RM is from collector identification 3, or 0 otherwise,

x_{11}= 1 if RM is from collector identification 2, or 0 otherwise,

x_{12} = 1 if RM is from collector identification 14, or 0 otherwise,

x_{13} = 1 if RM is from collector identification 18, or 0 otherwise,

x_{14} = 1 if RM is from collector identification 4, or 0 otherwise,

x_{15} = 1 if RM is from collector identification 13, or 0 otherwise, and

x_{16} = 1 if RM is from collector identification 19, or 0 otherwise.

Table 4.3 Analysis of Variance of regression model (Model C).

Source	**DF**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
Regression	17	6832.2	401.893	190.04	0.000
x_1	1	280.8	280.767	132.76	0.000
x_1^2	1	693.6	693.596	327.97	0.000
x_2	1	43.9	43.894	20.76	0.000
x_3	1	53.1	53.074	25.10	0.000
x_4	1	163.3	163.280	77.21	0.000
x_5	1	119.5	119.453	56.48	0.000
x_6	1	29.5	29.491	13.94	0.000
x_7	1	17.6	17.553	8.30	0.004
x_8	1	109.3	109.340	51.70	0.000
x_9	1	13.0	12.984	6.14	0.013
x_{10}	1	6.3	6.282	2.97	0.085
x_{11}	1	24.2	24.219	11.45	0.001
x_{12}	1	8.5	8.464	4.00	0.046
x_{13}	1	38.2	38.180	18.05	0.000
x_{14}	1	9.3	9.307	4.40	0.036
x_{15}	1	22.7	22.726	10.75	0.001
x_{16}	1	14.7	14.697	6.95	0.008
Error	1963	4151.4	2.115		
Total	1980	10983.5			

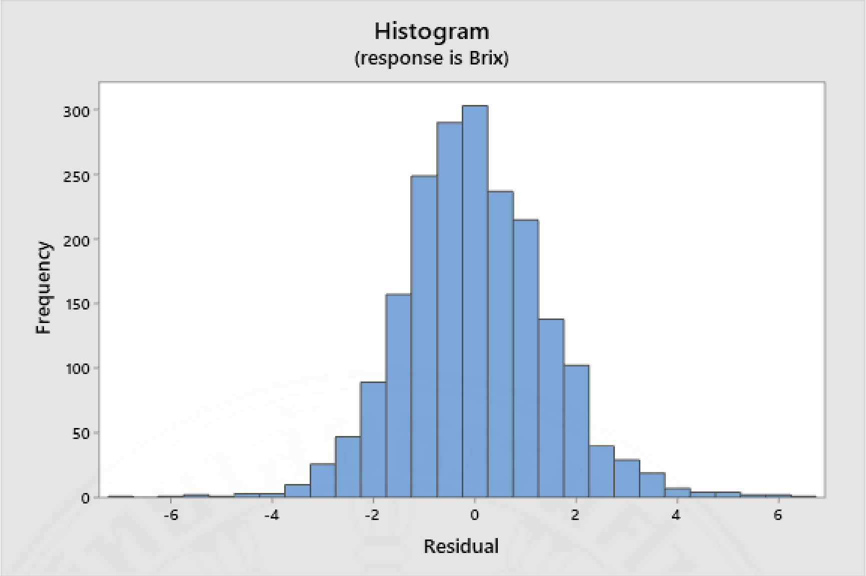


Figure 4.1 Histogram of residuals

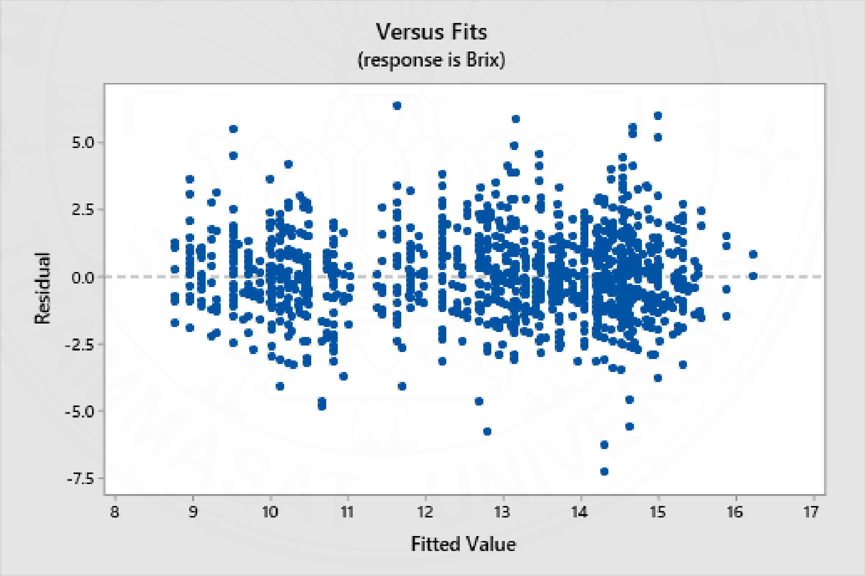


Figure 4.2 Scatter plot of residuals versus fits

4.2 Machine learning result

After performing the experiment according to the combination of hyperparameters which is generated by GS technique, the least value of MAE and MAPE of the validation set with their combination of hyperparameter for each ML models are provided in Table 4.4.

