A Comparison of Machine Learning Algorithms for Internet Cost in Satun Education Sandbox

1st Sasalak Tongkaw
Faculty of Science and Technology
Songkhla Rajabhat University
Songkhla, Thailand
ORCID:0000-0003-3668-2309

2nd Aumnat Tongkaw
Faculty of Science and Technology
Songkhla Rajabhat University
Songkhla, Thailand
ORCID:0000-0002-8516-2601

***Abstract*—Satun Province is a pilot in driving innovation in education in the area. Also known as Satun Education Sandbox. Sixteen schools participate in the project located in Satun Province, the south of Thailand. During the Covid-19 situation, many students study at home. The emergence of online learning increases the internet cost per family. Parents and teachers are aware of this cost. Moreover, the Satun Education Sandbox committee need to draw the strategies from the existed data. In this research, the data was collected from a survey of 2,594 people including teachers and students about the media that used in the teaching and learning during Covid-19 and the internet cost per family. This paper will compare machine learning algorithm, including LR, LDA, k-NN, CART, SVM, Naïve Bayes, SVM, RF, and MLP algorithms, to foresight the cost. After cleaning and some more precise configuration of the results, the details of the data set are described in detail. The model can then predict the expected cost of household internet use at home. It will be helpful for the Ministry of Education plan for further assistance to students in online learning.**

***Index Terms*—Machine Learning, Internet, cost, education**

I. Introduction

Satun Province has introduced educational innovations, including new concepts, methods, processes, teaching media, or management, which has been tried and developed until it is reliable that it can promote learning of learners applied in 16 schools in Satun Province. However, during COVID-19, the teaching and learning style in such schools has changed because it is online teaching but not all. However, online learning causes many problems. In particular, it causes Internet connection costs for each family. Online learning of students in the Satun Sandbox program has two types: fiber-to-home and cellular communication such as 3G, 4G, 5G using a notebook computer. desktop computer and mobile phone These Internet-related expenses are the responsibility of the student's family. Moreover, the government does not know how much it is worth and caused by what affected factors. This paper is a hands-on survey of students in schools participating in the Satun Sandbox program affected by teaching during covid-19 Let us find the relationship between different factors. That results in more internet costs.

978-1-6654-8584-5/22/$31.00 ©2022 IEEE

II. Objective

This research compared the ability of machine learning models including LR, LDA, k-NN, CART, SVM, Naïve Bayes, SVM, RF, and MLP algorithms to predict potential costs in this area by learning from the data from accurate surveys. This result is for the benefit of educational policy in planning aid and remedies and estimating compensation to reduce the burden of parents in having their children study online at home.

III. Machine Learning Techniques

Machine Learning is a way of writing algorithms, using those algorithms to learn a set of exercises. Moreover, convert it to a function, black box, or equation with several tests. That function will be used. It is processed through a function to make the results available. There are two types of learning: supervised and supervised. The learning information based on the information in training is called supervised learning. This research is the application of an instructor-led learning technique. Let us help estimate the cost incurred from internet usage during the past Covid-19 in Satun Province. This technique helps apply the survey data to help budget estimates for families with multiple school-age children to have to study through the online system at the same time. For example, a comparison of phishing detection [1]. Assessment of Supervised Learning outcomes is based on accuracy, and each problem may have a different approach. The view of use will be different. Moreover, because the information was obtained from the survey, make the data have a high estimation discrepancy. Algorithm comparison is another way to know prediction efficiency using machine learning with supervised learning [2].

A. Logistic Regression (LR)

Logistic regression analysis is a technique for analyzing the relationship of variables in the form of predicting the likelihood or probability of the occurrence or absence of an event of interest. It is often used in cases where the dependent variable is either a categorical variable or at least one independent variable is possible, either a quantitative or group variable. This method can predict the likelihood of an event of interest for a dependent variable. In most research,

Binomial Logistic Regression Analysis has been used in which the dependent variable had only two values, 0 and 1 with the mathematical model as the probability distribution function. In this study, 0 represents Internet costs for families less than 500 baht, and 1 represents internet costs per family of more than 501 baht.

