Thai Journal of Mathematics
Special Issue: The 17th IMT-GT ICMSA 2021
Pages 157–168 (2022)

http://thaijmath.in.cmu.ac.th

ISSN 1686-0209

A Comparison of Machine Learning Techniques for Classification in Bank Marketing Data

Waritpon Saengthongrattanachot[1], Anamai Na-udom[1,*] and Jaratsri Rungrattanaubol[2]

[1] *Depertment of Mathematics, Faculty of Science, Naresuan University*
e-mail : waritpon63@nu.ac.th (W. Saengthongratanachot); anamain@nu.ac.th (A. Na-udom)

[2] *Department of Computer Science and Information Technology, Faculty of Science, Naresuan University*
e-mail : jaratsrir@nu.ac.th (J. Rungrattanaubol)

Abstract Nowadays big data has played a crucial role in the context of data science. The components of big data originally consist of volume, velocity, and variety. Once a big set of data has been collected, the extraction of information in terms of classification is of interest for decision making. In this study, we consider the Bank Marketing dataset, which related to direct marketing campaigns of a Portuguese banking institution. It consists of 36,548 number of instances with 20 attributes and the goal is to predict if the clients will subscribe (yes/no) a term deposit. The main objective of this paper is to investigate the performance of the machine learning techniques on the Bank Marketing dataset. A series of machine learning techniques for classification used in this study are Decision Tree (DT), Random Forest (RF), Random Tree (RT), Naive Bayes classifier (NB), and k-Nearest Neighbor (kNN). A comparison is made through percentage of classification accuracy. The results show that the accuracy obtained from all techniques are in the range of 73.40% and 87.11%. The best accuracy 87.11% is obtained from RF method followed by DT with J48 algorithm with accuracy about 87.10%. The worst one 73.40% is produced by kNN method with k=3. Hence RF and J48 methods are suitable to use in classification problem in Bank Marketing data analytics. In addition, by reducing the input attributes from 20 to 5 relevant attributes, the simpler version of J48 model is obtained with the acceptable performance.

MSC: 49K35; 47H10; 20M12

Keywords: machine learning techniques; big data; classification; decision tree, random forest; k-nearest neighbor; Naive Bayes

Submission date: 15.03.2022 / Acceptance date: 31.03.2022

[*] Corresponding author.

Published by The Mathematical Association of Thailand.

1. Introduction

Big data is a term that describes large, fast or complex and hard to manage or impossible to process using traditional methods. It also refers to both of structured and unstructured data. Big data originally consists of three key concepts, then extended to five: volume, velocity, value, variety and veracity [1][2]. Currently, the usage of big data is mainly based on predictive analytics, user behavior analytics, or other methods to extract crucial information from big data for decision making in any business sectors. The analysis of big data is very challenging due to its size and complexity. Hence, the efficient methods like machine learning are of interest in big data analytics [3].

Machine learning (ML) is the study of computer algorithms that can improve automatically through observed pattern and the use of data [4][5][6]. ML algorithms construct a model based on input data or training data to predict output values within an acceptable range. As new input data is fed to the algorithms, ML algorithms learn and try to optimize their operation in order to improve performance or accuracy. ML algorithms are widely used in various applications such as in health sciences, business, sentiment analysis, speech recognition, and computer vision [7]. ML consist of two approaches, one is to classify data based on models which have been developed, the other objective is to predict the future outcomes. For example, consider fraud detection, every time someone makes a purchase using a credit card, ML algorithms immediately check their purchases to verify whether or not this might be a fraudulent transaction. They also predict whether it is fraudulent or not based on whether that purchase is consistent with the features of previous purchases. Search and recommendation systems are also a vast area of application for machine learning [8].

In general, ML algorithms are categorized into two main types. The first type is known as supervised learning, the machine is taught by example in which the goal is to predict some output variable that is associated with each input item. The algorithm identifies patterns in data, learns from observations and then makes prediction. The prediction process is corrected by the operators and repeated until the algorithm achieves such a high level of accuracy. The examples of supervised learning included classification, regression and forecasting [7][9][10][11]. A classification problem within supervised learning, and the function used to perform the classification task is called the classifier. The second major class of ML algorithms is called unsupervised learning. In this class, the ML algorithm studies data to identify patterns. There is no answer key or human operation to provide instruction. Such problems are solved by finding some useful structure or relationship between input data, in a procedure called clustering. Therefore, once we can discover this structure in the form of clusters, groups or subsets, this composition will be used for tasks like producing a useful summary of the input data. Unsupervised learning allows us to approach problems with little or no idea about the final result. If it is processed more data, its ability to make decisions on that input data will be improved and refined [5][8].