To verify the performance of the hyperparameter set, these hyperparameters are applied to the model with the new unseen dataset test set), and the result is shown in Table 4.5.

Table 4.4 Best hyperparameter set for validation dataset.

Model and its Hyperparameter value	No. of runs	Validation	
		MAE	MAPE
ANN: HN = 20, LR = 0.005095, TC = 159	4,410	1.043	8.60%
SVM: C = 1, Conv = 0.0901, Kernel Type = Anova	4,488	1.082	8.88%
DBN: HN = 11, N_i = 550, N = 10,000, LR_{RBM} = 0.1, LR = 0.00505 N_b = 90	729	1.035	8.54%

Table 4.5 The confirmation run result from the best hyperparameter set.

Model	Test set	
	MAE	MAPE
ANN	1.086	9.04%
SVM	1.047	8.86%
DBN	1.064	8.96%

4.3 Ensemble method result

The result from previous section is applied to the ensemble method to obtain more accuracy prediction value, and the result is shown in Table 4.6. Along with the comparison between each method. Also, the result comparison visualizations are as shown in Figure 4.3-4.8.

Table 4.6 Ensemble method results with individual prediction model result.

Model	Test set	
	MAE	MAPE
Regression	1.110	9.09%
ANN	1.086	9.04%
SVM	1.047	8.86%
DBN	1.064	8.96%
Ensemble (simple average)	**0.977**	**8.27%**
Ensemble (weight average)	**0.979**	**8.28%**

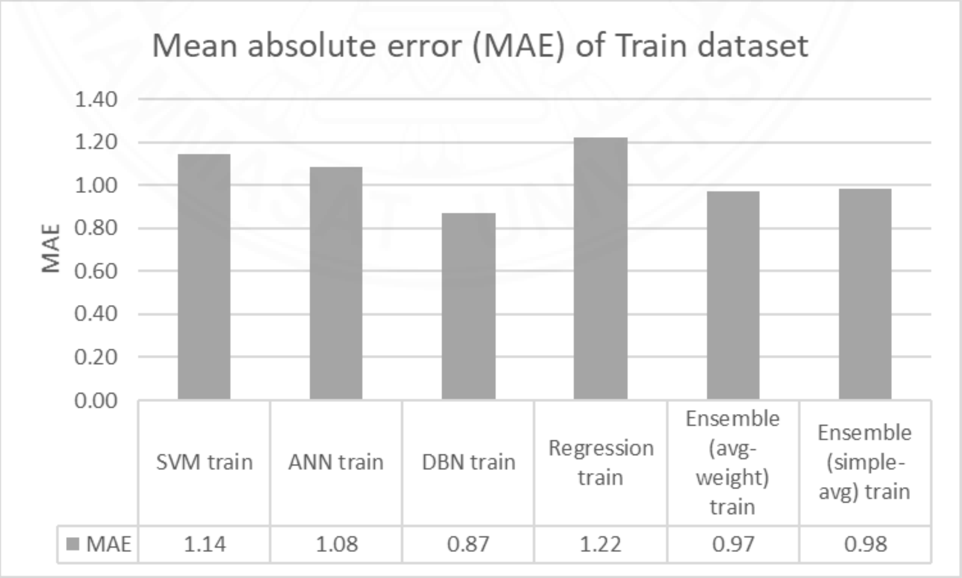


Figure 4.3 Mean absolute error of train dataset.

