A Comparison of Machine Learning Algorithms and their Applications

Jiraporn Charoenpong [1, 2, 3], Busayamas Pimpunchat [2 ,3], Somkid Amornsamankul [4, 5], Wannapong Triampo* [1 , 5]
Narin Nuttavut [1 , 5]

[1] The Centre of Excellence in Physics, CHE, 328 Si Ayuttaya Road, Bangkok, Thailand.
[2] *Department of Mathematics*, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10250, Thailand.
[3] The Centre of Excellence in Mathematics, CHE, Bangkok, 10400, Thailand.
[4] *Department of Mathematics*, Faculty of Science, Mahidol University, Bangkok 10400, Thailand.
[5] *Department of Physics*, Faculty of Science, Mahidol University, Bangkok 10400, Thailand.

email: Jira.f@outlook.com; Busayamas.pi@kmitl.ac.th; Somkid.amo@mahidol.ac.th; Wannapong.tri@mahidol.ac.th; Narin.nut@mahidol.ac.th

*Corresponding Author

***Abstract* - Machine Learning is one of Data mining. Machine Learning has recently been applied in many areas such as a disease like cancer, education, agriculture, environment, business, etc. In this work, we review the algorithms of frequently used Machine Learning for previously mentioned applications. The various software used for machine learning is analyzed and compared. The software includes freeware, open-source software and non-free software such as TensorFlow, Weka, RapidMiner, R Programming, etc. The benefit of most software is to shorten the time of creating the Machine Learning Model and some software is used as more friendly user design tools. Specific examples of Machine Learning applications to demonstrate the benefits of Machine Learning used in each research is presented. Comparison of algorithms and the advantages and disadvantages of each algorithm is analyzed.**

Keywords - Data mining; Machine Learning Algorithms; Application; Application Software

I. INTRODUCTION

Data mining is often used in large data to find interesting patterns in that data. The data sources can include databases, data warehouses, database online, other information repositories [1]. There are more and more people in the world, so the information is large. Data mining can handle large data problems by means of statistical methods or machine learning.

Machine learning is one of Data mining. Its work explores the study and creation of algorithms that can learn and predict data. The use of machine learning could reduce the complexity of the study, and it also helps to understand more about the analysis [2]. The prediction accuracy by using machine learning could obtain high accuracy [3] and be comfortable for analyzing large databases. There are more than 10 types of Machine learning as known [4] such as Decision Tree, Support Vector Machine. Each algorithm will have different data classification. So, the using of a different algorithm, the accuracy will be different. Software for Machine learning has both open and closed sources. Weka is one of open source software, and most of the research is done using such software to Machine learning for the convenience of work. Presently Machine Learning to a wide range of applications, whether medical, industrial and financial applications. There are many research studies on medical application such as prognosis disease or cancer. For example, the use of decision tree methods, which is one of the methods of Revealing determinant factors for early breast cancer recurrence of this method is found to have a predictive accuracy of 70% [5]. The use of a machine learning to apply with cancer has many studies. Because of the number of cancer patients is increasing, so the data of medical is increased as well.

In this review, machine learning algorithms (MLAs) for prediction are studied. The MLA of this review is Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), Artificial Neural Network (ANN), and Random Forest (RF). The algorithm to use for the application of Machine Learning including the advantage and limitation of each algorithm are presented.

A. Machine Learning Algorithms

A common framework of machine learning is to apply a group of modeling techniques or algorithms that can learn from data and make a judgement without human interference. For Big data problems, machine learning equips a scalable and modular strategy for data analysis [6]. Mainly Machine Learning can be classified into three main types; supervised learning, unsupervised learning, and reinforcement learning [7]. In supervised learning, the

system must learn inductively a function called a target function, which is an expression of a model describing the data. In supervised learning, there are two kinds of learning tasks: classification and regression. Classification models try to predict distinct classes, while regression models predict numerical values. Some of the most common techniques are Decision Trees, Rule Learning, and Instance-Based Learning. In unsupervised learning, the system tries to discover the hidden structure of data or between variables. In that case, training data consists of instances without any corresponding labels. Reinforcement Learning is the system attempts to learn through direct interaction with the environment. Reinforcement learning is mainly applied to autonomous systems, due to its independence in relation to its environment [8].

The method used widely of machine learning algorithms to use in cancer diagnostics is discussed in the following subsection:

1. Logistic Regression

Logistic regression is a widely used method and used to develop a regression model [9]. This method is used to analyze the connection between a single predictor, or several predictors [10]. And this method can also be used for both binary and multiclass classification. The principle of classification is the probability of occurrence to fitting data with logistics function. The value of the probability will be chosen by the logistic function, which has a value of 0 and 1 [11]. The result of logistic regression can be used to create a prediction model [12].

2. Decision Tree

Decision Tree is a tree similar hierarchical structure that composes of branches and three types of the root node, internal node and leaf node respectively that correspond to the sequence of decision rules [12]. Fig.1 shows the component of a decision tree. There are many types of Decision Trees. The dissimilarity between them is the mathematical model that is used in choosing the splitting attribute in extracting the Decision Tree rules. The research tests the three most generally used types: Information Gain, Gini Index, and Gain Ratio. Different decision tree algorithms are ready to use to classify the data, such as Gini index in CART, information gain in ID3 and gain ratio C4.5 [13].

