p-ISSN : 2443-2210
e-ISSN : 2443-2229

Studi Perbandingan Evaluasi Kinerja Metode Pembelajaran *Eager Learning* versus *Lazy Learning*

http://dx.doi.org/10.28932/jutisi.v10i3.9197

Riwayat Artikel
Received: 24 Juni 2024 | *Final Revision*: 2 September 2024| *Accepted*: 17 September 2024

Selvi Lukman✉ #1, Jimmy Agustian Loekito*2, Pin Panji Yapinus#3

[1] *School of Computer Science and Computer Engineering Department, Universitas Bina Nusantara*
Paskal Hyper Square, Jl. Pasir Kaliki No.25-27, Bandung, 40181, Indonesia
[1]selvi.lukman@binus.ac.id

[2,3] *Computer Engineering Department ,Universitas Kristen Maranatha*
Jl. Surya Sumantri No.65, Bandung , 40164, Indonesia
[2] jimmy.al@eng.maranatha.edu
[3]pin.py@eng.maranatha.edu
✉selvi.lukman@binus.ac.id

Abstrak — **Sebagian besar pendapatan utama sektor perbankan berasal dari nasabah simpanan jangka panjang. Banyak strategi pemasaran telah diterapkan untuk mengkaji karakteristik pengambilan keputusan para nasabah. Dalam hal ini, pembelajaran mesin sebagai cabang ilmu komputasi ilmiah dipergunakan untuk menemukan nasabah potensial yang terbaik terutama dalam melakukan prediksi untuk pengambilan deposito jangka panjang yang akan dilakukan nasabah. Pada penelitian ini dilakukan studi komparasi antara lazy learning dalam bentuk Random Forest (RF) dan easy learning dalam bentuk K-Nearest Neighbors (KNN). Perbedaan di antara keduanya bergantung pada konsep dari sifat komputasi pemrogramannya. Dalam penelitian studi kasus perbankan ini terbukti RF lebih unggul dibandingkan KNN dari segi akurasi yang mencapai 96%, presisi 93% dan F1-score 0.97. Oleh karena itu, performa terbaik RF telah dicapai berdasarkan kemampuannya untuk menangani problem non-linearitas dan ketahanan terhadap overfitting sehingga RF layak diterapkan untuk banyak aplikasi prediktif.**

Kata Kunci—Deposito Berjangka; Easy learning; Klasifikasi; Lazy Learning

Comparison Study on Performance Evaluation of Eager versus Lazy Learning Methods

Abstract — **The major revenue in the sector of banking is mostly initiated from long term deposit customers. Many strategies in marketing have been implemented for potential customers by examining their impacted characteristics in decision making. Therefore, machine learning as a scientific computing has drawn many interest in finding best potential customers especially in predicting whether a long term deposit should be subscribed or not. In this research, lazy and eager learning of Random Forest (RF) and K-Nearest Neighbours (KNN) is compared. The sharp distinction between them relies on the programming computation concepts. In this case study, RF is proven to be more superior than KNN in the term of Accuracy as much as 96%, Precision 93% and F1 score 0.97. Therefore, the ultimate performance of RF has been achieved based on its ability to handle non-linearities and resistance to overfitting. Hence it makes RF a suitable choice for many predictive applications.**

Keywords— Classification; Easy learning; Lazy Learning; Term Deposit

I. INTRODUCTION

As the tendency of exponential growth on interest rests, growing savings are expected to earn more than the average interest rate. Traditional savings accounts may not offer optimal interest rates and options like stocks involve a lot of risk. Therefore, for those who prefer a safe financial product, a term deposit account is preferable. A term deposit defines a fixed term investment of a financial institution. Term deposit investments usually sustain short-term maturities ranging from one month to some few years. The deposit provides many levels of required minimum amount and investors should understand that funding withdrawal can be performed after the term has come to the end. In some occurrences, early termination or withdrawal is allowable with prior notification and early termination will cause penalties. Commonly, an account holder deposits some funds at the bank, those funds are addressed to other businesses or consumers. The compensation is in the form of account balanced interest [1]. The fund holder is able to withdraw their money at any time. It leads to a difficulty in knowing priorly how much they may lend. Figure 1 depicts a common input and output cash flow in the bank.