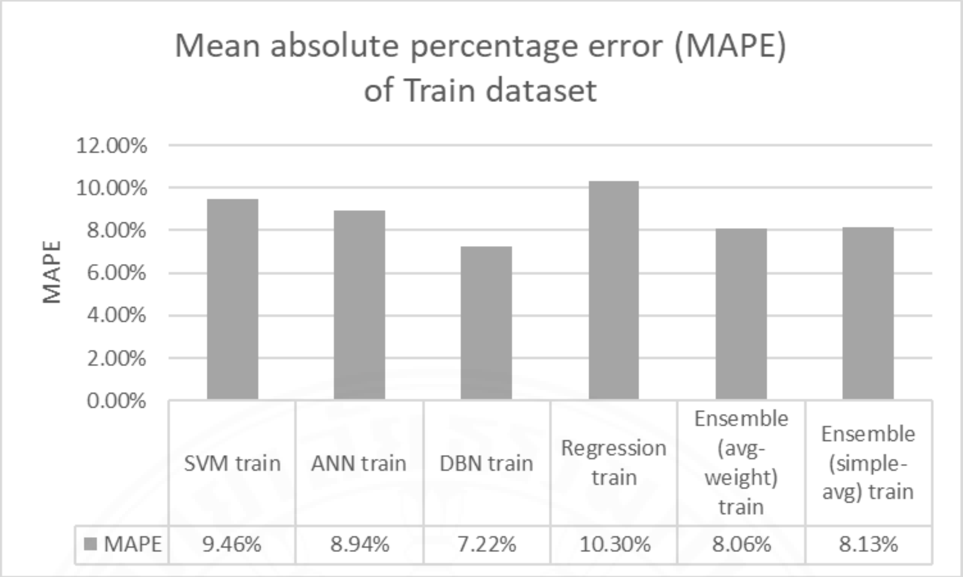


Figure 4.4 Mean absolute percentage error of train dataset.

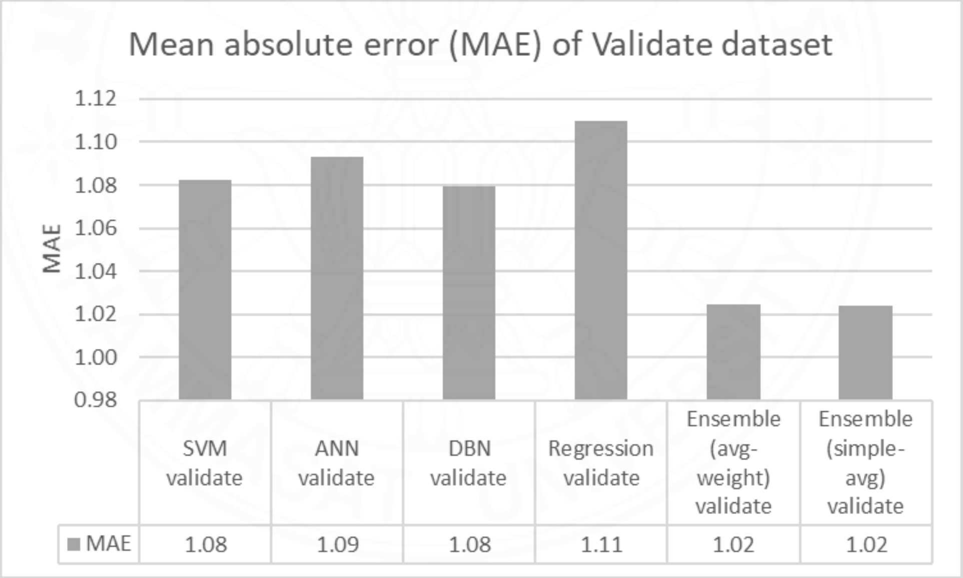


Figure 4.5 Mean absolute error of validate dataset

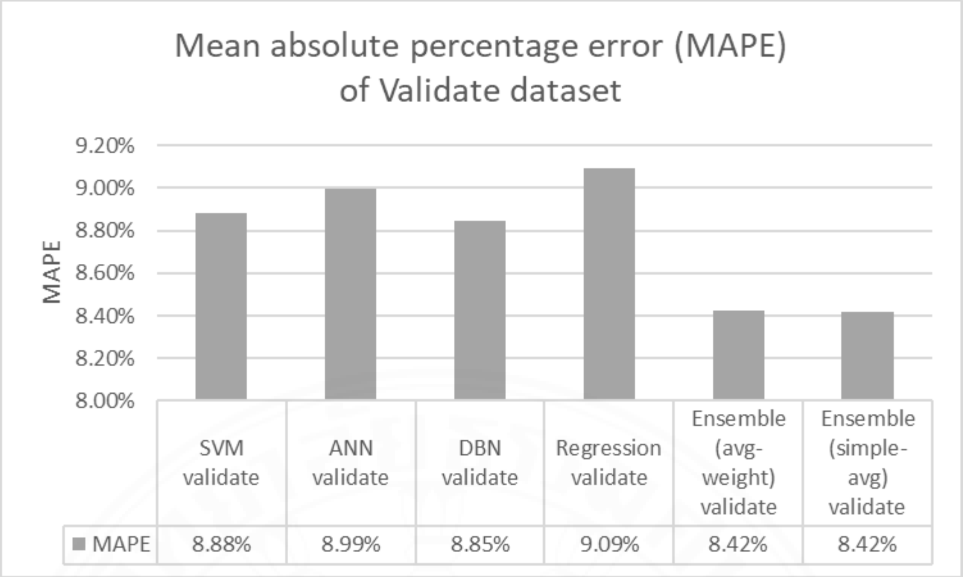


Figure 4.6 Mean absolute percentage error of validate dataset.