B. Linear Discriminant Analysis (LDA)

Discriminant analysis was a statistical method used to discriminate between 2 or more groups by analyzing from one dependent variable and one or more independent variables. In addition, to distinguish between groups, it can also tell some nature of the classification of groups. For example, it can tell which variable is less well classified. That is, the efficiency or weight of those variables can be classified. Cluster analysis uses predictive variables or independent variables together to predict the dependent variable. It is a statistical technique similar to Multiple Regression Analysis.

C. K-nearest Neighbors Classifier (K-NN)

K-Nearest Neighbor Algorithm) is a method used to classify classes. It uses to decide which classes can represent new conditions or cases by checking a specific number? The nearest neighbor algorithm finds cases or conditions that are the same or closest to each other. It will find the sum, count up, of the number of conditions or cases for each class and set new conditions that give the same class as the closest class. The "nearness" is determined by a distance metric which determines how far two records are $(x_1, x_2, ..., x_p)$ and $(u_1, u_2, ..., u_p)$. Two metrics are used, the Manhattan distance and the Euclidean distance.

In medicine, K-NN is used as an algorithm in clinical application for detecting diseases, classifying type of diseases, and testing of new medicines [3], [4], similar to K-NN, which can acquire new knowledge from old data to diagnose the nature of the symptoms of the original patient to diagnose new patients. Similar to this research, K-NN is used to estimate expenses. It can be estimated that household internet costs are less than 500 or more than 500 baht per family.

D. Classification and Regression Trees (CART)

CART, a machine-learning method for constructing prediction models from data, was invented by Breiman in 1984. CART is capable of both classification and regression. The decision tree generated by the CART algorithm is a binary tree consisting of two branches for each node. The models are created by recursively splitting the data space and fitting a simple prediction model to each partition. As a result, the partitioning can be graphically shown as a decision tree. The user must use the gain and gain ratio concept, as well as hypothetically measured facts, to build a decision tree. When dealing with a large number of values, however, the gain computation has its drawbacks. As a result, using the gain ratio is your best bet. Where x_k is the value of a feature and T is the test set, the gain ratio equation is illustrated in (1). $split(x_k, T)$ is the information obtained as a result of separating the test set on feature x_s [5].

$$gainratio(x_k, T) = \frac{gain(x_k, T)}{split(x_{k,T})} \tag{1}$$

$$split(x_k, T) = \sum_{i=1}^{n} \frac{|T_i|}{T} log_2 \left(\frac{|Ti|}{|T|} \right) \tag{2}$$

To construct decision points for classification problems, this method employs the gini index, a new statistic. The criterion function defaulted in sklearn.tree.DecisionTreeClassifier is "gini". Another criterion used in this research is 'entropy'. They use to measure the quality of a split.

E. Naïve Bayes Algorithm

In terms of Gaussian Naïve Bayes, GaussianNB, this study used the Naïve Bayes method. The Naïve Bayes supervised learning method uses a target and a label, and the group of internet cost was employed as a label in this study. A Naïve method is theoretically based on Bayes' theorem, which asserts that every pair of features is equal. Bayes' theorem is proposed class variable as y and dependent feature as x_1 to x_n

$$P(y|x_1, ..., x_n) = \frac{P(y)P(|x_1, ..., x_n|\, y)}{P(x_1, ..., x_n)} \tag{3}$$

For the input, we can apply this classification formula for constant as $P(x_1, ..., x_n)$.

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)) \tag{4}$$

The Maximum A Posteriori (MAP) method was also utilized to estimate $P(y)$ and $P(x_i|y)$ in this study.