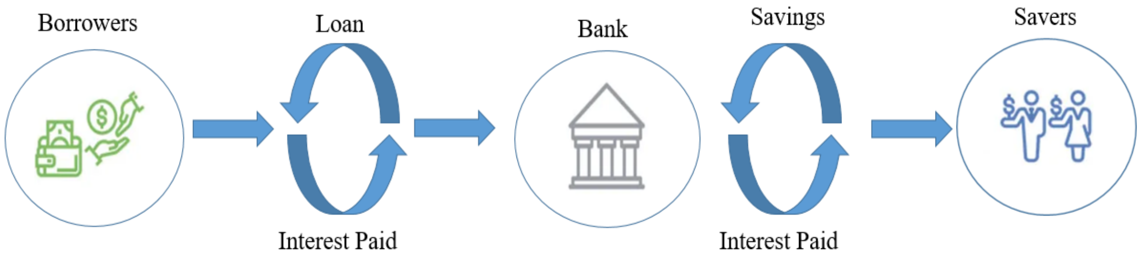


Figure 1. The common bank cash flow.

Term deposit is conducted in order to mitigate funding withdrawal in a fixed and certain period. A higher rate of interest is expected to be paid in return as the earned interest is usually slightly higher than what was paid on interest in standard savings or bearing checking accounts. However, increasing rates are caused by the money limitation in their time frame. When customers agree to place some funds in a term deposit, banks can locate their fund in other financial products with higher Rate of Return (RoR). Banks can also redirect the money out to their other clients in order to receive a higher interest rate from borrowers in which they will be compared to what banks are paying for the term deposit. The spread between the rate that banks usually deposit for their customers and the rate that banks charge to their borrowers is the net interest margin. Accordingly, net interest margin defines a banking profitability metric.

As banks are institutions that provide the possible term deposit lowest rate, loan fees should be higher to borrowers. This action is meant to increase their profitability or margins. Accordingly, there is a required balance that the bank requires to maintain. When the interest is too small, it will fail in attracting new venture capitalists into the term deposit accounts. However, when they charge a higher rate on loans, it will fail in attracting new borrowers. Therefore, improving term deposit marketing through marketing campaigns will enhance the promotion of term deposits to customers. Nowadays, many strategies are implemented for a simple automation or even a better decision of return investment [2].

Furthermore, large datasets are required to predict bank customers to subscribe to a term deposit after being contacted. In this study, classification algorithms in the machine learning field are implemented in the marketing campaigns in which the result will be presented in a binary of “Yes“ or “No“. Regarding the type of datasets, classification algorithms in machine learning describe the process of understanding and object grouping into categories. The utilized data in this research was obtained from the web https://archive.ics.uci.edu/ml/datasets/Bank+Marketing or site UCI Machine Learning Repository, entitled "Banking Marketing". The utilized data used is a primary dataset with a total of 17 variables consisting of age, education, job, material, housing, contact, loan, marital status, balance, default, duration, p-day, month, campaign, previous, p-outcome and 4521 records. The dataset was investigated by a banking institution in Portuguese for their marketing campaigns which depend on phone calls. Many contacts for the same client were needed in order to retrieve the bank term deposit subscription. Therefore, the classification algorithm of eager and lazy learning is aimed for a better subscription prediction or in this case is the output of variable y. For a clear explanation, Figure 2 presents the concept of classification algorithms in machine learning.

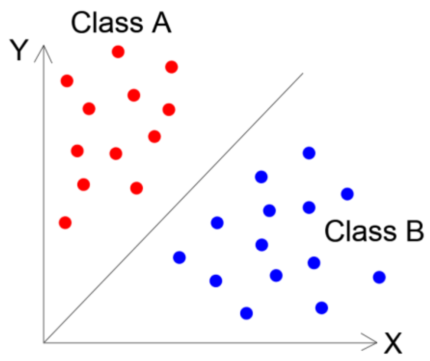


Figure 2. The concept of classification algorithm

This study presents a comparison of eager and lazy learning. The terminology of eager and easy learning refers to the generalization of training data in which lazy learning does not generalize until and unless it is needed. Whereas, eager learning implies instant learning as they also generalize before seeing the query. These distinctions impact the adaptability to new dataset, efficiency of learning and flexibility of the resulting models. In its early development, eager learning faced some challenges of real time prediction and large datasets that lead to the lazy learning procedure in addressing the prior issues [3][4]. An insight about eager and lazy methods is depicted in Figure 3.