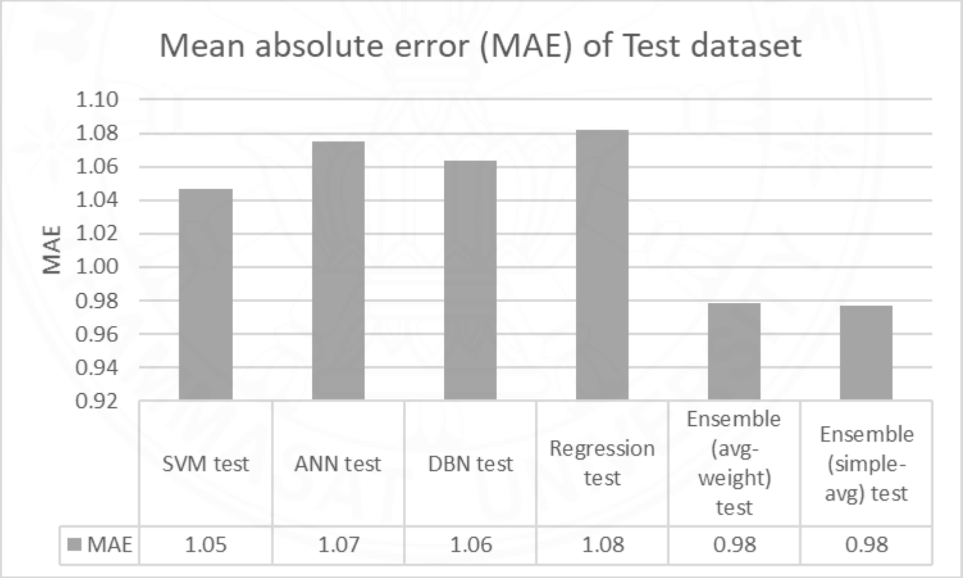


Figure 4.7 Mean absolute error of test dataset.