F. Support Vector Machine (SVM)

A support vector machine (SVM) is a computer technique that uses supervised learning to assign labels to objects. For example, by reviewing hundreds or thousands of fraudulent and nonfraudulent credit card activity records, an SVM can learn to distinguish fraudulent credit card activity. An SVM may also learn to detect handwritten numerals by evaluating a vast collection of scanned images of handwritten zeroes, ones, and other symbols. SVMs are now being used in a growing number of biological applications with great success [6].

G. Random Forest : RF

A random forest is a group of models known as ensemble learning, where the principle is to train the same model multiple times (multiple instances) on the same data set. Each training session will select different parts of the data that are being trained. Then decide on those models to vote on which class is the most chosen. Make work efficiency higher More precisely. The Random Forest model is prevalent in machine learning.

By performing classification with a small number of trees, it is possible to enhance the random forest classifier. The

approach iteratively removes certain irrelevant features based on the number of essential and unimportant features. It then formulates a novel theoretical upper limit on the number of trees that should be added to the forest to ensure that classification accuracy improves [7].

H. Multi Layer Perceptron (MLP)

The ANN Multi-Layer Perceptron is made up of nodes or units that are organized in two or more layers, with the input layer left out. While there are no nodes on the same layer, real value weights are connected to some nodes [8]. MLP's architecture could be mathematically stated as follows (5)

$$a_{i,q}^{l} = \sum_{n_j^m \epsilon S_i^l} w_{ij}^{lm} y_{j,q}^{m}, l > 0 \tag{5}$$

$$y_{i,q}^{l} = f(a_{i,q}^{l}), l > 0 \tag{6}$$

Where $a_{i,q}^{l}$ denotes node n_i^l's activation for a specific pattern q and layer l. Node n_j^m is the source node for n_i^l, while w_{ij}^{lm} is the weight vector that connects nodes n_i^m and n_i^l. $y_{j,q}^{m}$ represents the pattern q's source input, while the variable S_i^l describes the set of source nodes. n_0^l stands for bias nodes. The second idea is that $y_{i,q}^{l}$ is the output of an activation function with the input of the node's activation; the sigmoid function in (7) is most usually utilized.

$$f(x) = \frac{1}{(1 + e^{-x})} \tag{7}$$

The intermediate layer is sometimes referred to as the hidden layer, and hidden nodes exist inside the layers. The number of hidden layers and hidden nodes impact the MLPs' training functionality. Too few can result in problems not being solved, whereas too many might result in longer training times or insufficient generalization abilities. For researchers utilizing MLPs, the best-hidden layers and nodes that might solve precision and low minimal approximated error are sought. Depending on the node withholding data set, optimizations could be done by growing and pruning networks.

IV. Research Method

This research employed an empirical study by using the original data from the survey. We divide the testing process into two sections: a) accuracy comparison results and classification and regression tree results.

A. Data sets

This research collected data from a survey of students, parents, and teachers in the Educational Innovation Area covering 26 Tumbons 6 Ampores in Satun Province. Sixteen schools were participating in this activity. Between 4-27 December 2021. There were a total of 2,628 participants with 26 features, X1=City; X2=Class Level; X3=Mobile; X4=Tablet; X5=Notebook; X6=Paper; X7=Desktop; X8=Number of Devices; X9=Number of Students; X10=Enough; X11=Internet; X12=Google Classroom; X13=Paper Work; X14=ClassStart; X15=Line; X16=Facebook Messenger; X17=Email X18=Meet X19=PowerPoint; X20=YouTube; X21=Facebook; X22=Quizz; X23=Zoom; X24=Team; X25=Webex, and X26=Cost level. The cost level is the label. Detecting and removing duplicate records is one of the most critical responsibilities in data cleansing. Any incomplete records will be deleted. Only 2,594 active records left.

B. Criteria of Comparison

Model function used in Jupyter python version 3.7 includes: LogisticRegression(), LinearDiscriminantAnalysis(), KNeighborsClassifier(), DecisionTreeClassifier(max depth=5), GaussianNB(),SVC(),RandomForestClassifier(max depth=5, n estimators=10, max features=1), Results and Discussions, and MLPClassifier(alpha=1, max iter=1000). The result include the accuracy in particular model with different n splits, 10,20,30, and 40. The random seed is 7. For the decision tree classifier, this paper test in two criteria, first with the default criterion 'gini' and the criterion 'entrophy'.