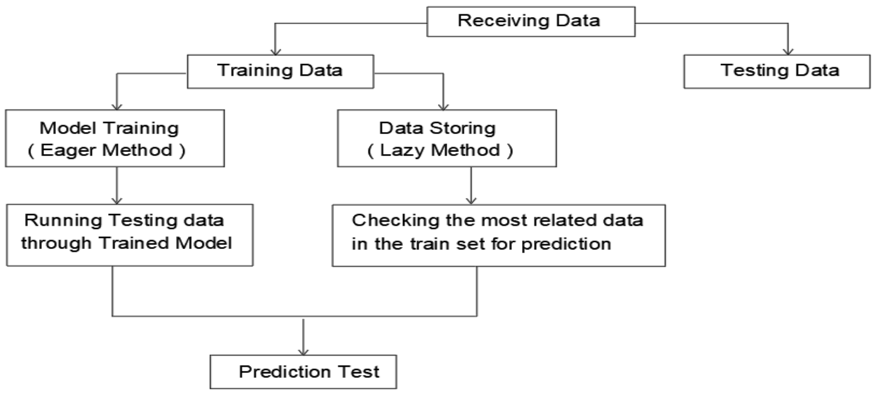


Figure 3. The distinction between eager and lazy method

Choosing the best classification method requires a deep investigation of being eager or being lazy learning algorithms, since their generalization behaviour will affect the performance of memory, miscalculation, overhead introduction and dependencies that can hinder parallelism [4][5]. Therefore, this study attempts a comparison evaluation of a lazy and eager classification method upon datasets of term deposit for a reliable bank campaign as the choice between them is not easy. A single rule set to conduct predictions tasks utilization may not take advantage of specific test instance characteristics. Meanwhile, establishing a special rule set for every test instance may result in excessive efforts in computational procedures [6]. For improving the efficiency of the banking domain industry, an investigation of utilizing classification models used a test of classifying ability for the learned model of Term Frequency Inverse Document Frequency (TF-IDF) and this study has provided expected feedback from customers [7].

Another research has contributed classification methods to identify categorized customers who provide a satisfying probability in subscribing long term deposits using telemarketing datasets. However, this research should attempt a sort of lazy learning model to do the comparison since eager learning would likely produce the similar results. Improving the

Success Rate of Bank Direct Marketing Campaign has been studied by other authors and GLM (Generalized Linear model), R-Part Tree and rule based algorithms were utilized in determining the important characteristics that will give an impact on selection of potential buying customers [8]. The performance between them were compared and it can be shown how R-Part Tree and Rule based algorithms were superior compared to GLM. Therefore, a generalization procedure must be conducted when training the dataset even if the model has provided a generalized behaviour. A well-adapted Class Membership-Based (CMB) classifier bank telemarketing campaign was implemented in building an optimal model for predicting potential customers in their campaign launching. The model can predict prospective future customers more precisely than the previous works. Not a surprise that this model exploited nominal variables in decision function. However, when memory usage is pivotal during prediction, choosing to be lazy and eager might be a great help to achieve its performance [9]. More information topic upon prediction of customer willingness for long term deposit using Neural Network was also analysed and it was observed that increasing network size would not increase the accuracy [10]. However, those existing literatures have successfully predicted the long term deposit for customer willingness in banking industries. More research is required to understand the importance of choosing classification algorithms especially when optimization is involved. Lazy iteration is more complicated to understand than eager iteration as a thorough navigation consists of a labyrinth of methods that must be conducted in order to follow the path of execution [11].

The K-Nearest Neighbours (KNN) algorithm is a well-known machine learning method that is commonly utilized for regression and classification tasks. The idea is based on the same data points which provide a tendency on similar values or labels. KNN algorithm stores the entire training dataset as a reference during the training phases [12]. Furthermore, Figure 4 depicts the flowchart of a simple KNN [13].