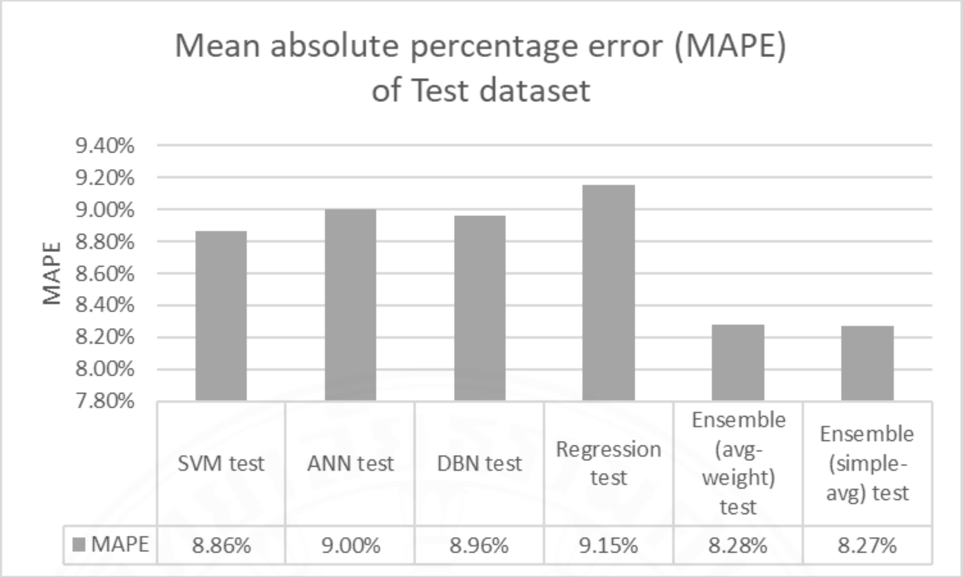


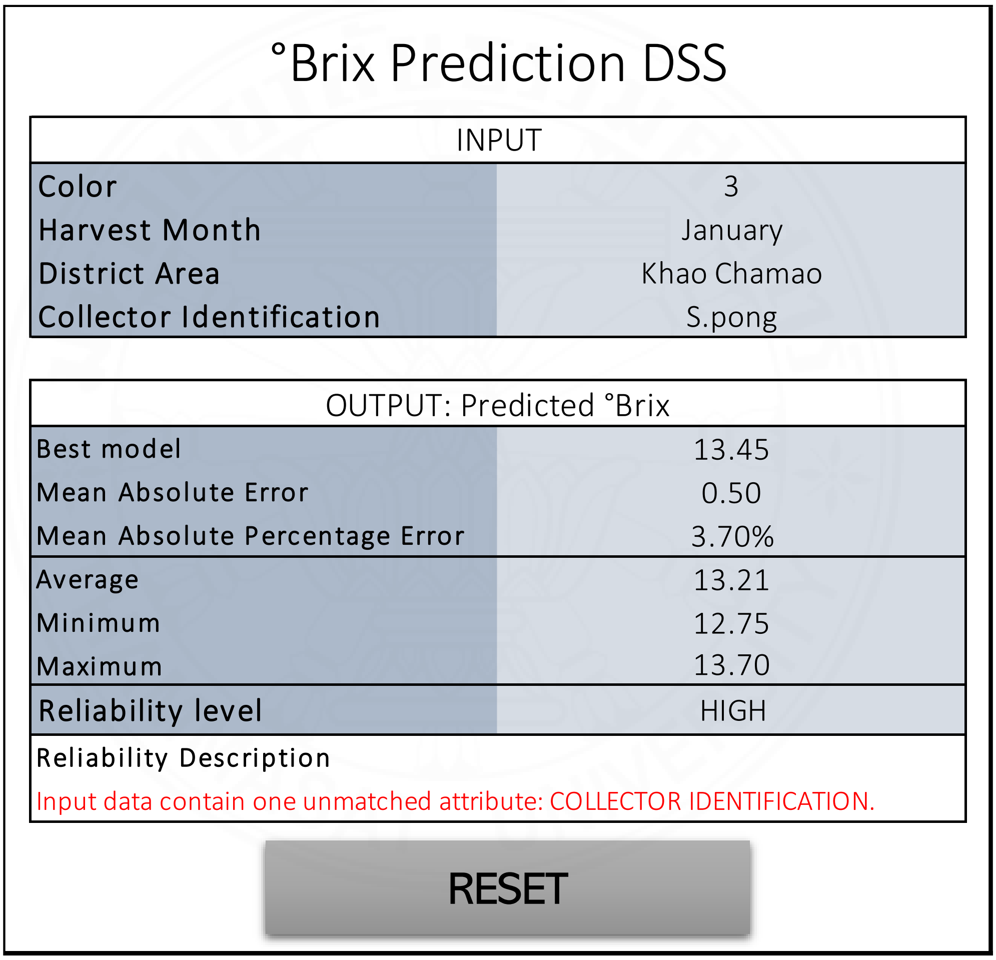
Figure 4.8 Mean absolute percentage error of test dataset.

From the results, the performance of ML models is satisfactory, especially the performance of SVM. After considering the number of runs, DBN is preferable since it requires a much smaller number of runs than SVM and ANN, while achieving relatively the same performance. A major disadvantage of ML models is their “black-box” nature. The black-box algorithms do not allow human modeler to understand the internal structure of the algorithm, even if the desired result is obtained. Therefore, significance of input variables are evaluated from the data analysis.

The analysis in section 3.1 shows that the harvest month has a significant impact on the sweetness of the pineapple. The months that that yield very sweet pineapple are May and June, while the months that produce little sweet are October and November. The next factor is the color of the pineapple. With a low grade of color or the most yellow, the sweetness is the highest. When the color is less yellow, the pineapple is as sweet. Lastly, the geographical factors also impact the sweetness of the pineapple due to the differences in temperature, rainfall, soil characteristic, and fertilizer.

4.4 Decision Support System for the Industrial User

A decision support system (DSS) is developed as an information system that supports the PM preparation process for the canned pineapple producer. The DSS is a simple and easy-to-use spreadsheet database containing historical data of key input variables, actual °Brix, and its prediction values from four final prediction models. It requires four key input variables including color, harvest month, district area, and collector identification of the incoming RM. The output contains the predicted °Brix from the best prediction model and its estimated errors (MAE, MAPE), as well as average, minimum, and maximum predicted °Brix from all four models. The user interface of the DSS is as shown in Figure 4.9.



°Brix Prediction DSS

INPUT	
Color	3
Harvest Month	January
District Area	Khao Chamao
Collector Identification	S.pong

OUTPUT: Predicted °Brix	
Best model	13.45
Mean Absolute Error	0.50
Mean Absolute Percentage Error	3.70%
Average	13.21
Minimum	12.75
Maximum	13.70
Reliability level	HIGH

Reliability Description

Input data contain one unmatched attribute: COLLECTOR IDENTIFICATION.

Figure 4.9 User interface of the decision support system (DSS)

In addition, the DSS reports the reliability level of the predicted values. The reliability level is determined by comparing the user input data with the existing historical records. The reliability level is classified into four levels as shown in Table 14. A very high reliability level indicates that the four key input variables from the user exactly match with the historical records. In other words, there exists at least one record of the RM with this color, harvested in this month, from this district area, and supplied by this collector in the historical data that are used to construct the prediction models. A high reliability level means that the user input data has one unmatched attribute, as reported in the reliability description. For example, the database contains at least one record of incoming RM with the same color, harvested in the same month, and from the same district area. However, the incoming RM is supplied by a different (or new) collector. A medium reliability level is reported when the user input data contain two unmatched attributes. Finally, the low reliability level indicates that the user input data do not match with the existing records.

Table 4.7 Reliability description.