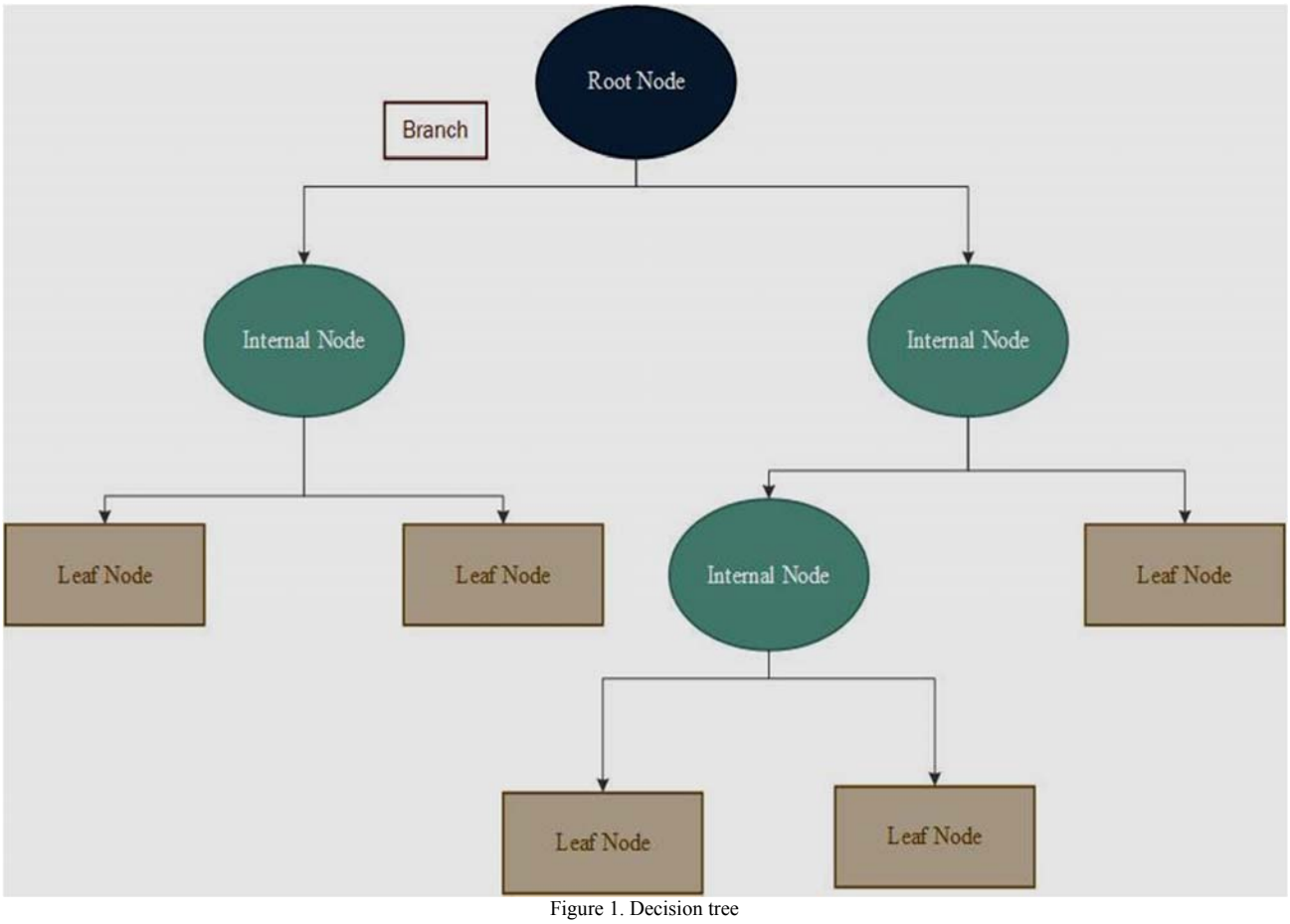

Figure 1. Decision tree

3. Support Vector Machine (SVM)

This method is based on statistical theory and structural risk minimization. The principle of this method is to providing training data as an N-dimensional vector space. SVM creates a hyperplane for separating the training data into different types. Classification of training data in this method use hyperplane, which are linear and maximum margin. Margin is defined as the sum of the distance of the hyperplane to the nearest training data. SVM will choose a hyperplane that provides the highest margin. [11,14,15].

4. Naïve Bayes (NB)

This method is one of the most commonly used and a statistical pattern recognition method. This method is based on the assumption that the probability of each attribute does not depend on other attributes. In other words, the probability of occurrence of various events is independent. The probability can be calculated from the probability of each class or group which is the highest probability will be chosen [11]. The output of Naïve Bayes is a probability which is based on probability Bayes 'rule [16]. The NB method is considered suitable for large data [17].

5. Artificial Neural Network (ANN)

ANN is similar to the black box which related to an output set with input set through artificial neural units. ANN is used to approximate the function [18]. Fig 2 illustrates each circular node as an artificial neuron. Layers are made up of a number of interconnected nodes similar to the neural network in animals.

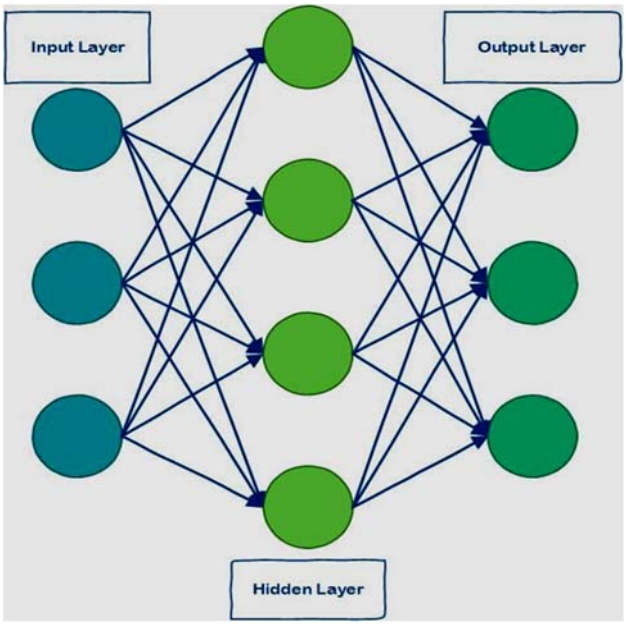

Figure 2. Artificial Neural Network

ANN are shown the pattern of the network via the input layer, which communicates to one or more hidden layers. The hidden layer is linked to the output layer. An arrow is represented a connection from the input layer to output layer. It can model highly non-linear systems in which the connection among the variables is unknown or very complicated [19]. This method will imitate the functionality of the neurons of the human brain [20]. It uses trillions of neurons to change information through electrical pulses [21].

6. Random Forest

Random Forest is an ensemble learning method that merges tree predictor. This technique implicates the generation of an ensemble of trees that vote for the most popular class. Random Forest has two distinctive characteristics; firstly, the induction error converges as the number of trees in the forest increases and the last one, the technique does not endure from overfitting [22, 23].

From Table I summaries the overview of the advantages and disadvantages of each of the algorithm, which is discussed in this review. Some algorithms may be complicated but yield an accurate result. While some algorithms are easy to implement but there are limitations or simple mistakes. For example, the decision to play tennis with weather conditions decision factors including outlook, humidity, and raining. The decision diagram is shown in Figure 3. From Figure 3, it can be explained that the outlook is overcast, the result showing that tennis can be played. But if the outlook is sunny, we must consider the next factor: humidity. If humidity is normal then, the result showing that tennis cannot be played. The results of the Decision Tree are not complicated and can be easily understood. At the same time, it is also quick to make a decision. But there are disadvantages in a case when there is a lot of information, the factor to make a decision increase substantially. Such problems result in an error in the training set can mislead to the wrong final decisions [25, 27, 28].

Logistic Regression is a model created with data in 2 classes or called binary classification. For example, the diagnosis of breast cancer by the model will predict the chance of breast cancer. There are 2 classes of malignant and benign. The results by using logistic regression will show the probability of a cancer chance has a value between 0 and 1. The results of this method are easy to understand because the results are not complicated and are straightforward. Logistic regression is a model that is suitable for uncomplicated data and the data must not be too large. Because if the data has too large, it will cause complexity and the parameter estimation procedure of logistic regression can lead to inaccurate estimates of parameters [23,24]. Support Vector Machine is one currently of the most popular methods. There are many

advantages to this method, whether it is overfitting less than other methods, and it can also handle large data as well.

But it has disadvantages as well because the support vector machine is suitable for large data. Data training is slow, and its methods are complicated. So, it may be difficult to understand the structure of the algorithm [25, 26, 28]. Naive Bays has results and simple algorithms that are easy to understand for those who start studying in this field. Since Naive Bays uses calculations about the probability that the calculations aren't complex. But there are also disadvantages are attributes or variables must be statistically independent. The limitations of this method are found that the classes must be mutually exclusive that means there is no way that any of the data could fall into 2 different classes at once [25-28]. Artificial Neural Network has a basic idea about the bioelectric network in the brain. The working principle of this method is quite complex, so it may be difficult to understand in the algorithm. The artificial neural network has many advantages, such as its complicated algorithm, which can handle noise and missing data problems, and it can also handle very large data. Since the workflow of this method is quite complex, it may be difficult to understand in the algorithm, and if the data is very large, this method will take a lot of time to calculate as well [25-28]. Random Forest is a randomly selected decision tree several trees. And bring the results of each tree together. By randomly using the Bootstrap method. The advantages of this method are not difficult to understand in the process of working. Because this method has a lot of random trees If the data is large, it will make the time to create a model increase as well [23, 24].

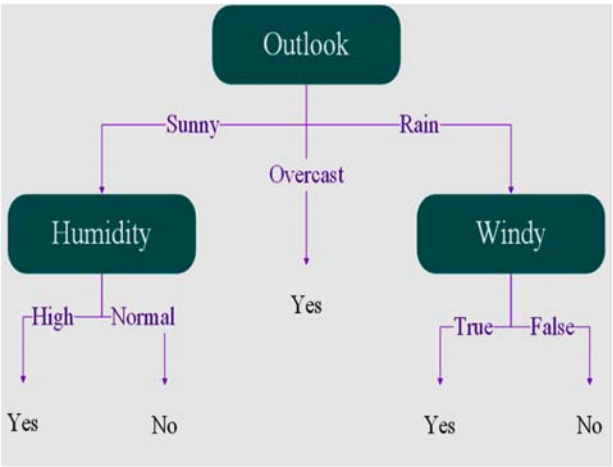

Figure 3. The example of the decision tree.

II. PERFORMANCE EVALUATION OF MACHINE LEARNING

Effectiveness of Machine Learning is important for deployment. Evaluation of Model is an assessment of machine learning algorithms that are suitable for data. If the efficiency of the model is high, it may mean that the method chosen is suitable for the data. The most common method of measuring performance is matrix confusion algorithm because it has uncomplicated procedures and easy to understand results.