V. Results and Discussions

This section will show the results and discussions about the results follow the criteria setting. It is divided in two sections: accuracy comparison amoung models, and the decision tree classifier with different criterion.

a) Accuracy comparison results: This section will show the comparison of KFold by collecting the results from n splits: 10, 20, 30, and 40. By using LR, LDA, KNN, CART, NB, SVM, RF, and MLP algorithm. The boxplot will show in Fig. 1. The numbers in the table show the accuracy score (standard division).

TABLE I
Algorithm Compare Accuracy Results

Algorighm	KFold-10	KFold-20	KFold-30	KFold-40
Logistic Regression (LR)	0.692366 (0.035775)	0.692018 (0.040206)	0.690430 (0.049757)	0.693666 (0.051966)
Linear Discriminate Analysis (LDA)	0.693910 (0.034069)	0.691634 (0.041275)	0.691232 (0.053452)	0.692885 (0.052544)
K-Nearest Neighbors (k-NN)	0.640713 (0.052228)	0.646547 (0.052378)	0.641099 (0.063379)	0.635493 (0.076754)
Classification and RT (CART)	0.688509 (0.040548)	0.688930 (0.044496)	0.680812 (0.047760)	0.689411 (0.063137)
Naïve Bayes algorithms (NB)	0.666515 (0.050887)	0.666574 (0.055381)	0.667598 (0.067578)	0.668191 (0.066864)
Support Vector Machine (SVM)	0.686978 (0.040972)	0.689723 (0.041062)	0.686982 (0.055380)	0.689429 (0.061879)
Random Forest Classifier (RF)	0.649235 (0.071218)	0.658882 (0.057278)	0.613463 (0.099166)	0.660931 (0.072425)
Multi Layer Perceptron (MLP)	0.696613 (0.033198)	0.694350 (0.040821)	0.696623 (0.053060)	0.695986 (0.059558)

TABLE 1 shows that the maximum accuracy is the MLP model with KFold-10, gives 69.66% accuracy result, followed by LDA 69.39% accuracy result and LR 69.23%, respectively. In similar way MLP also have minimum standard division with KFold-10, follow by LDA and LR. The MLP also had the highest accuracy among all the algorithms compared. Even if KFold is increased to 20, 30, or 40.

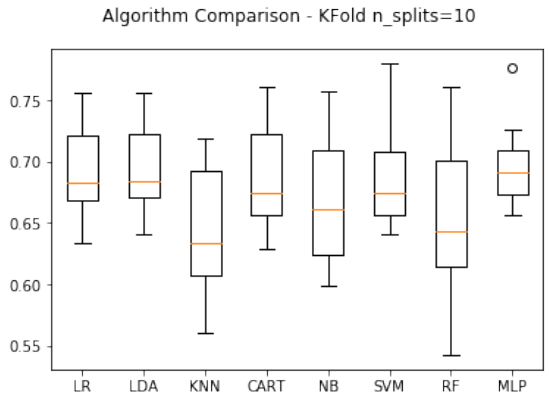

Fig. 1. Algorithm Comparison KFold nsplits =10

b) Classification and Regression Trees results: Fig. 2 shows the correlation matrix. It can report that the type of Internet connection (X11), and paper work assignment (X13) affect the internet cost per family. There are two internet systems at students' homes: Wi-Fi at home connected to an ISP that provides high-speed Internet via fiber optic cables and cellular networks such as 3G 4G, which, if higher speeds are required, users need to pay more. It is possible to a contract with a home Wi-Fi ISP provider. Inconvenient for users' expenses will be regular expenses. However, using the service via the cellular network is more convenient to use the service. Nevertheless, the cost is higher, thus affecting the monthly internet expenses more. The result from decision tree classifier is show in Fig. 3. The features that classification as a root of the tree is enough perspective (X10), the participant own opinion that they have adequate support for learning at home or not. Another layer feature is paper work assignment (X13), the learning at home use paper work or not. The decision tree with two criterion gave the same features results.