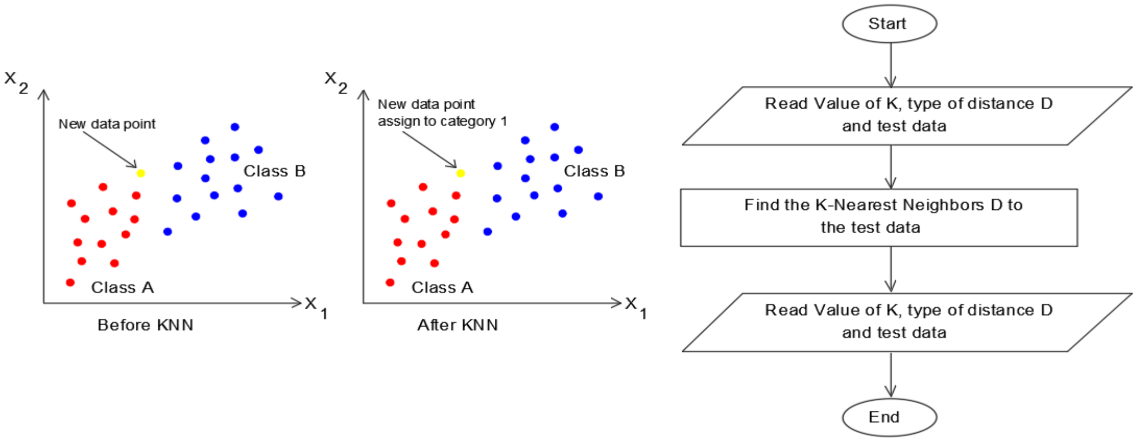


Figure 4. The concept of a simple KNN.

In a simple KNN as depicted in Figure 4, the number of K neighbours are selected before calculating their distances using either Euclidean distance or Manhattan distance. Euclidean measures the straight-line distance between two points and this procedure is mostly suitable for continuous features. Meanwhile, Manhattan distance or L1 normalization calculates the sum of absolute differences and it is suitable for high-dimensional data [14]. Among these K neighbours, the number of points is owned by each category before it is calculated, therefore the new point can be assigned to the most prevalent category. KNN as a lazy learning algorithm does not construct generalization from the training data or explicit model. Moreover, the entire training data will be stored and performed the computation when a new query is available. Accordingly, KNN is adaptable to changes in the dataset as it is able to handle nonlinear and noisy data. Another advantage of KNN is it does not make any assumptions about underlying structure of data or distribution unlike other algorithms that require the datasets to be linear, normal, or independent. However, its main drawback is the requirement of computational resources, loads of memory to process and store the overall training dataset. When the data dimensionality is increasing, the distance is calculated and the searching path for the nearest neighbours will become more time consuming and expensive. This limitation leads to inefficiency of KNN, especially in high-dimensional datasets. Furthermore, KNN can be sensitive to redundant or irrelevant features as it can directly affect the distance measurement and the classification accuracy.

On the other hand, Random Forest (RF) as an eager algorithm works especially effective when handling the well-labelled and structured datasets or whereas the relationships between labels and features is intelligible. A simple Random forest is depicted in Figure 5.

p-ISSN : 2443-2210
e-ISSN : 2443-2229

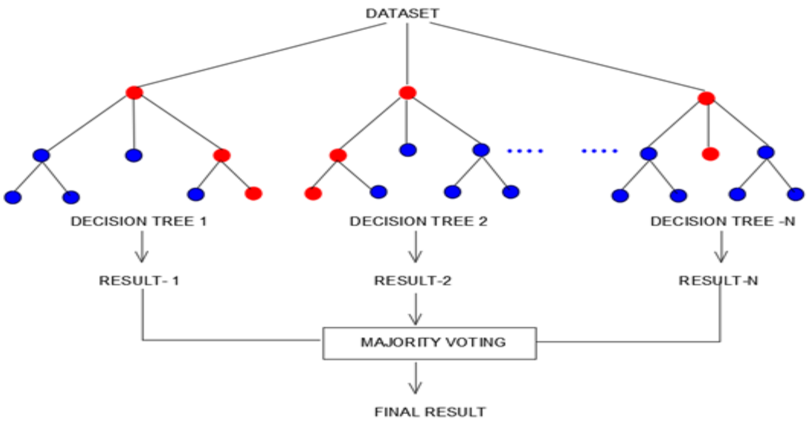


Figure 5. The concept of a simple Random Forest

Eager Learning models such as Random Forest can accurately and quickly regress or classify new instances as they do not consume the whole training dataset during prediction. Predictions can be conducted without continuous access to the training data or even offline. Therefore, this feature is beneficial for scenarios in isolated environments or limited connectivity. Optimization during the training phase is also possible as it can lead to an improved performance. Hence, it will yield a well adapted dataset with intelligible relationships and patterns. Furthermore, as they do not depend on training data during prediction. They are also simple in deploying facilities to integrate the performance to various systems and interpretation [15].