The present work focuses on supervised ML and the main aim is to compare prediction accuracy among various ML classification techniques on the marketing bank dataset. A series of ML techniques used in this study are Decision Tree (DT), Random Tree (RT), Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbor (kNN). A comparison is made mainly through percentage of classification accuracy based on k-fold evaluation. In addition, some other related metrics including precision, recall and F-measure are presented in the result.

2. MACHINE LEARNING ALGORITHMS

As mentioned before, this study aims to investigate and compare the performance of different supervised ML algorithms on a case study of Bank Marketing data [12]. The details of each algorithm are given below.

2.1. DECISION TREE AND RANDOM TREE

Decision Tree (DT) algorithm is one of the most popular algorithms that have been extensively used in the context of data science. It is a flow chart like tree structure that uses a branching method to illustrate every possible outcome of a decision. Each node within the tree represents a test on a specific attribute and each branch is referred to the outcome of the test. DT is used for classifying problems and works well in both of categorical and continuous output variable or class. Random Tree (RT) works exactly like decision tree with one exception is that for each split only a random subset of attributes is available. The random tree nodes use bootstrap sampling with replacement to generate sample data and then the sample data is used to grow a tree model. During tree growth, RT will select part of attributes and use the best one to split a tree node. This process is repeated when splitting each tree node [13].

2.2. RANDOM FORESTS

Random Forests (RF) or random decision forests is an ensemble learning method. It combines multiple algorithms to generate better results for classification. The algorithm starts with a decision tree and the input is fed to the top. It then moves down the tree while data is segmented into smaller sets based on specific attributes. To classify a new object based on its attributes, each tree is classified, and the tree votes for that class. The forest chooses the classification having the most votes (over all the trees in the forest) [14].

2.3. NAÏVE BAYES

A Naïve Bayes (NB) classifier uses Bayes theorem and classifies every value as independent of any other values or it assumes that the presence of a particular feature in a class is unrelated to the presence of any other features. Though, if these features are related to each other, a NB classifier would consider all these properties independently when calculating the probability of a particular outcome. A Naïve Bayes model is easy to build and useful for massive datasets. Despite its simplicity, it has been shown to outperform even highly sophisticated classification methods [15].

2.4. k-NEAREST-NEIGHBOUR

The k-Nearest-Neighbour (kNN) algorithm estimates how likely a data point is to be a member of one group or another. It essentially looks at the data point around any specific data point to determine what group it actually in. This algorithm can be applied to both classification and regression problems. It is a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbours by using a distance function performs this measurement. To perform kNN, the optimal k must be empirically specified which makes kNN algorithm computationally expensive. Further, all features should be normalized prior to proceed kNN [13].

3. DATA ANALYTICS AND PREDICTION

This paper aims to construct the predictive model using decision tree techniques, Naive Bayes and k-Nearest Neighbour [5] [6] [13] [14] to classify whether the client subscribed a term deposit or not, using the Bank Marketing dataset [12], and then comparing their performance and find the related factors or attributes. The process of constructing the model consists of four steps and are presented as follows.

3.1. DATA UNDERSTANDING

The dataset used in this paper is a secondary dataset about the marketing bank data obtained from UCI [12], which is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls and more than one contact to the same client was taken, to access if the bank term deposit would be yes or no (subscribed or not). The original dataset contains 36,548 records, 20 input attributes and 1 output (or class of yes/no). The meaning and characteristics of each attribute is shown in Table 1. There are 10 qualitative and 10 quantitative attributes. Within 36,548 records, there are 4,640 records for a class of yes (Minority class) and 26,629 records for no (Majority class). The dataset is imbalanced with the ratio between number of minority and majority class is about 1:7 (12.2%, 88.8%). The construction of the predictive model normally tries to predict the class based on the input dataset, when the number of records of certain class are much larger than the other class, the predictive model may get biased towards the prediction, i.e. Majority class [16] [17].