TABLE I. THE ADVANTAGE AND LIMITATION OF MACHINE LEARNING ALGORITHM OF THIS REVIEW

Method	Advantages	Disadvantages
Logistic Regression (LR). [23,24]	The results of this method are easy to understand and it can be practically useful in establishing the relationships between other events.	Logistic regression can accept a large number of independent variables. Therefore, the parameter estimation procedure of logistic regression can lead to inaccurate estimates of parameters. Because there is data too large.
Support Vector Machine (SVM). [25,26,28]	Durable to noise. Easy to command. It can be used to model non-linear. Overfitting is unlikely to occur.	Training is slow. Difficult to determine optimal parameters when training data is Difficult to understand structure of algorithm.
Naïve Bays (NB). [25-28]	Not difficult to understand. No effect of order on training. It take less computational time.	Classes must be mutually exclusive (non-overlapping). Attributes or variables must be statistically independent. Less accurate because a attribute and a class frequencies affect accuracy.
Decision Tree (DT) [25,27,28]	Not difficult to understand. Fast to forecast. No effect of order on training.	Classes must be mutually exclusive (non-overlapping). Error in the training set can lead to wrong the result final decisions.
Artificial Neural Network (ANN) [25-28]	Tolerant to noise and missing data. Instances can be classified by more than one output.	Requires a lot of computer power because sample size must be large. So, it's time-consuming. Difficult to understand structure of algorithm.
Random Forest (RF) [23,24]	Not difficult to understand and It provides accurate predictions for many types of applications. Parameters can be set easily therefore,eliminates the need for pruning the trees.	For data including categorical variables with different number of levels, random forests are biased in favor of those attributes

The method is described as follows.

Confusion Matrix

A confusion matrix of binary classification is a two by two table formed by counting the number of the four outcomes of a binary classifier. There are two classes, which call positive and negative, the confusion matrix consists of four cells, which can be labeled TP, FP, FN, TN as in Table 2 [29].

TABLE II. TRUE AND FALSE POSITIVES AND NEGATIVES

Confusion Matrix		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FP
	Negative	FN	TN

Where TP (True Positive): the number of positive instances that are classified as positive. TN (True Negative): the number of negative instances that are classified as negative. FP (False Positive): the number of negative instances that are classified as positive. FN (False Negative): the number of positive instances that are classified as positive. The prediction accuracy can be obtained from this matrix as follows [30]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

Sensitivity or recall is the number of correctly predicted positive samples among all real positive samples. Similarly, specificity measures the capability of a predictor in negative samples [31,32]. The measure of sensitivity and specificity are following:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$$

To provide more insight. So, we give an example [33]: Examples of breast cancer diagnoses, which have two alternatives, are malignant and benign. This data contains 699 patients consist of benign 458 and malignant 241 patients. From the prediction, the model can be evaluated using a confusion matrix, as shown in Table III.

TABLE III. EXAMPLE OF A CONFUSION MATRIX

Confusion Matrix		Predicted Class		
		Positive	Negative	Total
Actual Class	Positive	438 (TP)	20 (FP)	458
	Negative	18 (FN)	223 (TN)	241
Total		456	243	699

From Table III, it was found that 458 patients correct predicted that benign and wrong predicted of 20 patients (the patients are benign, but the model is predicted to be malignant). And, it was found that 223 patients correct predicted that malignant and wrong predicted of 18 patients (the patients are malignant, but the model is predicted to be benign). The model can be evaluated by using the accuracy value which is approximately 94%.

III. MACHINE LEARNING ALGORITHMS APPLICATIONS

Machine Learning can be applied in many things such as industry, agriculture, medical, etc. Machine Learning used for the analysis of the importance of clinical parameters and their combinations for prognosis, e.g. prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning and support, and for the overall patient management [34]. In the past many years, there is a lot of research to use Machine Learning Algorithm to apply in cancer. In the survival analysis for cancer. This method is used to find the survival rate of cancer patients. Application of Machine Learning to Survival will help predict survival rates. This application can be developed for survival prediction in rare pathologies and have the potential to serve as the basis for the creation of decision support tools in the future [35].

At present, the software helps us to work more conveniently and save more time. And some types of software are also available for free. In ML, there are several types of software available for use. Weka is a tool for the initial processing of data. Advantages of Weka Firstly, Weka is an open source. Secondly, it solves the ML problem. Finally, it runs on many platforms [36]. It also allows users to select the most appropriate algorithm for the model to obtain the most accurate value [37]. Another tool that is used frequently, the R programming. R is used in machine learning for classification, regression, and clustering. There are a variety of packages in R to make ML easier to use, such as rpart [38] is a package that helps in recursive partitioning for classification, regression and survival trees. Another important package is e1701 [39]. It can be used in support vector machines, shortest path computation, bagged clustering, and Naive Bayes classifier.

A. Machine Learning Applications in Cancer Studies:

From survey research of machine learning application with cancer. The first matter to be studied is diagnostic. Comparison of the performance of different machine learning algorithms has been studied in many types of research. This research uses public information from the UCI. This data is a breast cancer test with two classes for decision making, with malignant and benign. This information is available for 699 patients, consisting of 241 of malignant and 458 of benign. The main purpose of this research is to compare and evaluate with different machine learning. By using SVM, DT (C4.5), NB and k-NN to predict the diagnosis of breast cancer and comparative

efficiency [40]. The results of such research showed that SVM gives the highest accuracy (97.13%) and the lowest error rate (2%) compared with other algorithms. It can be said that the SVM algorithm is a method that provides the most accurate of this dataset for prediction diagnosis.

Using ANN and LR to develop a model to predict liver cancer [41] is one of the interests by using 70% of the data to training and 30% for testing the model. The result shows that the accuracy of ANN better than the LR. In the case of survival and recurrence, there is interesting research as well. The one paper report is the prediction model of survival and recurrence for breast cancer by using machine learning [42]. The result of predict survival shows that the ANN algorithm obtained the best accuracy (83.6%) with using live-one-out cross-validation as a test method. Meanwhile in the prediction of recurrence of breast cancer found that the DT algorithm has achieved the best accuracy (96.57%) by using validation like that the validation of prediction of survival. The comparison of the methods of machine learning is still many in many types of research.

There is this research that uses 3 machine learning algorithms (DT (C4.5), SVM, ANN) for medical data analysis which is a big data to find hidden knowledge. The data for this study is the data of breast cancer patients admitted to the Iran Center program, which has a total of 1189 patients. The results of this study showed that SVM predicted breast cancer models better than other methods with the highest correctness of 95.7% and DT with the lowest predictive value of 93.6% [43]. Next research is also a study of breast cancer. This research uses 7 machine learning algorithms consisting of NB, Trees Random Forest, 1-Nearest Neighbor, AdaBoost, SVM, RBF Network and Multilayer Perceptron. The objective of this research is to apply the seven methods of machine learning to predict the model of survival of breast cancer. The data from the study of Cancer Registry Organization of Kerman, Iran with 900 breast cancer patients, consisting of 803 living patients and 97 patients died. The results of this study showed that the Trees Random Forest predicted the best breast cancer survival model with 96% accuracy. And 1-Nearest Neighbor predicted the model to be less accurate than other methods, 91% of accuracy [44].

B. Machine Learning Applications in Education

At present, educational problems have many problems. Therefore, researchers use machine learning to play a role in solving problems. In the study, machine learning is used in many research areas, such as the following researches. This research studies about student archetypes in an online course. The researcher conducted a questionnaire on Google Forms which consisted of 69 questions with 632 students to answer the questionnaire. The tool used in this research is the decision tree. The results of the research found that students studying with a higher than bachelor's degree will study through e-learning rather than undergraduates, men tend to study twice as much as women [45]. In many schools, the school has set the score criteria and can be used to indicate student performance.

There is research on the use of Machine Learning to predict student performance. This research studied the predictions of student performance by using various methods of machine learning. For predicting student performance, it is found that the ANN method provides the most accurate value, followed by Decision Tree and the least accurate method is NB. This research suggests that the factor that affects the accuracy of predictions is influence from the attributes of information. The benefits of the research study are to develop and improve the learning process and teaching to achieve good performance for students [46]. Education is also important in improving student performance because of the problem of studying with students may also be reduced.