The comparison of KNN as lazy learning and RF as eager learning provides many pros and cons especially in terms of interpretability. During training procedures, the models can be observed when they are making decisions. However, when the dataset is small, eager learning might be useful. Several tasks in eager learning models can be transferred more easily than lazy learning. Unlike the training procedure in lazy learning which involves predictions search for the same instances, eager learning does not need additional overhead. This is beneficial as it leads to faster processing. In terms of choosing the right method of classification, this work specifically concentrates on using two kinds of classification methods: Random Forest (RF) and K-Nearest Neighbours (KNN) to predict whether customers are likely to subscribe to a term deposit or not. This system is used to select the right customers for marketing campaigns. Hence, the efficiency of the marketing efforts will be enhanced as well. This work is structured as follows. Section II analyzes the dataset and proceeds to Section III in which experimental results and discussion are presented. Finally, Section IV gives conclusions and future works.

II. RESEARCH METHOD

The utilized dataset in this research is a secondary dataset of a campaign in direct marketing of in Portuguese Banking Institution. The campaigns were conducted based on marketing phone calls which required more contacts for the same client. In order to provide the information whether the product of bank term deposit should be subscribed or not subscribed, classification methods are implemented to predict whether the client will subscribe a term deposit in a form of binary output [16]. In this research, dataset collection is performed to gather representative and comprehensive real scenarios. The utilized dataset should be relevant and free from error, therefore the dataset is thoroughly checked before doing label encoding from categorical data to numerical data. Data pre-processing estimates scales and translates all features individually by giving range on the training set. Data splitting is mostly pivotal in the pipeline of every modelling task to support a fair evaluation of a predictive model. Grid search and random search are also conducted in this study. The distinction between them lies on the availability of prior knowledge, ranging from hyper parameters and systematic schemes. In this section, a flowchart of this research is presented in Figure 6.

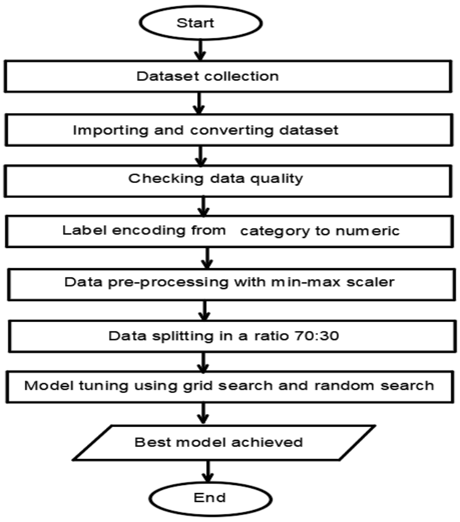


Figure 6. The flowchart of data analysis

In this research, Python Language is utilized. After libraries from Python are summoned, Null-checks are implemented to make the code readable by clearly indicating the considered possibility of a value. In a relational database, a null value is utilized when the value in a column is missing or unknown. A null describes neither an empty string for character or date time data types. Data types checking are conducted to confirm if data entered provides a correct data type, otherwise it will be rejected by the system. Data counting as depicted in Figure 7 is aimed to understand the attributes and features, hence the distribution data can be easily understood. Specific features and attributes can support trends identification, anomalies detection and compute probabilities in statistical analysis. Among 4521 campaigns in this research, the dataset is counted as illustrated in Figure 7. As depicted in Figure 7, the injected dataset in this research consists of the customer's marital status, customer's housing loan and the p-outcome from the previous marketing campaign. Three classes of marital status of married, single and divorce are counted in the datasets. In the outcomes of the previous marketing campaign, the result can only be 4 values of success, failure, unknown and others. Accordingly, the default variable is recorded to be unknown, since it may be possible that customers are not disclosing this information to the banking representative institutions. Meanwhile, the housing dataset is counted whether the customers have housing loans or not.