3.2. DATA PREPARATION

The dataset obtained from [12] is imbalanced. In principles there are many approaches to deal with class imbalance problem [16] [17] [18]. However, to make it simple and to fulfill our objective of this study, which is to make a comparison of ML techniques, we just simply apply under-sampling technique on the majority class using simple random sampling without replacement to make the dataset balanced before constructing the model. After random sampling on the majority class no (26,626 records), we obtained 4,640 records of no. Therefore, the dataset, used for constructing the predictive models and a comparison of ML techniques, contains an equal number of records between yes and

TABLE 1. The meaning and characteristics of each attribute.

Attribute name	**Meaning (Characteristics)**
age	Age in year (numeric)
job	Type of job (12 categories)
marital	Marital status (4 categories)
education	Education (8 categories)
default	Has credit in default? (3 categories)
housing	Has housing loan? (3 categories)
loan	Has personal load? (3 categories)
contact	Contact communication type (2 categories)
month	Last contact month of year (12 categories)
day	Last contact day of the week (7 categories)
duration	Last contact duration, in seconds (numeric)
campaign	Number of contacts performed during this campaign (numeric)
pdays	Number of days that passed by after the client was last contacted (numeric)
previous	Number of contacts performed before this campaign (numeric)
poutcome	Outcome of the previous marketing campaign (3 categories)
emp.var.rate	Employment variation rate quarterly indicator (numeric)
cons.price.idx	Consumer price index monthly indicator (numeric)
cons.conf.idx	Consumer confidence index monthly indicator (numeric)
euribor3m	Euribor 3-month rate daily indicator (numeric)
nr.employed	Number of employees quarterly indicator (numeric)
class	**Has the client subscribed a term deposit? (yes or no)**

no, with the total of 9,280 records. From this point, this dataset is analyzed, visualized, and used to construct the predictive models.

The displays of bar charts on some categorical input attributes classified by yes and no reveal the effects of their values on the output as shown in Figure 1, which are job and education. Similarly the box plots of some numeric attributes classified by yes and no are presented in Figure 2, which are duration, nr.employed and emp.var.rate. It can be noted that duration, nr.employed and emp.var.rate attributes are likely related to the output.

3.3. MODELING

In this study we use a k-fold cross validation concept to construct the predictive model. A k-fold cross validation is one of popular techniques for a model construction. When constructing the model, the dataset is divided into a training set and a test set. A training set is used to construct the predictive model and a test set is for the model evaluation and performance. This paper applies the concept of k-fold cross validation in designing a training set and test set. In k-fold, the dataset is divided into k groups using a random sampling without replacement technique, then the k-1 groups are combined to

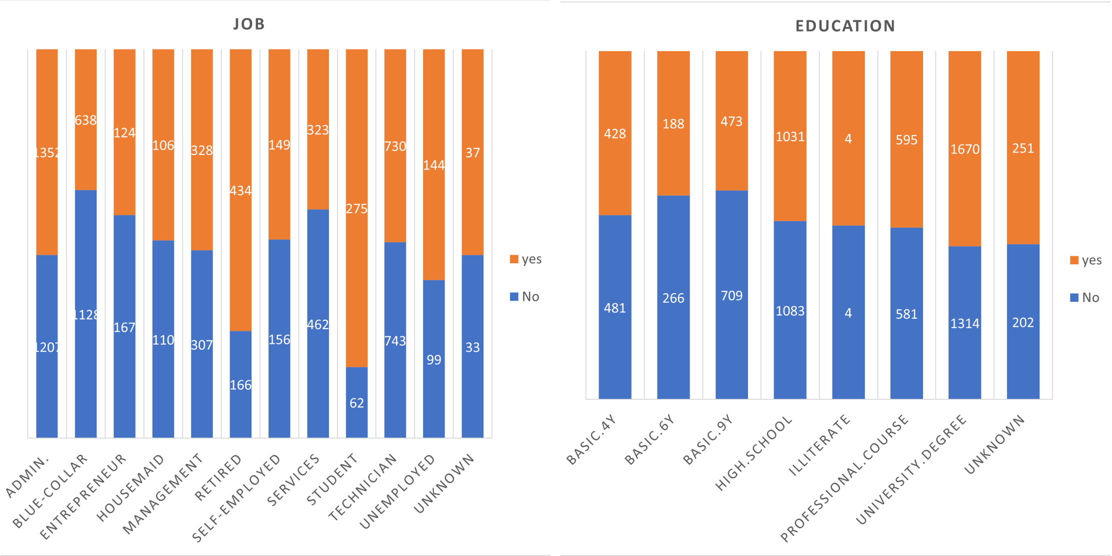


Figure 1. Bar charts of some categorical attributes classified by yes/no.