Automated Student Advisory:

There is a research study on the Automated Student Advisory to introduce first-year students for the appropriate educational landscape. This research is to help improve student performance and quality by reducing the failure rate of first-year students. Machine learning algorithms selected by this research are Decision tree (c4.5), k-Means clustering, by using Decision Tree for predicting the department that is suitable for students and k-Means clustering for division a group of students. The results of the prediction showed that the accuracy for the prediction using decision tree C4.5 was 98.025% [47]. One of the problems that can be encountered frequently in many schools is the dropout of students.

There is research that studies the cause of the student's resignation using Machine Learning: the case study of a high school in Denmark. This research uses data from the MaCom Lectio study administration system, which has a total of 72,598 students, it was found that the student dropped out 17,399 people. In the process of creating the model, this research has divided 36,299 for creating models for prediction and 36,299 for testing models. Machine learning applications in research are Random forest, Decision tree (CART), Support Vector Machine, Naive Bayes. The results showed that Random forest was the most accurate predictive model, 93.5% and the least accurate was Naive Bayes, which as 85.6%. The benefits of this research may help teachers or staff in schools to find ways to prevent or reduce the number of student dropout [48].

Studies in School Drop-Out Rates:

There is another research that studies the drop-out by developing tools that can alert schools about students who are at risk of drop-out. Because student dropout is a big problem that may affect families, schools, and society. This research has created a model for prediction by using a Random Forest, which is one machine learning to develop

an advance warning system that identifies students who are at risk of dropout. The data used in the study were 165,715 high school students in South Korea. There were 2,050 students who drop out. The results of the study showed that the random forest gave 95% accuracy and the class teacher could observe the behavior change of students as much as possible [49].

At present, online learning is very popular because it allows students to access education more easily. Using Machine Learning to assess online students who have a school-record at-risk group. This research may help instructors adjust teaching methods to suit students to reduce the number of students at risk. This research divides students into two groups namely successful and failed. Those with an overall score equal to or higher than 60 are considered successful and those who score less than 60 on a certain scale are considered failed. Three machine learning algorithms used in this work are K *, Naïve Bayes and Decision Tree (C4.5). The study found that K* provides better results compared to other algorithms, with the highest accuracy being 82% [50]. Because online learning allows students to have easy access to educational contents, but teachers cannot know that students may not understand the teaching content. Student opinion is important. The instructor can edit or add content about the lesson so that the learner can understand the lesson more. This research selected to study information about the Massive Open Online Course (MOOC) to design teaching and learning that will align with student interest.

Types of Machine Learning Algorithms:

There are 5 types of machine learning algorithms that this research uses: k-Nearest Neighbors, Gradient Boosting Trees, Support Vector Machines, and Logistic Regression. The results of the research found that the ability of the instructor to contribute to the student's desire is a key factor to participate in the class. And the results of using machine Learning model showed that Gradient Boosting Trees produces the best accuracy [51].

There is a research report that evaluates the performance of schools and students, 9 countries, using the Program for International Student Assessment (PISA) test. This research has developed and applied machine learning to analyze the factors of PISA 2015 student scores in nine countries, including Australia, Canada, France, Germany, Italy, Japan, Spain, United Kingdom, and the United States. This research uses tree-based methods complement linear regression models to analyze the factors. The results found that the variables used, such as Socio-economic index, anxiety, motivation, gender is not found to be directly related to student success. And in this research, in terms of machine learning, is used to make the relevant factors more clear [52].

C. Machine Learning Applications in Industry

In manufacturing, the use of machine learning in the industry to play a role. SVM is one of the methods used in ML. The SVM method is used in research on the forecasting of electricity consumption in oil plants. This research studies the paper industry. Eucalyptus is also a raw material for paper production, especially eucalyptus globulus (Blue gum). Eucalyptus globulus gives the highest yield and the best pulp quality in pulp production. The purpose of this research is to use advanced statistical learning techniques to determine the amount of the inside-bark volume associated with standing E. globulus trees destined for pulp manufacture. This research selected support vector machines (SVM) and multilayer perceptron (MLP) to create a non-linear model. The data for this study were age, height, outside-bark volume 42 trees for E. globulus. The results of the experiment showed that the best model for estimating the internal volume of the system was the SVM [53].

ML tools are also used to predict machinery in factories, such as predicting mechanical wear in a smartphone factory to compare the values that give the most accurate by using ML 3 algorithm i.e. RF, SVM, feed-forward backpropagation. In the research, RF values are the most accurate [54]. Developing methods to improve the forecasting efficiency of the model is also found in many researches.

Neural Network with Decision Trees:

The following researches have used the Neural Network in conjunction with Decision Tree. This research has studied the use of ANN, DT (C4.5), and Neural Decision Tree, which provides training data as input to the Neural Network and output assent to decision tree for create the model. By using three methods to create a model for petroleum production prediction and compare to find the best model. Using data from the Saskatchewan industry and resources, Canada, which is the dataset on oil production consists of 320 data records. The results of the research showed that the accuracy of the classification of both Neural Decision Tree and Decision tree (C4.5) methods is lower than the ANN method. The researchers commented that the accuracy of the Neural Decision Tree predicted at low may occur from data not covering enough problems [55]. Electric energy is an important energy for human life. When the population increases, energy consumption will also increase.

Several research studies have used various methods to predict future resource usage. We give examples of research that studies the use of electricity in Turkey and uses machine learning to predict electricity usage. Using monthly electricity usage data from the Electricity Generation Company of Turkey from 1970 to 2011. This research has chosen a tool for prediction using ANN and Support Vector Regression, also known as SVM for regression, to develop

the best model for predicting electricity consumption by 2010 and 2011. The study found that in the future, Turkey will have more electricity. Therefore, Turkey is an ideal country to invest in energy. As for the results used in the prediction, it was found that the Support Vector Regression model was better predicted than the ANN model [56].

Crude Oil Price Prediction:

Oil is an important part of the global economy. This research, therefore, studied new methods for crude oil price prediction by using a new SVM-based method for time series forecasting for crude oil price prediction. This research studies the monthly spot prices of West Texas Intermediate (WTI) crude oil from January 1970 to December 2003. In this research using root mean square error to evaluate the prediction performance. The study found that support vector machines can perform very well on time series forecasting. The RMSE value of SVM is 2.1921 which is less than the Back Propagation Neural Network (BPNN) and Auto Regressive Integrated Moving Average (ARIMA) [57].

D. Machine Learning Applications in Agriculture

There are many factors in agriculture. For example, the higher population, the expanding urban community, resulting in reduced cultivation areas. Including the demand for agricultural products to convert to higher energy making traditional agriculture even more productive to meet the needs, In addition, many agricultural products have been lost since harvesting. This research has studied the harvesting of 7 types of crops, which includes rice, wheat, corn, soybean, peanuts, rapeseed, potato, all 31 regions in China. Research objectives for analyzing the grain loss in the harvest and using logistic regression to create models for prediction. This research divides the data into two parts, the first 60% for the training dataset and 40% for testing dataset. The results showed that Logistic regression provided the accuracy of the model for prediction was 86.25%. And this research using the stochastic gradient descent method to optimizes the structure parameters of the logistic regression model. It was found that the predictions of grain harvest loss were as high as 92.53%. And the results of the model show that the different harvest time, mode of transportation, threshing way on grain harvesting losses caused by the smaller, better small harvesting loss [58]. In this research, there is a development of learning models using convolutional neural networks which evolved from traditional neural networks. The purpose of this research is to especially recognize images for identifying plant diseases through simple leaf images of healthy or diseased plants. The data used for the study were 87,848 photographs of leaves of healthy and infected plants were used for training and testing. The most successful model of this research is the VGG convolutional neural networks model, with a success rate of 99.53% from the efficiency of the neural network [59].