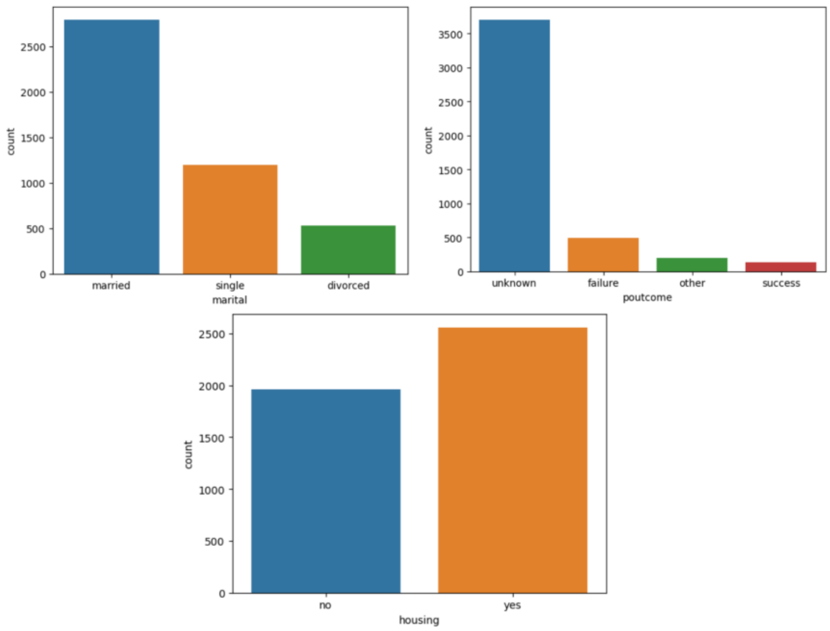


Figure 7. Dataset Counting

p-ISSN : 2443-2210
e-ISSN : 2443-2229

The further process is the label encoding. It is a simple method that is utilized to transform the categorical columns into numerical ones so that they can be summoned to machine learning. Determining variables whether they are dependent or independent are also pivotal in order to add some knowledge in data understanding. This process is followed by a resampling procedure as duplicate random records from the minority class will cause overfishing [17][18]. Data fishing or dredging describes analyses that are conducted without predefined research questions. Before the dataset is divided in a ratio of 70:30, Min-Max Scalers are utilized for feature scaling so data is in the range of [0, 1]. It is mainly beneficial when the distribution of the data is not Gaussian or the relationship between variable values should be values [19].

The performance of the classification model is usually presented in a confusion matrix that contains the prediction of a categorical label for each input instance. The confusion matrix is meant to display numbers of inaccurate and accurate instances in the prediction model. The confusion matrix supports a comprehensible analysis of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions. The calculation of them will facilitate a more profound knowledge of a model's accuracy, recall, precision as notified in (1), (2), (3) and (4) for the entire effectiveness of class distinction [20][21]. This matrix is also helpful in handling an uneven class distribution in a dataset especially for model's evaluation beyond basic accuracy metrics. The terminology number of instances can be classified as follows.

- True positives (TP) defines the prediction of an accurate positive data point as notified in (1)
- True negatives (TN) defines the prediction of an accurate negative data point
- False positives (FP) defines the prediction of inaccurate positive data point
- False negatives (FN) defines the prediction of inaccurate negative data point

Performance Metrics based on Confusion Matrix is illustrated as follow:

$$Accuracy = \frac{TP}{TP + TN + FP + FN} \quad (1)$$

$$\Pr ecision = \frac{TP}{TP + FP} \quad (2)$$

$$\operatorname{Re} call = \frac{TP}{TP + FN} \quad (3)$$

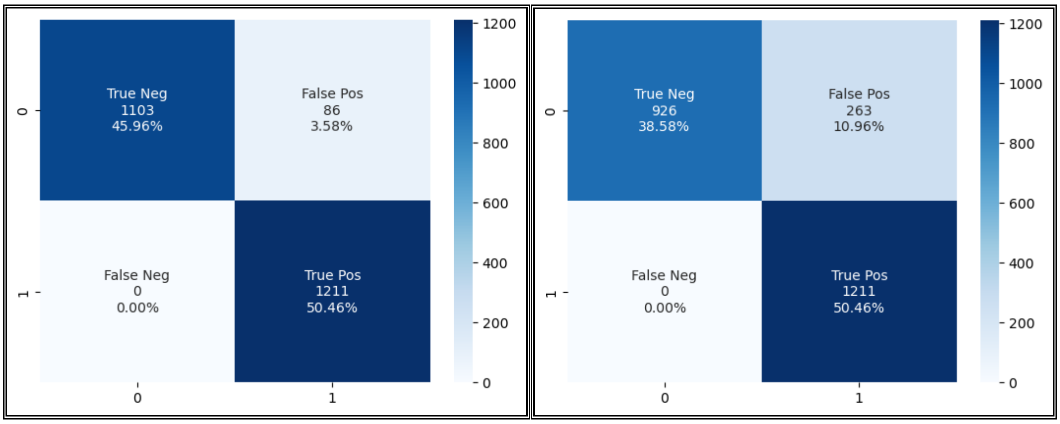
$$F1 - score = \frac{2.precision.recall}{\Pr ecision + \operatorname{Re} call} \quad (4)$$

$$Specifity = \frac{TN}{TN + FP} \quad (5)$$

The confusion matrix is an essential instrument for evaluating the effectiveness of classification models. Confusion matrices can be implemented for either binary and multi class classification scenarios. Not to forget to mention that well-informed decisions regarding model performance are helpful particularly when dealing with imbalanced class distribution.