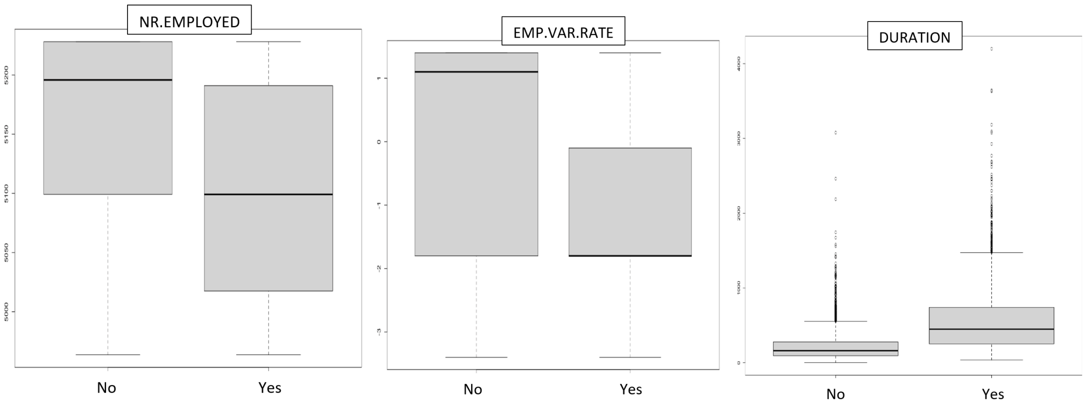


Figure 2. Box plots of some numeric attributes classified by yes/no.

be a training set and one left group is a test set. Repeating construction the model and test by rotating group members for the training and test set k times, the performance is an average of accuracy on the test data. In this study, we set k to 10 as suggested in[13]. The process of constructing the models based on 10-fold cross validation is shown in Figure 3. Thus, the models for each ML technique are constructed and evaluated 10 times. By setting k to 10 is equal to the splitting ratio 90:10 on a training and test set. The accuracy of each ML algorithm is calculated by Equation (1).

$$\boldsymbol{Accuracy} = \frac{1}{10}\sum_{i=1}^{10} Accuracy\_on\_test_i \tag{3.1}$$

We use WEKA (Waikato Environment for Knowledge Analysis) Version 3.8.5 [19], which is a free ML tool developed at University of Waikato, New Zealand to construct the models. Decision-tree algorithms selected are J48, Random Forest (RF) and Random Tree (RT). Moreover, we used Naive Bayes classifier (NB) and k-Nearest Neighbor (kNN)

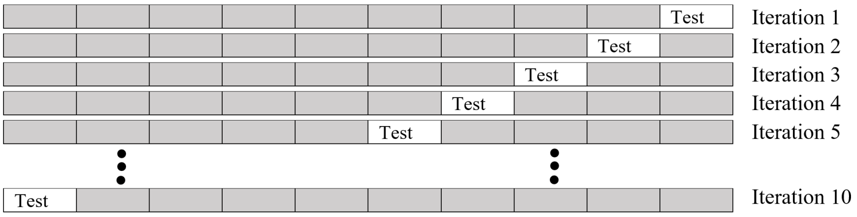


Figure 3. A process of 10-fold cross validation.

with varying k to 3, 5 and 7. Firstly, we constructed the models with all input attributes (20 attributes), later by studying the predictive models we constructed the models with 5 related input attributes. All parameters of algorithm, especially DT, RF and RT, are set as the default value of WEKA as implemented in[20].

3.4. EVALUATION

The performance of predictive models is evaluated with four criteria: Accuracy, Precision, Recall and F-measure. The calculations of them are related to the Confusion Matrix, which is a summary of prediction results. The number of correct (TP and TN) and incorrect (FN and FP) predictions are summarized with count values and broken down by each class (i.e. yes/no) as displayed in Table 2.