Separation of Weeds from Sugar Beets:

Next research, the main objective of this research is to evaluate the accuracy of SVM and ANN to separate weeds and sugar beets according to the shape properties obtained from image data. The image data that is used to study is the digital image used in this project, made from the sugar beet farm of Shiraz University, based on the shape of the plant for weed detection. The results showed that the accuracy of ANN classification was 92.92%, which accurately classified weeds 92.50%. While the accuracy is higher when using SVM as a classification with an overall accuracy of 95%, with 93.33% accurate weed classification, which provides higher classification accuracy than ANN [60].

Another problem encountered is the condition of the changing world that affects crop harvesting. We find research that studies this field. By research selected in the areas of Illinois, Indiana, Iowa, Kentucky, Michigan, Minnesota, Ohio, and Wisconsin. The purpose of this work was to develop a novel approach for augmenting deep neural networks, which term semiparametric neural networks (SNNs). The study found that machine learning helps in forecasting and summarizing the effects of climate change on agriculture. And forecasting for simulating future weather conditions found that the weather affects less harsh harvest than expected [61]. Soil pollution also affects agriculture as well. The heavy metal is the main cause of soil contamination. Pollution in the soil affects plants, making plants grow slowly, fewer plants and dying.

Composting from Sewage Sludge:

There is a work that studies about predicting the effects of composting from sewage sludge under various circumstances and predicting soil and plant pollution, including pollution and quality of grain. This research has created various regression models to predict important variables for soil fertility and accumulation of heavy metals in soil, roots, and grains. This research shows that Decision trees are tools that provide highly accurate results and are easy to interpret, helping to easily see the processes that occur over many years of plants and soil. By the forecasting model giving the correlation coefficients of the regression tree for the prediction between 0.7343 and 0.9749 [62].

E. Machine Learning Applications in Environment Studies

The rapid increase of the population affects many aspects. In the environment, it is affected by pollution in the air. Particulate Matter with aerodynamic diameter ≤ 2.5 μm) or called PM2.5 can penetrate into the gas exchange area of the circulatory system. In China, PM2.5 is classified as a risk factor that causes death. And related to respiratory and

cancer. Research has been conducted on the use of random forests to predict PM2.5 concentrations in China. The objective is to compare the efficiency of random forests model with two traditional regression models (Generalized Additive Model and Non-Linear Exposure-Lag-Response Model). The results showed that the random forest predicted PM2.5 concentrations better than the generalized additive model and non-linear exposure-lag response model. This research may help in assessing the effects of air pollution on humans in the future [63]. This research studies about PM2.5, but it has a different purpose than previous research. The goal of this research is to develop a hybrid model that combines weather analysis and wavelet transform to improve ANN accuracy. And the goal is to predict the daily average concentration of PM2.5 in the next two days. The results of the use of a hybrid model to improve the accuracy of ANN found that the root means squared error of the hybrid model can be reduced by about 40%. When PM2.5 exceeds the specified threshold, the results showed that the hybrid model could reach 90% of the average dust detection [64]. The fine dust, also known as PM2.5, has been recognized as important air pollution that influences the health risks of the population. In this study, the efficiency of multiple linear regression algorithms, Bayesian Regularized Neural Networks, and Vector Support with Radial Basis Function Kernel, Least Absolute Shrinkage and Selection Operator, Multivariate Adaptive Regression Splines, Random forest, eXtreme Gradient Boosting, and Cubist.) This research studies data from remote sensing dataset system in British Columbia, Canada, to predict ground-level PM2.5concentrations. Based on the Cubist study, random forests and eXtreme Gradient Boosting are algorithms that perform better than other algorithms. And the most effective model is Cubist [65].

Air Pollution and Airborne Contaminants:

Hazardous air pollution or airborne contaminants means any substances that may cause additional deaths or serious illnesses or that may cause current or potentially harmful to human health. This research aims to create a regression model of air quality using Support Vector Regression (SVR) in Avilés, Spain. The data that this research has studied are nitrogen oxides (NOx), carbon monoxide (CO), sulfur dioxide (SO2), ozone (O3) and dust (PM10) for the years 2006–2008. The study found that using SVM using the PUK function to create models results in the best predictions for the air quality's problem. The benefits of this research may be to use the results of predictions to deal with and solve problems related to air pollution in the future [66]. Wastewater is water used in households and industries without treatment. Excessive wastewater results in worsening water quality and eventually becoming polluted. Wastewater that is waiting to enter the waste treatment process can cause a high concentration of total nitrogen, which will affect the quality of the water.

High Concentration of Total Nitrogen:

There is research that studies the high concentration of total nitrogen. The objective of this research is to create a learning model of two types of machines: artificial neural networks (ANNs) and support vector machines (SVMs) to predict the concentration of total nitrogen at the 1day period from a wastewater treatment plant in Ulsan, South Korea. The results showed that both tools can be used to create both models well, but SVM provides more accurate predictions than ANN, while the sensitivity of ANN shows that the model has good results of wastewater treatment [67]. Wastewater treatment has many methods such as Grease Trap, Stabilization Pond, Aerated Lagoon, etc. The cost of wastewater treatment will vary. There are researches that use machine learning to find the cost of wastewater treatment. This research presents a new method which is the basis of Machine Learning. To develop to be more effective with more complex information. And the goal of this research is to create a model to generate high-performing energy cost models for wastewater treatment plants by using machine learning. The data used in the study are 317 wastewater treatment plants in northwestern Europe. The results of this research using NN and RF to predict that the pollution load greatly contributes to the cost of energy and the cost of energy has a negligible impact on the cost of wastewater treatment. In the use of Machine Learning in this research, it was found that the newly developed method is the Machine Learning and Computation Modeling (MLMC) which gives better results than traditional techniques and provides better performance than relevant research [68].

F. Machine Learning Applications with Business:

In business, machine learning can be applied to evaluate or forecast sales. It may help companies reduce production costs [69]. Companies can find ways to increase sales by using machine learning to find the most relevant factors. By studying researches using machine learning application, we found interesting research. Currently, there are many users of the bank. Machine learning is taking part in this research. This research aims to extract interesting information from the database. Interesting data could increase the profitability of the organization by using an artificial neural network algorithm and using R which is software for data processing. The data is divided into two sets. First, the set has two types of credits like good and bad. Second, the set has two types of customers like active. There is a total of 12 inputs. The results of the study showed that the validity of the first dataset was 72% and 98% of the second dataset. [70].

Banks:

Another interesting research studies about banks. This research studies data from customer data from Chinese commercial bank, which has 50,000 records. This research

has taken data through data processing such as missing data. From data processing, there are 46,406 records for prediction. In this research, SVM and Logistic Regression algorithms are used to predict models. The results of the research found that SVM provides better logistic regression with accuracy as 98.95% [71]. There is a research report that uses 14 million records to study. Information that is studied is information about banking customer's behavior from Santander Bank This research divides data into 2 sets, with 13 million as data for training and 1 million for testing. By using two methods of machine learning, NB and SVM. The result fount that NB is more efficient than SVM. By NB giving precision, recall, f-measure to 4%, 49%, and 7.3% respectively. While SVM gives precision, recall, f-measure to 0.18%, 10%, 0.3% respectively [72].

Predictive Models:

For using Machine Learning to create predictive models. Some research uses more than one forecasting model to find the best model. This research has created a binary classifiers model using the machine and deep learning for the predicting loan default probability for model comparisons. This research has created 7 models, including 4 logistic regression models, The random forest model, 4 gradient boosting model and deep learning models. The criteria used to measure the performance of models are AUC and RMSE. The study found that the random forest model, the gradient boosting model, gave a high AUC value, but the gradient boosting model was the model that gave the highest AUC and lowest RMSE than 6 models [73]. Machine Learning is suitable for solving big data problems. At present, there are a number of people who are interested in insurance, which makes the insurance business a greater number of customers. The use of information to machine learning is another method that is often used. This research applied Random Forest Regression to forecast customer profitability. The objective is to compare the efficiency of the Random Forest (RF) with other methods (Linear Regression (LR), Decision tree (DT), Support Vector Machine (SVM) and Generalized Boosting Model (GBM)). The information used in the study will be considered. Region, age, insurance status, gender, and customer source Is the most important factor in predicting the profits of insurance companies The results of the RF model comparison with other models found that RF gave the least RMSE value of 653.27 and Linear Regression, giving the highest RMSE value of 3122.39, while considering the R-squares, RF is the highest value. 99.01% and minimal, Linear Regression is 77.56% [74].