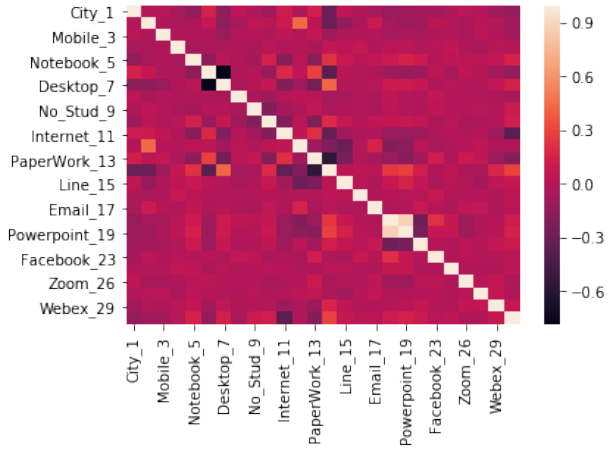

Fig. 2. Correlation Matrix result

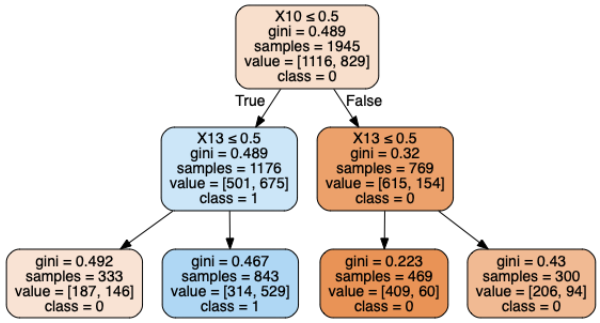

Fig. 3. Decision Tree Results, criterion = gini

VI. Conclusion

This article compares the machine learning algorithms includings LR, LDA, k-NN, CART, SVM, Naïve Bayes, SVM, RF, and MLP to estimate household Internet costs. The best algorithm for estimating is MLP, where KFold=10 gives 69.66% accuracy result, followed by LDA 69.39% accuracy result and LR 69.23%, respectively. As a result of finding the correlation matrix between variables, it turns out that the most critical factor affecting household Internet costs was the type of Internet connection. Next is Class Start and homework assignments as paperwork. The best model of prediction, in this case, is the MLP, which can estimate the cost of Internet per family based on the family context in Satun Province. Government education could be suggested by supporting internet connection from home. It will reduce the cost per family down and increase the learning of students per family.

Acknowledgment

The authors would like to thank members of Satun Education Sandbox, teachers, students, parents, and all other participants in Satun Education Sandbox, Satun province.

References

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, ser. eCrime '07. Association for Computing Machinery, pp. 60–69. [Online]. Available: https://doi.org/10.1145/1299015.1299021

[2] Q. Zhao, S. Yu, F. Zhao, L. Tian, and Z. Zhao, "Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments," vol. 434, pp. 224–234.

[3] J. Ren, "ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging," vol. 26, pp. 144–153.

[4] D. lin, A. V. Vasilakos, Y. Tang, and Y. Yao, "Neural networks for computer-aided diagnosis in medicine: A review," vol. 216, pp. 700–708. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216308815

[5] D.-W. Sun, *Computer Vision Technology for Food Quality Evaluation*. Academic Press.

[6] W. S. Noble, "What is a support vector machine?" vol. 24, no. 12, pp. 1565–1567. [Online]. Available: https://www.nature.com/articles/nbt1206-1565

[7] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," vol. 27, no. 8, pp. 4012–4024.

[8] A. J. Shepherd, *Second-Order Methods for Neural Networks: Fast and Reliable Training Methods for Multi-Layer Perceptrons*. Springer Science & Business Media.