Table 2. A Confusion Matrix

		Predicted Values	
Actual Values		**Positive (yes)**	**Negative (no)**
	Positive (yes)	*TP*	*FN*
	Negative (no)	*FP*	*TN*

The meaning of each criterion and its formula is described as follows.

1. **Accuracy** (Acc) is given as the percentage of total correct predictions divided by the total number of records, as shown in Equation (2).
2. **Precision** is an ability of the model to identify only the relevant data points. Mathematically, precision is defined as the number of true positives divided by the number of true positives plus the number of false positives, as shown in Equation (3).
3. **Recall** is an ability of a model to find all the relevant cases within a dataset. Mathematically, recall is defined as the number of true positives divided by the number of true positives plus the number of false negatives, as shown in Equation (4).
4. **F-measure** is the harmonic mean of precision and recall, taking both metrics into account, as shown in Equation (5).

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)} \tag{3.2}$$

$$Precision = \frac{1}{2} * (\frac{TP}{(TP + FP)} + \frac{TN}{TN + FN}) \tag{3.3}$$

$$Recall = \frac{1}{2} * (\frac{TP}{(TP + FN)} + \frac{TN}{TN + FN}) \tag{3.4}$$

$$F - measure = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{3.5}$$

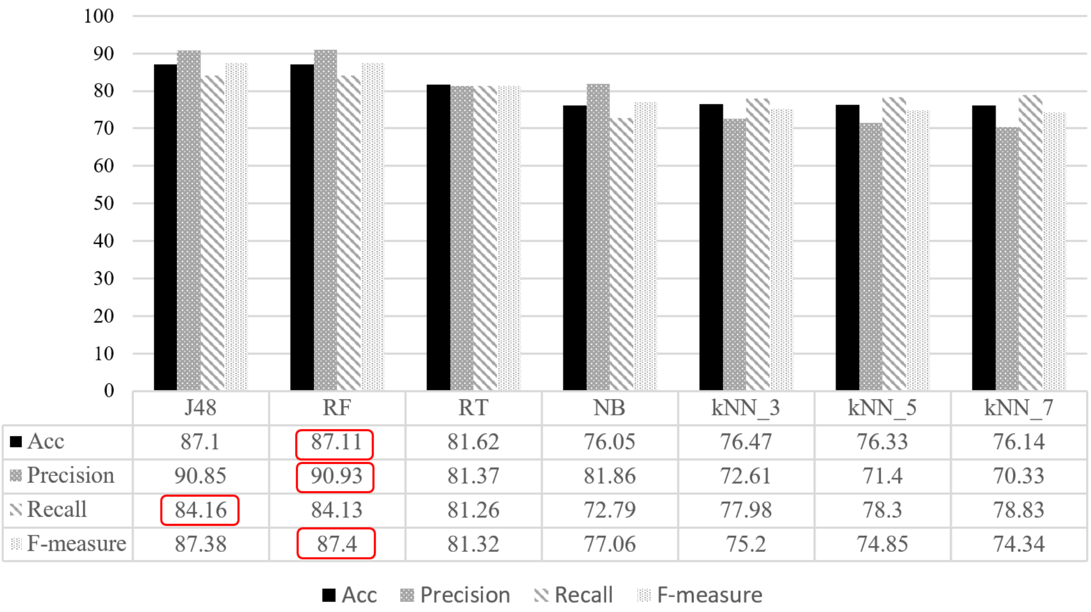
It can be noted that the value of Precision and Recall used in this study are the average of Precision of 'yes' and Precision of 'no', similarly applied to Recall, since the dataset is balanced after applying under-sampling technique as discussed in section 3.2.

4. RESULTS AND DISCUSSION

As mentioned before, the 10-fold cross validation is used in the study, then the model is repeatedly constructed 10 times on each ML technique. Hence, the values of each measurement displayed in Figure 4 and Table 3 are the average of 10 predictive models evaluated on the test set as described in Figure 3 and Equation (1). In this paper, DT algorithm used here is called J48. Hence, the term J48 refers to DT technique. A result of the predictive models constructed with 20 input attributes from each technique is presented in Figure 4. It shows that RF and J48 perform best with almost the same accuracy 87.11% and 87.1%, followed by RT with 81.86%. Whereas NB and three various kNN perform worse with similar accuracy about 76%. For the other values (Precision, Recall and F-measure), they are in a similar pattern as accuracy, only that the recall of J48 outperforms RF.