FOREX Trading:

There is a research article that brings Machine Learning Algorithms to compare to find the best model, such as trading through a simulation set of trading in the FOREX market. The research is conducted using 6 machine learning algorithms to find a model that shows the highest trading profits [75]. In the sales business, There is research on employees who make good sales, will be analyzed to the character and various factors. Such analysis will be used to apply for jobs in order to find people with similar characteristics with employees who perform well. The purpose of this research is to study about selecting employees to improve sales efficiency by using Machine Learning. Machine Learning Algorithm used in research includes Support vector machines, Decision trees, Logistic regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis. The results of the study showed that the support vector machines approach was able to develop the accuracy of more than 10% and give the highest accuracy at 82.61%. In addition, the Support vector machines use only three out of ten features include agreeableness, openness to experience and sociability [76].

Finance:

This research is about finance. This research uses machine learning techniques to sample credit card data at large account levels and compare models for predicting bank risk with 3 methods of machine learning, including Decision Tree, Random Forest, Logistic Regression. The data used in the study is that consumer data from the United States large credit bureau, which includes more than 500 million individual account records, was recorded over a period of 6 years. For creating a forecast model to separate the accounts into two categories: bad account and good account. The results of the study of machine learning showed that decision trees and random forests gave better results than logistic regression. And logistic regression incorrectly predicts the delay of payment of credit card fees [77]. Housing appraisal is one of the most important trading decisions that affect national real estate policies. This study uses machine learning algorithms as a research method to develop a home price prediction model and improve the accuracy of predictions more efficiently. The data used in this research is the residence of 5359 townhouses in Fairfax County, Virginia, collected by the Multiple Listing Service. Results in terms of forecasting with Machine Learning Algorithm found that RIPPER's performance is better than C4.5, Naïve Bayesian and AdaBoost. The RIPPER model is most effective in predicting house prices in Fairfax County, Virginia. Therefore, this study shows that the use of Machine Learning in forecasting is accurate and reliable [78]. The machine learning applications survey as shown in Table IV-1 to IV-4.

TABLE IV-1. SURVEY RESEARCH ON MACHINE LEARNING APPLICATIONS

Author	Year	Paper	Machine Learning Algorithms	Size of data	Evaluation	Best Algorithm	Software
Ahmed et.al	2013	Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence	C4.5,ANN,SVM	1189	95.7	SVM	Weka
Cirkovic et.al.	2015	Prediction models for estimation of survival rate and relapse for breast cancer patients	Naïve Bayes, DT,SVM,logistic regression,ANN	-	83.6, 96.67	ANN, DT	Weka
Asri et.al.	2016	Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis	(SVM), Decision Tree (C4.5), Naive Bayes and (k-NN)	699	97.13	SVM	Weka
Rau et.al.	2016	Development of a web-based liver cancer prediction model for type II diabetes patients by using an ANN	ANN, Logistic Regression	2060	87.3	ANN	Weka and Statistica
Montazeri et.al.	2016	Machine learning models in breast cancer survival prediction	NB , Trees Random Forest, 1-Nearest Neighbor, AdaBoost, SVM, RBF Network (RBFN), and Multilayer Perceptron (MLP)	900	96	Trees Random Forest,	Weka
Amirah et.al	2015	A Review on Predicting Student's Performance using Data Mining Techniques	Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine.	-	98	Neural Network	-
Aly et.al	2013	Automated Student Advisory using Machine Learning	Decision tree(C4.5), k-Means Clustering	-	98.028	Decision tree (C4.5)	-
Sara et.al	2018	Student and school performance across countries: A machine learning approach	Multilevel Regression Trees	-	-	-	-
Jae et.al	2015	High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study	Random forest, CART, SVM, Naive Bayes	72,598	93.5	Random Forest	Weka
Erkan et.al	2019	Dropout early warning systems for high school students using machine learning	Random Forests	165,715	95	Random Forest	R programming

TABLE IV-2. SURVEY RESEARCH ON MACHINE LEARNING APPLICATIONS

Author	Year	Paper	Machine Learning Algorithms	Size of data	Evaluation	Best Algorithm	Software
Khe et.al.	2012	Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with Information System(IS) 100	K-Star, Naïve Bayes and Decision Tree (C4.5)	625	82%	K-star (K*)	-
Chiara et.al	2018	Understanding Student Engagement in Large-Scale Open Online Courses: A Machine Learning Facilitated Analysis of Student's Reflections in 18 Highly Rated Massive Open Online Courses	k-Nearest Neighbors, Gradient Boosting Trees, Support Vector Machines, Logistic Regression, and Naïve Bayesian	24,612	-	Gradient Boosting Trees	Python programming
Chan et.al.	2013	Application of the Neural Decision Tree approach for prediction of petroleum production	Artificial Neural Network (ANN), Decision Tree	320	76.25	Artificial Neural Network	Waka
Gamze et.al	2012	Forecasting Electricity Consumption with Neural Networks and Support Vector Regression	Artificial Neural Networks and Support Vector Regression	504	-	Support Vector Regression	-
Nieto et.al.	2012	Support vector machines and neural networks used to evaluate paper manufactured using Eucalyptus globulus	support vector machines (SVM) and multilayer perceptron (MLP)	142	R-Squared is 97.26%	Support Vector Machines	
PetkoviC et.al.	2009	Electrical Energy Consumption Forecasting in Oil Refining Industry Using Support Vector Machines and Particle Swarm Optimization	Support Vector Machines (SVMs) and Particle Swarm Optimization (PSO)	-	-	Support Vector Machines	-
Xie et.al.	2006	A New Method for Crude Oil Price Forecasting Based on Support Vector Machines	Support Vector Machines	408	-	Support Vector Machines	-
Wu et.al.	2017	A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests	ANN, SVM for regreesion, RFs	-	-	Random Forests	-

TABLE IV-3. SURVEY RESEARCH ON MACHINE LEARNING APPLICATIONS

Author	Year	Paper	Machine Learning Algorithms	Size of data	Evaluation	Best Algorithm	Software
Huang et.al	2017	Analysis of the grain loss in harvest based on logistic regression	Logistic Regrssion	5400	86.25%	Logistic Regrssion	-
Ferentinos et.al	2018	Deep learning models for plant disease detection and diagnosis	Convolutional Neural Networks (CNNs)	87,848	99.53	Convolutional Neural Networks (CNNs)	-
Bakhshipour et.al.	2018	Evaluation of support vector machine and artificial neural networks in weed detection using shape features	SVM,ANN	600	92.92	ANN	-
Crane et.al	2018	Machine learning methods for crop yield prediction and climate change impact assessment in agriculture	Semiparametric Neural Networks	-	-	Semiparametric Neural Networks	R programming
Perex et.al	2017	Decision Trees for the prediction of environmental and agronomic effects of the use of Compost of Sewage Slugde	Decision Tree	-	97.49	Decision Tree	Weka
Chen et.al.	2018	A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information	Random Forests	-	36-90	Random Forests	R programming
Feng et.al	2015	Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation	Artificial Neural Network (ANN)	-	-	Artificial Neural Network (ANN)	-
Xu et.al.	2018	Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5	multiple linear regression, Bayesian Regularized Neural Networks, Support Vector Machines with Radial Basis Function Kernel, Least Absolute Shrinkage and Selection Operator, Multivariate Adaptive Regression Splines, Random forest, eXtreme Gradient Boosting, and Cubist.	-	-	Cubist	-