III. Results and Discussion

Lazy learning and eager learning are two contrasting methods in machine learning that refers to the handling of model prediction and construction. This study investigates the comparison of KNN as lazy learning and RF as eager learning. The distinction between them refers to the computation of prediction and instance specific information of the necessity of pre-computing the model during the training procedures. Choosing between eager and lazy learning relies on many factors such as the dataset size, adaptability to new data, trade-off between memory usage and prediction speed and adaptability to new data. The confusion matrix as depicted in Figure 8 presents the performance of KNN as lazy learning and RF as eager learning accordingly.



RF algorithm

KNN algorithm

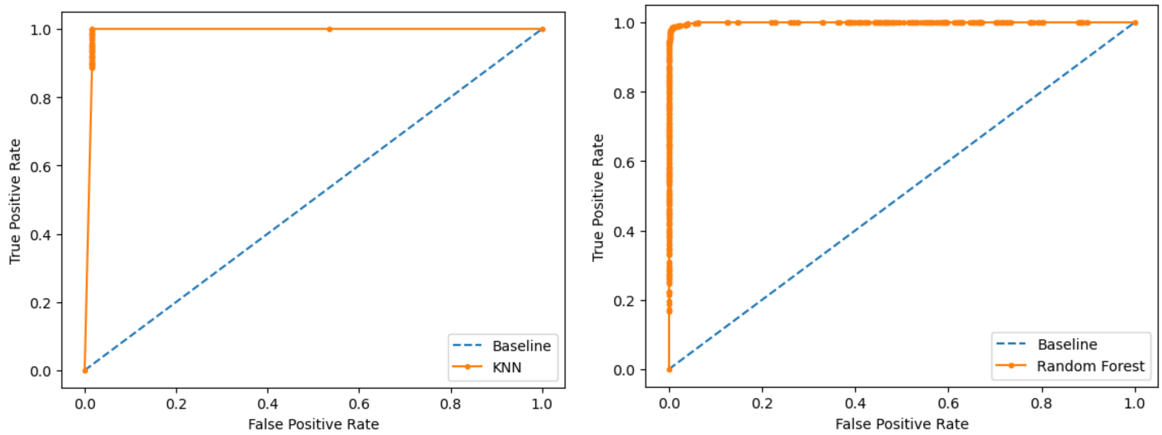
Figure 8 The Comparison of Confusion Matrix between RF and KNN

Figure 7 depicts that KNN has yielded true negative values as much as 38.58%, meanwhile RF yields true negative values as much as 45.96%. A true negative describes an outcome when the model correctly predicts the negative class. As depicted in Figure 7, KNN has resulted in true positive values as much as 50.46 % in which RF has resulted in the same values as well. A true positive describes an end-result where the model precisely predicts the positive class. Furthermore, false positive values in KNN are less than false positives in RF even though they both produce none false negatives. The terminology of false positives indicates an end-result where the model imprecisely predicts the positive class, meanwhile a false negative refers to an end-result where the model imprecisely predicts the negative class. For a better insight, Table 1 presents a comparison performance metric based on the confusion matrix as follows.

TABLE 1
PERFORMANCE METRICS BASED ON CONFUSION MATRIX

Accuracy	KNN	0.89
	RF	0.96
Precision	KNN	0.82
	RF	0.93
Recall	KNN	1.00
	RF	1.00
F1-Score	KNN	0.90
	RF	0.97