Although RF (87.11%) outperforms J48 (87.10%), only a little higher accuracy of 0.01, as shown in Figure 4. Regarding to ML algorithms explained in section 2, DT or J48 algorithm is much simpler than RF. Hence, the tree structure of J48 predictive model appeared to be less complex than RF. Since the J48 model has an acceptable accuracy with less complex in tree structure, further investigation on modification of the J48 model is proceeded.

The J48 predictive tree model displayed in Figure 5, however, is still quite complicated especially at the bottom branches, the simplified J48 tree is further studied to obtained the most relevant attributes to be used in constructing predictive models for each technique again. Based on DT algorithm, the relevant attributes normally appear at the top node of the tree. As depicted in Figure 5, the top four levels of the tree (zoom into the tree), it contains five attributes, which are nr.employed, duration, cons.price.idx, cons.conf.idx and emp.var.rate, which in this study are considered as the most relevant factors for the output. These five attributes are used in constructing the predictive models for each ML techniques. Then repeating modeling with these five input attributes for each ML technique based on 10-fold is performed. The performance of predictive models using 20 attributes (base) and the adjusted models using 5 attributes (adjust) are recorded and compared as displayed in Table 3.



	J48	RF	RT	NB	kNN_3	kNN_5	kNN_7
■ Acc	87.1	87.11	81.62	76.05	76.47	76.33	76.14
▨ Precision	90.85	90.93	81.37	81.86	72.61	71.4	70.33
⧅ Recall	84.16	84.13	81.26	72.79	77.98	78.3	78.83
░ F-measure	87.38	87.4	81.32	77.06	75.2	74.85	74.34

Figure 4. The performance of predictive models from each technique.

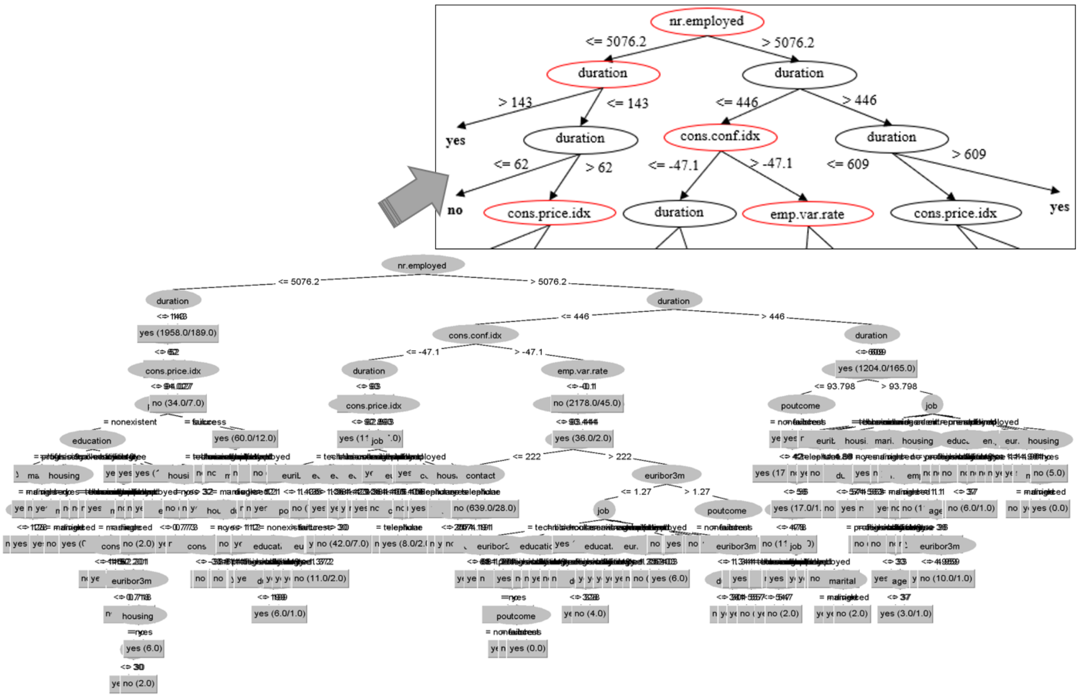


Figure 5. An illustrative result of the decision tree J48.