TABLE IV-4. SURVEY RESEARCH ON MACHINE LEARNING APPLICATIONS

Author	Year	Paper	Machine Learning Algorithms	Size of data	Evaluation	Best Algorithm	Software
Suarez et.al	2011	Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain)	Support Vector Regression	36 smonth	-	Support Vector Regression	-
Guo et.al	2018	Machine learning for energy cost modelling in wastewater treatment plants	Random Forest, Neural Network, Machine Learning and Computation Modeling (MLMC)	-	-	MLMC	R programming
Torregrossa et.al	2015	Prediction of effluent concentration in a wastewater treatment plant using machine learning models	Support Vector Machine, Artificial Neural Network	-	-	Artificial Neural Network	-
Addo et.al	2018	Credit Risk Analysis Using Machine and Deep Learning Models	Elastic Net (extension from linear regrssion), Random Forest, Deep Learning, The Gradient Boosting	117,019	-	The Gradient Boosting	R programming
Fang et.al.	2016	Customer profitability forecasting using Big Data analytics: A case study of the insurance industry	Linear Regression, Decision Tree, SVM ,generalized boosted model, RF	-	99.01	Random Forest	-
Gerlein et.al	2016	Evaluating machine learning classification for financial trading: an empirical approach	The naïve Bayes , The K*, Decision Tree (C4.5), Logistic Model Tree	7,701	52.84	Logistic Model Tree	Weka
Delgado et.al	2011	Improving sale performance prediction using support vector machines	Support vector machines, Decision trees, Logistic regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis	-	82.61	Support vector machines	-
Butaru et.al	2016	Risk and risk management in the credit card industry	Decision Tree, Random Forest, Logistic Regression	500 million	-	Decision Tree, Random Forest	Weka
Park et.al	2015	Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data	C4.5, RIPPER (Repeated Incremental Pruning to Produce Error Reduction), Naïve Bayesian, and AdaBoost (Adaptive Boosting)	5,359	99.75	RIPPER	Weka

IV. DISCUSSION AND CONCLUSION

Based on the research study on the use of MLAs, each research has chosen a different algorithm. We have summarized the similarities and differences of the research studied in this review, as shown in Table 5.

TABLE V. THE SIMILARITIES AND DIFFERENCES OF MACHINE LEARNING

Similarities	Differences
Big Data The research shows that the data is large. The ML can handle large data. So, the problem with ML is usually applied to big data.	Algorithm and Validation From the research study, we found that the Algorithm and validation are selected based on the suitability of the information.
Evaluation model The research is often used to evaluate the efficiency of the model with a confusion matrix. The model chosen is the most accurate accuracy.	Attributes, Parameters Data that has too many attributes or parameters will affect the model. Each algorithm gives different results.

Table V describes the similarity in research that the use of machine learning for applications is often applied to Big Data. Because machine learning can deal with the problems that occur in Big Data, whether it is dealing with noise problems in data such as missing data etc. Another similarity is the evaluation of the model. Most researches often choose to use a confusion matrix for evaluating the performance of the model because the confusion matrix has simple steps. And the results are easy to understand. From the study, it is found that the algorithms selected for each research use different methods, which is based on the most accurate. The method that provides the most accurate is often selected for use in that research. For some data, there are many attributes. Some methods of machine learning will cut the least important attribute. This means that if we use the same data but different methods. The result may be given the number of attributes or parameters unequal.

Algorithms in many research use information in a variety of ways to find the most accurate models. The study found that most research uses the SVM algorithm because it is easy to use and durable to noise. And DT because it is easy to use, and the results can be easily understood. Both algorithms provide higher accuracy than other algorithms. The software used in this research is WEKA. It popular because it is easy to use, and the results of the model are easy to understand for the user.

This review is a survey of research on the application of MLAs. The number and type of information could affect the accuracy in prediction. By using software to help model Machine Learning can save time. Certain software also allows users to develop models of models. Therefore, the using of many algorithms in software to check accuracy is important because it will help compare how to know the appropriate algorithm with the data.

ACKNOWLEDGMENT

This research project was funded by the Thailand Center of Excellence in Physics (ThEP) and the Centre of Excellence in Mathematics (CEM), Faculty of Science, Mahidol University.

REFERENCES

[1] J. Han, M. Kamber, J. Pei, Data mining, Concepts, and Techniques 3rd ed., USA, 2011

[2] Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V., Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, 350–361., 2017.

[3] Zheng, B., Yoon, S. W., & Lam, S. S., Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476–1482., 2014.

[4] Das K., Behera N. R., A Survey on Machine Learning: Concept, Algorithms and Applications, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, 2017.

[5] B. Ashok, Dr. P. Aruna, Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier, Journal of Engineering Research and Applications, 2016.

[6] Ip, R. H. L., Ang, L.-M., Seng, K. P., Broster, J. C., & Pratley, J. E. (2018). Big data and machine learning for crop protection. Computers and Electronics in Agriculture, 151, 376–383., 2018.

[7] Russell S., Norvig p., Artificial Intelligence: A Modern Approach, 3rd ed., Prentice Hall, Upper Saddle River, New Jersey, USA, 2010.

[8] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I., Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116., 2017.

[9] Zekić-Sušac M., Šarlija N, Has A., Bilandžić A., Predicting company growth using logistic regression and neural networks, Croatian Operational Research Review International Scientific Journal, 229–248, 2016.

[10] Reeda P., Wub Y., Logistic regression for risk factor modeling in stuttering research, Journal of Fluency Disorders Vol.38, 88–101, 2013.

[11] Belavagi, M. C., & Muniyal, B., Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection. Procedia Computer Science, 89, 117–123., 2016.

[12] Dharm S., Naveen C.and Jully S., Analysis of Data Mining Classification with Decision tree Technique, Global Journal of Computer Science and Technology Software & Data Engineering, Vol. 13 Issue 13, 2013.

[13] Richter, A. N., & Khoshgoftaar, T. M., A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artificial Intelligence in Medicine., 2018.

[14] Tsai, C.-F., Hsu, Y.-F., Lin, C.-Y., & Lin, W.-Y. , Intrusion detection by machine learning: A review. Expert Systems with Applications, 36(10), 11994–12000., 2009.

[15] Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., … Frieboes, H. B., Prediction of lung cancer patient survival via supervised machine learning classification techniques. International Journal of Medical Informatics, 108, 1–8., 2017.

[16] Etaiwi, W., Biltawi, M., & Naymat, G., Evaluation of classification algorithms for banking customer's behavior under Apache Spark