Reliability level	Description
Very high	Input data exactly match existing records.
High	Input data contain one unmatched attribute: COLOR.
	Input data contain one unmatched attribute: HARVEST MONTH.
	Input data contain one unmatched attribute: COLLECTOR IDENTIFICATION.
	Input data contain one unmatched attribute: DISTRICT AREA.
Medium	Input data contain two unmatched attributes: COLOR and HARVEST MONTH.
	Input data contain two unmatched attributes: COLOR and COLLECTOR IDENTIFICATION.
	Input data contain two unmatched attributes: COLOR and DISTRICT AREA.
	Input data contain two unmatched attributes: MONTH and COLLECTOR IDENTIFICATION
	Input data contain two unmatched attributes: MONTH and DISTRICT AREA
	Input data contain two unmatched attributes: DISTRICT AREA and COLLECTOR IDENTIFICATION
Low	Input data do not match existing records.

Note: C denotes as color, M denotes as month, D denotes as district area, and Q denotes as collector.

The predicted °Brix of the RM from the DSS can be used to estimate the PM °Brix that should be prepared. This is based on the relationship between °Brix and weight of the two input ingredients and that of the final product as expressed in Eq. 9.

$$B_1W_1 + B_2W_2 = B_3W_3 \tag{9}$$

where B_1, B_2, B_3 are the °Brix of the RM, PM, and the final product, respectively, and W_1, W_2, W_3 are their weights.

Based on the weight of incoming RM measured at the receiving operation, the weight of PM to be prepared, and that of the canned pineapple, the PM °Brix, B_2, can be determined from Eq. 10.

$$B_2 = \frac{B_3W_3 - B_1W_1}{W_2} = \frac{B_3W_3 - B_1W_1}{W_3 - W_1} \tag{10}$$

After the canned pineapple producer uses the developed DSS, the number of production process delay caused by PM °Brix preparation process is significantly reduced.

CHAPTER 5
DISCUSSION AND CONCLUSION

To support the laboratory technicians, as well as to reduce preparation time and cost, prediction of the °Brix is a desirable task. Accurate prediction value can guide the technicians to prepare the PM more efficiently. This study proposed three machine learning techniques to construct prediction model including ANN, SVM, and DBN as well as multiple linear regression to find that which one is the most suitable tools for making a prediction. One of the most important elements that impact the performance machine learning technique is to find an appropriate set of hyperparameter in order to obtain lower error as much as possible. So, the applied method is the grid-search which is one of the most popular tuning strategies. However, grid-search is an exhaustive technique which time-consuming if the range of value is high this is why the modeler must specified the range of value and number of increasing step from lower to upper bound due to the fact that the range of every hyperparameter has a wide range, and the experimental runs cannot be done at every possible set of hyperparameter. This may lead to the situation that the result may stuck in the local optima if the modeler does not specify the range of the boundary at the global optimal area.

Therefore, the obtained result from machine learning technique may not its best performance of the model, but since the value of their MAE is just around 1.1 and MAPE lower than 10%, this may conclude that it provides great result but may not their best result. In addition, the ensemble method is applied to improve the prediction performance, which the MAE and MAPE reduce from 1.1 to 0.98 and 8.86% to 8.27%, respectively. However, there is no significant difference between each prediction model performance. Moreover, at this level, it can help the laboratory technicians to reduce the cycle time for preparing the right amount of PM to be filled. For the further study is to develop the more powerful tuning methods for hyperparameter to reduce an error and reduce the number of experimental runs.

To make use of the results of this study, a DSS is developed to provide the user the predicted °Brix for future incoming RM based on four key inputs. The DSS also indicates the reliability level of the predicted °Brix. The database embedded in the DSS may need to be updated as additional data become available. This requires retraining and fine tuning the ML models and generating predicted °Brix for the new data. The update may be performed once a quarter, every six months, or no longer than once a year. Feedback from the user (laboratory technician) on the prediction performance and its reliability could also trigger the DSS update.

REFERENCES

Ajith, A. (2005). Artificial neural networks. Handbook of measuring system design, 901-908.

Akande, K. O., Owolabi, T. O., Twaha, S., & Olatunji, S. O. (2014). Performance comparison of SVM and ANN in predicting compressive strength of concrete. IOSR Journal of Computer Engineering, 16(5), 88-94.

Alonso, J., Castañón, Á. R., & Bahamonde, A. (2013). Support Vector Regression to predict carcass weight in beef cattle in advance of the slaughter. *Computers and electronics in agriculture*, *91*, 116-120.

Amraei, S., Mehdizadeh, S. A., & Sallary, S. (2017). Application of computer vision and support vector regression for weight prediction of live broiler chicken. Engineering in agriculture, environment and food, 10(4), 266-271.

Anderson, D., & McNeill, G. (1992). Artificial neural networks technology. Kaman Sciences Corporation, 258(6), 1-83.

Ascione, F., Bianco, N., De Stasio, C., Mauro, G. M., & Vanoli, G. P. (2017). Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. Energy, 118, 999-1017.

Aulia, M. N., Khodra, M. L., & Koesoema, A. P. (2020). Predicting Macronutrient of Baby Food using Near-infrared Spectroscopy and Deep Learning Approach. In IOP Conference Series: Materials Science and Engineering (Vol. 803, No. 1, p. 012019). IOP Publishing.