Table 3 shows the performance of the predictive models with 5 attributes (adjust), comparing with the model with 20 attributes (base). The accuracy of adjusted J48 and RT drops slightly, whereas the RF drops significantly. On the other hand, the accuracy of adjusted NB and three kNN increases, it indicates that these adjusted models perform better. Table 3 also displays that the accuracy of NB increases slightly, whereas three kNN

TABLE 3. The performance of adjusted predictive models with top five relevant attributes.

	Accuracy		**Precision**		**Recall**		**F-measure**	
	base	adjust	base	adjust	base	adjust	base	adjust
J48	**87.10**	86.91	90.85	**91.76**	**84.16**	83.28	**87.38**	87.33
RF	**87.11**	82.54	90.93	**82.28**	**84.13**	82.21	**87.40**	82.24
RT	**81.62**	81.48	81.37	**82.07**	**81.26**	80.61	81.32	**81.33**
NB	76.05	**76.98**	81.86	**85.80**	**72.79**	72.43	77.06	**78.55**
kNN_3	76.47	**84.90**	72.61	**88.31**	77.98	**82.26**	75.20	**85.18**
kNN_5	76.33	**86.11**	71.40	**89.66**	78.30	**83.33**	74.85	**86.38**
kNN_7	76.14	**86.50**	70.33	**90.72**	78.83	**83.30**	74.34	**86.85**

increase significantly. Among the modified models (adjust), J48 has the most accuracy followed by kNN_7 and kNN_5. All precision values of adjusted models are better than the based models. Although, J48 (adjust) has a slight less accuracy than the J48 (base), the structure of J48 tree with 5 attributes has essentially less complex than the J48 tree with 20 attributes, as a comparison displayed in Figure 5 and Figure 6.

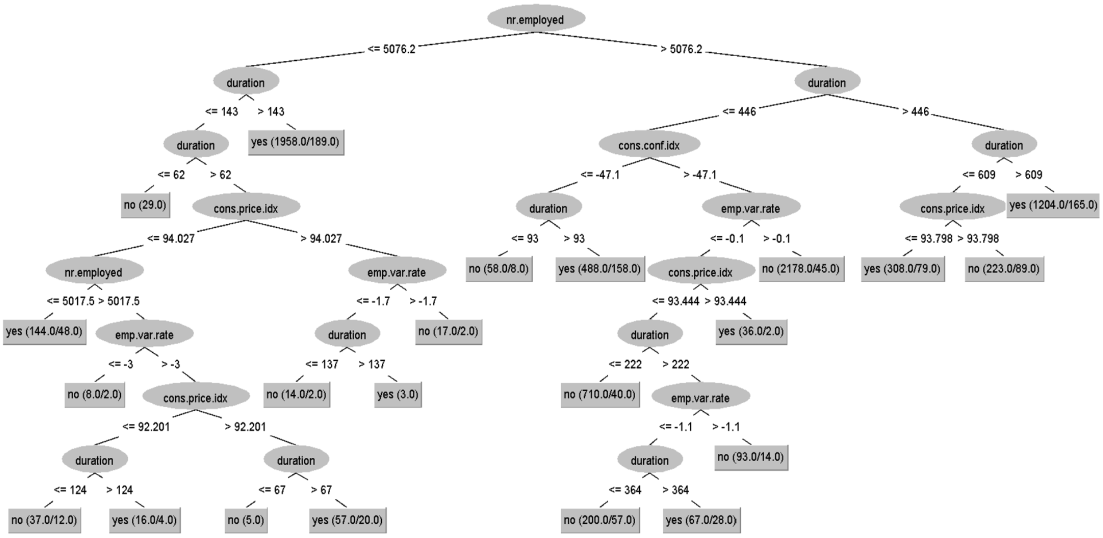


FIGURE 6. An illustrative result of the decision tree J48 with modification.

When considering the ease of deployment or usage of the predictive models, by reducing the input attributes, the decision tree becomes easier to understand and interpret, even though it has a slight less accuracy comparing to the base J48. In term of application, the tree on Figure 6 can be straightly developed as a predictive program for users.