Data Processing System. Procedia Computer Science, 113, 559–564., 2017.
[17] Joaquín A. and Javier G. C., Improving the Naive Bayes Classifier via a Quick Variable Selection Method Using Maximum of Entropy, entropy, 2017.
[18] Jahangir M. and Jagath J. K., Parameter estimation using artificial neural network and genetic algorithm for free-product migration and recovery, Water Resources Research, Vol. 34, No. 5, Pages 1101-1113, 1998.
[19] Cruz, J. A., & Wishart, D. S., Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics, 2, 59–77., 2006.
[20] Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J., Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine, 11(2), 47–58., 2013.
[21] Gupta, Y., Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, 305–312, 2018.
[22] Prajwala. T. R., A Comparative Study on Decision Tree and Random Forest Using R Tool, International Journal of Advanced Research in Computer and Communication Engineering, 2015.
[23] Zanifa O., Fredrick M., Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised Learning, International Journal for Infonomics, Vol.3,2010.
[24] Reed, P., & Wu, Y., Logistic regression for risk factor modeling in stuttering research. Journal of Fluency Disorders, 38(2), 88–101., 2013.
[25] Zekić-Sušac, M., Šarlija, N., Has, A., … Bilandžić, A., Predicting company growth using logistic regression and neural networks. Croatian Operational Research Review, 7(2), 229–248., 2016.
[26] Fatima, M. and Pasha, M., Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, 9, 1-16., 2017.
[27] William, W., Ware, A., Basaza-Ejiri, A. H., & Obungoloch, J., A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. Computer Methods and Programs in Biomedicine, 164, 15–22.,2018.
[28] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I., Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17., 2015.
[29] Bibault, J.-E., Giraud, P., & Burgun, A., Big Data and machine learning in radiation oncology: State of the art and future prospects. Cancer Letters, 382(1), 110–117., 2016.
[30] M. Bramer, Principles of Data Mining (Third Edition), Undergraduate Topics in Computer Science, Springer Verlag London Ltd., 2016.
[31] Visa S., Ramsay B., Ralescu A., Knaap E., Confusion Matrix-based Feature Selection, Conference: Proceedings of The 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011, Cincinnati, Ohio, USA, April 16-17, 2011.
[32] Jiao, Y., Du, P., Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quantitative Biology, 4(4), 320–330., 2016.
[33] Sumbaly R., Vishnusri N., Jeyalatha S., Diagnosis of Breast Cancer using Decision Tree Data Mining Technique, International Journal of Computer Applications Vol. 98-No.10, 2014.
[34] Sokolova M., Lapalme G., Performance Measures in Classification of Human Communications., Advances in Artificial Intelligence. AI 2007. Lecture Notes in Computer Science, vol 4509. Springer, Berlin, Heidelberg, 2007.
[35] George D. M., Andriana P., Machine Learning in Medical Applications, Springer-Verlag Berlin Heidelberg, 300-307, 2001.
[36] Frank E., Hall M., Holmes G., Kirkby R., Pfahringer B., Witten H. I., Trigg L., Weka-A Machine Learning Workbench for Data Mining, In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA, 2010.
[37] Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H., Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479–2481., 2004.
[38] Therneau T., Atkinson B., An Introduction to Recursive Partitioning Using the RPART Routines., https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf., 2018.
[39] Meyer D., Support Vector Machines—the Interface to libsvm in package e1071, https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf., 2018.
[40] Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T., Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, 83, 1064–1069., 2016.
[41] Rau, H.-H., Hsu, C.-Y., Lin, Y.-A., Atique, S., Fuad, A., Wei, L.-M., & Hsu, M.-H., Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. Computer Methods and Programs in Biomedicine, 125, 58–65., 2016.
[42] Cirkovic, B. R. A., Cvetkovic, A. M., Ninkovic, S. M., & Filipovic, N. D, Prediction models for estimation of survival rate and relapse for breast cancer patients., 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)., 2015.
[43] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR., Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. J Health Med Inform, 2013.
[44] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A., Machine learning models in breast cancer survival prediction. Technology and Health Care, 24(1), 2016.
[45] Topîrceanu, A., & Grosseck, G. Decision tree learning used for the classification of student archetypes in online courses. Procedia Computer Science, 112, 51–60., 2017.
[46] Amirah Mohamed Shahiria, Wahidah Husaina, Nur'aini Abdul Rashida, A Review on Predicting Student's Performance using Data Mining Techniques, Procedia Computer Science 72, 414 – 422, 2015.
[47] Aly M. W., Hegazy F. O., Heba Mohmmed Nagy Rashad, Automated Student Advisory using Machine Learning, International Journal of Computer Applications (0975 – 8887) Volume 81 – No 19, 2013
[48] Sara B.-N., Halland R., Igel C., and Alstrup S., High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study, European Symposium on Artificial Neural Networks, Computational Intelligence, and Machine Learning. Bruges (Belgium), 2015.
[49] Jae Y.C., Lee S., Dropout early warning systems for high school students using machine learning, Children and Youth Services Review 96 ,346–353, 2019.
[50] Erkan E., Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with Information System 100, International Journal of Machine Learning and Computing, Vol. 2, No. 4, 2012.
[51] Khe Foon Hew, Chen Qiao, and Ying Tang, Understanding Student Engagement in Large-Scale Open Online Courses: A Machine Learning Facilitated Analysis of Student's Reflections in 18 Highly Rated MOOCs, International Review of Research in Open and Distributed Learning Volume 19, No.3, 2018.
[52] Chiara M., Geraint J., Tommaso A., Student and school performance across countries: A machine learning approach, European Journal of Operational Research 269, 1072–1085, 2018.
[53] PetkoviĆ MR, RapaiĆ MR, JakovljeviĆ BB. Electrical Energy Consumption Forecasting in Oil Refining Industry Using Support Vector Machines and Particle Swarm Optimization. WSEAS Transactions on Inf. Sci. Appl., 6: 1761-1770., 2009.
[54] Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests. Journal of Manufacturing Science and Engineering, 139(7),2017.
[55] X. Li, C.W. Chan, H.H. Nguyen, Application of the Neural Decision Tree approach for prediction of petroleum production, Journal of Petroleum Science and Engineering, 2013.
[56] Gamze Oğcu, Omer F.Demirel, SelimZaim, Forecasting Electricity Consumption with Neural Networks and Support Vector Regression, Procedia - Social and Behavioral Sciences, Vol. 58, Pages 1576-1585, 2012.
[57] Xie W., Yu L., Xu S., Wang S., A New Method for Crude Oil Price Forecasting Based on Support Vector Machines. In: Alexandrov

V.N., van Albada G.D., Sloot P.M.A., Dongarra J. (eds) Computational Science – ICCS 2006. ICCS 2006. Lecture Notes in Computer Science, vol 3994. Springer, Berlin, Heidelberg., 2006.

[58] Huang, T., Li, B., Shen, D., Cao, J., & Mao, B., Analysis of the grain loss in harvest based on logistic regression. Procedia Computer Science, 122, 698–705., 2017.

[59] Ferentinos, K. P., Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318, 2018.

[60] Bakhshipour, A., & Jafari, A., Evaluation of support vector machine and artificial neural networks in weed detection using shape features. Computers and Electronics in Agriculture, 145, 153–160., 2018.

[61] Crane-Droesch, A., Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters., 2018.

[62] Perez-Alonso, D., Peña-Tejedor, S., Navarro, M., Rad, C., Arnaiz-González, Á., & Díez-Pastor, J.-F. Decision Trees for the prediction of environmental and agronomic effects of the use of Compost of Sewage Slugde (CSS). Sustainable Production and Consumption, 12, 119–133., 2017.

[63] Chen, G., Li, S., Knibbs, L. D., Hamm, N. A. S., Cao, W., Li, T., Guo, Y. A machine learning method to estimate PM 2.5 concentrations across China with remote sensing, meteorological and land use information. Science of The Total Environment, 636, 52–60., 2018.

[64] Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., & Wang, J. Artificial neural networks forecasting of PM 2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmospheric Environment, 107, 118–128., 2015.

[65] Xu, Y., Ho, H. C., Wong, M. S., Deng, C., Shi, Y., Chan, T.-C., & Knudby, A. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. Environmental Pollution, 242, 1417–1426.2018.

[66] Suárez Sánchez, A., García Nieto, P. J., Riesgo Fernández, P., del Coz Díaz, J. J., & Iglesias-Rodríguez, F. J. Application of an SVM-based regression model to the air quality study at the local scale in the Avilés urban area (Spain). Mathematical and Computer Modelling, 54(5-6), 1453–1466.2011.

[67] Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y. M., Park, J., … Cho, K. H. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. Journal of Environmental Sciences, 32, 90–101. 2015.

[68] Torregrossa, D., Leopold, U., Hernández-Sancho, F., & Hansen, J. Machine learning for energy cost modeling in wastewater treatment plants. Journal of Environmental Management, 223, 1061–1067., 2018.

[69] Tsoumakas, G. A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review., 2018.

[70] Patil S. P., Dharwadkar V. N., Analysis of Banking Data Using Machine Learning, International conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud), 2017.

[71] He B., Shic Y., Wan Q., Zhao X., Prediction of customer attrition of commercial banks based on SVM model., 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014, 2014.

[72] Etaiwi, W., Biltawi, M., & Naymat, G. Evaluation of classification algorithms for banking customer's behavior under Apache Spark Data Processing System. Procedia Computer Science, 113, 559–564., 2017.

[73] Addo, P. M., Guegan, D., & Hassani, B. Credit Risk Analysis Using Machine and Deep Learning Models. SSRN Electronic Journal. 2018.

[74] Fang, K., Jiang, Y., & Song, M. Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. Computers & Industrial Engineering,2016.

[75] Gerlein, E. A., McGinnity, M., Belatreche, A., & Coleman, S. Evaluating machine learning classification for financial trading: An empirical approach. Expert Systems with Applications, 54, 193–207. 2016.

[76] Delgado-Gómez, D., Aguado, D., Lopez-Castroman, J., Santacruz, C., & Artés-Rodriguez, A. Improving sale performance prediction using support vector machines. Expert Systems with Applications, 38(5), 5129–5132. 2011.

[77] Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. Journal of Banking & Finance, 72, 218–239. 2016.

[78] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928–2934., 2015.