Barradas Filho, A. O., Barros, A. K. D., Labidi, S., Viegas, I. M. A., Marques, D. B., Romariz, A. R., ... & Marques, E. P. (2015). Application of artificial neural networks to predict viscosity, iodine value and induction period of biodiesel focused on the study of oxidative stability. Fuel, 145, 127-135.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Boughorbel, S., Tarel, J. P., & Boujemaa, N. (2005). Conditionally positive definite kernels for svm based image recognition. In 2005 IEEE International Conference on Multimedia and Expo (pp. 113-116). IEEE.

Cai, Y. D., Ricardo, P. W., Jen, C. H., & Chou, K. C. (2004). Application of SVM to predict membrane protein types. Journal of Theoretical Biology, 226(4), 373-376.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, 11, 2079-2107.

Chia, K. S., Rahim, H. A., & Rahim, R. A. (2012). Prediction of soluble solids content of pineapple via non-invasive low cost visible and shortwave near infrared spectroscopy and artificial neural network. Biosystems Engineering, 113(2), 158-165.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.Prokop, A., Helling, H. J., Hahn, U., Udomkaewkanjana, C., & Rehm, K. E. (2005). Biodegradable implants for pipkin fractures. Clinical Orthopaedics and Related Research, 12(32), 226-233.

da Costa, N. L., Llobodanin, L. A. G., de Lima, M. D., Castro, I. A., & Barbosa, R. (2018). Geographical recognition of Syrah wines by combining feature selection with Extreme Learning Machine. Measurement, 120, 92-99.

Djekic, I., Mujčinović, A., Nikolić, A., Jambrak, A. R., Papademas, P., Feyissa, A. H., ... & Tonda, A. (2019). Cross-European initial survey on the use of mathematical models in food industry. Journal of Food Engineering, 261, 109-116.

Escobar, C. A., & Morales-Menendez, R. (2018). Machine learning techniques for quality control in high conformance manufacturing environment. *Advances in Mechanical Engineering, 10*(2)

Ferlito, S., Atrigna, M., Graditi, G., De Vito, S., Salvato, M., Buonanno, A., & Di Francia, G. (2015). Predictive models for building's energy consumption: An Artificial Neural Network (ANN) approach. In *2015 xviii aisem annual conference* (pp. 1-4). IEEE.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. Parallel distributed processing: Explorations in the microstructure of cognition, 1(282-317), 2.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

Jin, L., Kuang, X., Huang, H., Qin, Z., & Wang, Y. (2005). Study on the overfitting of the artificial neural network forecasting model. ACTA METEOROLOGICA SINICA-ENGLISH EDITION-, 19(2), 216.

Karhunen, J., Raiko, T., & Cho, K. (2015). Unsupervised deep learning: A short review. Advances in Independent Component Analysis and Learning Machines, 125-142.

Kashwan, K. R., & Velu, C. M. (2013). Customer segmentation using clustering and data mining techniques. International Journal of Computer Theory and Engineering, 5(6), 856.

Kerdpiboon, S., Kerr, W. L., & Devahastin, S. (2006). Neural network prediction of physical property changes of dried carrot as a function of fractal dimension and moisture content. Food research international, 39(10), 1110-1118.

Kuremoto, T., Kimura, S., Kobayashi, K., & Obayashi, M. (2014). Time series forecasting using a deep belief network with restricted Boltzmann machines. Neurocomputing, 137, 47-56.

Längkvist, M., Coradeschi, S., Loutfi, A., & Rayappan, J. B. B. (2013). Fast classification of meat spoilage markers using nanostructured ZnO thin films and unsupervised feature learning. Sensors, 13(2), 1578-1592.

Lashgari, M., Maleki, A., & Amiriparian, J. (2017). Application of acoustic impulse response in discrimination of apple storage time using neural network. International Food Research Journal, 24(3).

Lawrence, S., Giles, C. L., & Tsoi, A. C. (1997). Lessons in neural network training: Overfitting may be harder than expected. In AAAI/IAAI (pp. 540-545).

Lease, M. (2011). On quality control and machine learning in crowdsourcing. Human Computation, 11(11).

Li, H., Xu, Q., He, Y., Fan, X., & Li, S. (2020). Modeling and predicting reservoir landslide displacement with deep belief network and EWMA control charts: a case study in Three Gorges Reservoir. Landslides, 17(3), 693-707.

Lin, S. W., Ying, K. C., Chen, S. C., & Lee, Z. J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert systems with applications, 35(4), 1817-1824.

Lin, Y., Guo, H., & Hu, J. (2013). An SVM-based approach for stock market trend prediction. In The 2013 international joint conference on neural networks (IJCNN) (pp. 1-7). IEEE.

Meng, Y., Lan, Q., Qin, J., Yu, S., Pang, H., & Zheng, K. (2019). Data-driven soft sensor modeling based on twin support vector regression for cane sugar crystallization. Journal of food engineering, 241, 159-165.

Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert systems with applications, 28(4), 603-614.

Mohan, P., & Patil, K. K. (2017). Crop production rate estimation using parallel layer regression with deep belief network. In 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT) (pp. 168-173). IEEE.

Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621-633.

Mukarev, M. I., & Walsh, K. B. (2012). Prediction of Brix values of intact peaches with least squares-support vector machine regression models. Journal of Near Infrared Spectroscopy, 20(6), 647-655.

Piotrowski, A. P., & Napiorkowski, J. J. (2013). A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. Journal of Hydrology, 476, 97-111.

Rai, P., Majumdar, G. C., DasGupta, S., & De, S. (2005). Prediction of the viscosity of clarified fruit juice using artificial neural network: a combined effect of concentration and temperature. Journal of Food Engineering, 68(4), 527-533.

Sanaeifar, A., Bakhshipour, A., & de la Guardia, M. (2016). Prediction of banana quality indices from color features using support vector regression. Talanta, 148, 54-61.

San-Payo, G., Ferreira, J. C., Santos, P., & Martins, A. L. (2019). Machine learning for quality control system. Journal of Ambient Intelligence and Humanized Computing, 1-10.

Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. Visualization in Engineering, 6(1), 1-20.

Shao, H., Jiang, H., Zhang, X., & Niu, M. (2015). Rolling bearing fault diagnosis using an optimization deep belief network. Measurement Science and Technology, 26(11), 115002.

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. Expert systems with applications, 28(1), 127-135.

Singla, R., Chambayil, B., Khosla, A., & Santosh, J. (2011). Comparison of SVM and ANN for classification of eye events in EEG. Journal of Biomedical Science and Engineering, 4(1), 62.

Sucipto, S., Niami, M. W., Hendrawan, Y., Al-Riza, D. F., Yuliatun, S., Supriyanto, S., & Somantri, A. S. (2018). Prediction of water content, sucrose and invert sugar of sugarcane using bioelectrical properties and artificial neural network. International Food Research Journal, 25(6).

Thailand Board of Investment (BOI)Thailand Investment Review (TIR). September 2019. Feed The World Through Innovation. Retrived from BOI Website: https://www.boi.go.th/upload/content/TIR_SEPT2019.pdf

Topuz, A. (2010). Predicting moisture content of agricultural products using artificial neural networks. Advances in Engineering Software, 41(3), 464-470.

Torkashvand, A. M., Ahmadi, A., & Nikravesh, N. L. (2017). Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network

(ANN) and multiple linear regressions (MLR). Journal of integrative agriculture, 16(7), 1634-1644.

Tsakanikas, P., Karnavas, A., Panagou, E. Z., & Nychas, G. J. (2020). A machine learning workflow for raw food spectroscopic classification in a future industry. Scientific Reports, 10(1), 1-11.

Tsen, A. Y. D., Jang, S. S., Wong, D. S. H., & Joseph, B. (1996). Predictive control of quality in batch polymerization using hybrid ANN models. AIChE Journal, 42(2), 455-465.

United States Department of Agriculture (USDA) Foreign Agricultural Service (FAS). April 2020. Thailand: Food Processing Ingredients. Retrieved from FAS Website: https://apps.fas.usda.gov/newgainapi/api/Report/DownloadReportByFileName?fileName=Food%20Processing%20Ingredients_Bangkok_Thailand_03-30-2020

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. Renewable Energy, 105, 569-582.

Wang, X., Chen, T., & Xu, H. (2020). Thickness Distribution Prediction for Tectonically Deformed Coal with a Deep Belief Network: A Case Study. Energies, 13(5), 1169.

Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., ... & Lei, M. (2019). Machine learning in materials science. InfoMat, 1(3), 338-358.

Winkler, R., & Makridakis, S. (1983). The combination of forecasts. Journal of the Royal Statistical Society. Series A (General), 150-157.

Wolpert, D.H., & Macready, W.G. (1997). No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1, 67-82.

Wu, M., & Chen, L. (2015). Image recognition based on deep learning. In 2015 Chinese Automation Congress (CAC) (pp. 542-546). IEEE.

Yeh, C. Y., Huang, C. W., & Lee, S. J. (2011). A multiple-kernel support vector regression approach for stock market price forecasting. Expert Systems with Applications, 38(3), 2177-2186.

Zenoozian, M. S., Devahastin, S., Razavi, M. A., Shahidi, F., & Poreza, H. R. (2007). Use of artificial neural network and image analysis to predict physical properties of osmotically dehydrated pumpkin. Drying Technology, 26(1), 132-144.

Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. European journal of operational research, 116(1), 16-32.

BIOGRAPHY

Name	Mr. Kasidit Singthong
Date of Birth	June 6, 1997
Education	2020: Bachelor of Engineering (Industrial Engineering) Sirindhorn International Institute of Technology, Thammasat University