5. CONCLUSIONS

In this paper, a comparative study is carried out with supervised machine learning algorithms to classify the decision of the clients whether they will subscribe a term deposit. The data used in this paper is the Bank Marketing dataset which is related to direct marketing campaigns of a Portuguese banking institution since the original dataset is imbalanced, we applied under-sampling technique on majority class ('no') to get the balanced dataset for constructing models and comparing performance of the models from each ML technique. The results reveal that RF has the most accuracy, followed by DT J48 with just 0.01 difference, NB and three kNN models perform quite poor when using all 20 attributes. With further investigation, we proposed that by selecting top five relevant attributes or factors using J48 and the relevant attributes are nr.employed, duration, cons.price.idx, cons.conf.idx and emp.var.rate. The results obtained from the refined model reveals that kNN performs significantly better when using only relevant attributes, whereas, the accuracy of J48 model with relevant attributes slightly decreases (i.e. about 0.19). The results also show that we can reduce the number of attributes from 20 to 5 on the J48 model to obtain the much simpler predictive models with an acceptable performance.

Acknowledgments

We would like to thank the referees for their comments and suggestions on the manuscript. This work was supported by the Thailand Science Research and Innovation.

References

[1] H. Martin, L. Priscila, The Worlds Technological Capacity to Store, Communicate, and Computer Information, Science. 332 (2011) 6065.

[2] S. Sagiroglu, D. Sinanc, Big data: A review, International Conference on Collaboration Technologies and Systems (CTS). (2013) 42-47.

[3] EMC Education services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Willey, 2015.

[4] M. Tom. Machine Learning, McGraw Hill, New York, 1997.

[5] J. Dean, Big Data, Data Mining, and Machine Learning, John Wiley & Sons, New Jersey, 2014.

[6] B. Boehmke, B. Greenwell, Hands-On Machine Learning with R, CRC Press; Taylor & Francis Group LLC, New York, 2020.

[7] Z. Comert, A.F. Kocamaz, Comparison of Machine Learning Techniques for Fetal Heart Rate Classification, Acta Physica Polonica A. 132 (2017) 451454.

[8] S. Burk, G.D. Miner, Its All Analytics!: The Foundations of AI, Big Data, and Data Science Landscape for Professionals in Healthcare, Business, and Government, Productivity Press, 2020.

[9] V. S. Kublanov, A. Y. Dolganov, D. Belo, H. Gamboa, Comparison of machine learning methods for the arterial hypertension diagnostics. Applied bionics and biomechanics, 2017.

[10] A. S. Sunge, H. L. H. S. Warnar, Y. Heryadi, E. Abdurachman, B. Soewito, F. L. Gaol, Prediction Diabetes Mellitus Using Decision Tree Models. International Congress on Applied Information Technology 4(1), (2019) 189-198.

[11] Y. Religia, G.T. Pranoto, E.D. Santosa, (2020). South German Credit Data Classification Using Random Forest Algorithm to Predict Bank Credit Receipts. JISA: Jurnal Informatika dan Sains, 3(2), (2020) 62-66.

[12] D. Dheeru, G. Casay, UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/bank+marketing], Irvine, CA: University of California, School of Information and Computer Science, 2019.

[13] P. N. Tan, M. Steinbach, V. Kumar, A. Karpatne, Introduction to Data Mining, Pearson Education, New York, 2019.

[14] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, New York, 2014.

[15] I. H. Witten, E. Frank, M. A. Hall, C.Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

[16] H. He, E. A. Garcia, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, 21(9), (2008).

[17] K. Upadhyay, P. Kaur, D.K. Verma, Evaluating the Performance of Data Level Methods Using KEEL Tool to Address Class Imbalance Problem. Arabian Journal for Science and Engineering, (2021) 1-14.

[18] H. Yang, S.J. Fong, Optimized very fast decision tree with balanced classification accuracy and compact tree size, The 3rd International Conference on Data Mining and Intelligent Information Technology Applications, (2011) 57-64.

[19] WEKA. Waikato Environment for Knowledge Analysis Machining Learning Software in Java. Retrieved from website: https://www.cs.waikato.ac.nz/ml/weka/ on 7th March, 2022.

[20] A. Verma, Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA. International Research Journal of Engineering and Technology, 5(13), (2019) 54-60.