As can be observed in Table 1, accuracy answers the question on how often the model is right and the higher the accuracy, the better the model will be. Accuracy is helpful when the case is dealing with balanced classes. In Table 1, the achieved accuracy of KNN is 89 % and RF achieves higher as much as 96%. Meanwhile precision measures true positive predictions within all positive predictions provided by the model in which Table 1 shows that KNN achieves 82% precision values and RF achieves 93%. Recall shows another dimension of the model quality. Recall is usually called True Positive Rate (TPR) and sensitivity, the perfect recall number is 1.0 as it indicates the capability of the model to find all instances of the target class in the dataset. Accordingly, both KNN and RF have achieved the perfect recall number. Furthermore, F1 score is a measure of the harmonic mean of recall and precision as it is mostly used for binary and multi-class classification. An F1 score closer to 1 implies a better model performance, therefore the FI score for KNN 0.90 and RF 0.97 indicate that the model correctly identifies all positives without any false positives or negatives. This research also presents the curve of ROC (Receiver Operating Characteristic) and AUC (Area Under Curve). A probability curve is presented by ROC and the measure or degree of separability is presented in AUC. The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis, therefore the greater the AUC, the better the model in predicting classes. For a better visualisation of those curves, Figure 9 depicts the ROC and AUC as follows.



KNN algorithm

RF algorithm

Figure 9. The Comparison of ROC and AUC curves of KNN and RF.

As observed in Figure 9, it is noticeable that the closer those curves to the graph upper left corner, the greater accuracy test. For any diagnostic technique, the AUC must be higher than 0.5 generally and it must be higher than 0.8 to be acknowledged as acceptable. Accordingly, the ROC curve with the highest AUC is comprehended to have a better diagnostic performance.

IV. Conclusion

This study investigates a comparison of performance evaluation between eager learning and lazy learning method to be utilized in a term deposit classification in a bank marketing campaign. The machine learning approach is proven to be reliable to streamline the lending process as it involves data analytics to evaluate risks, creditworthiness and other factors that determine the probability of a borrower defaulting on their loan. The option between using eager learning and lazy learning is pivotal in shaping immediate resource availability that lead to prediction performance. Therefore, in this research KNN as a lazy learning and RF as an eager learning is compared. The result has shown that RF is more superior than KNN in terms of Accuracy as much as 96%, Precision 93% and F1 score 0.97. The strength of RF lies in creating different trees with distinct sub-features from the actual features, hence making it easy to provide a well-learned prediction.

References

[1] C. S. T. Koumetio, W. Cherif, and S. Hassan, “Optimizing the prediction of telemarketing target calls by a classification technique,” *Proc. - 2018 Int. Conf. Wirel. Networks Mob. Commun. WINCOM 2018*, 2018, doi: 10.1109/WINCOM.2018.8629675.

[2] M. A. T. Rony, M. M. Hassan, E. Ahmed, A. Karim, S. Azam, and D. S. A. A. Reza, “Identifying Long-Term Deposit Customers: A Machine Learning Approach,” *2nd Int. Informatics Softw. Eng. Conf. IISEC 2021*, pp. 1–6, 2021, doi: 10.1109/IISEC54230.2021.9672452.

[3] S. C. Bellad *et al.*, “Prostate cancer prognosis-a comparative approach using machine learning techniques,” *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, no. Iciccs, pp. 1722–1728, 2021, doi: 10.1109/ICICCS51141.2021.9432173.

[4] G. S. Fentie, A. D. Alemu, and B. S. D. Mohapatra, “International Journal of Computer Science and Mobile Computing A Comparative Study on Performance Evalution of Eager versus Lazy Learning Methods,” *Int. J. Comput. Sci. Mob. Comput.*, vol. 3, no. 3, pp. 562–568, 2014, [Online]. Available: www.ijcsmc.com

[5] I. M. Galv´an, J. M. Valls, M. Garc´ıa, and P. Isasi, “A Lazy Learning Approach for Building Classification Models,” *Int. J. Intell. Syst.*, vol. 29, no. 2, pp. 495–524, 2014, doi: 10.1002/int.

[6] S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, “Customer profiling using classification approach for bank telemarketing,” *Int. J. Informatics Vis.*, vol. 1, no. 4–2, pp. 214–217, 2017, doi: 10.30630/joiv.1.4-2.68.

[7] N. G. Cholli, “Machine Learning Classification Models for Banking Domain,” *SSRN Electron. J.*, pp. 2442–2447, 2019, doi: 10.2139/ssrn.3363076.

[8] R. Vaidehi, “Predictive Modeling to Improve Success Rate of Bank Direct Marketing Campaign,” *Int. J. Manag. Bus. Stud.*, vol. 6, no. 1, pp. 22–24, 2016.

[9] S. C. K. Tékouabou, Ş. C. Gherghina, H. Toulni, P. N. Mata, M. N. Mata, and J. M. Martins, “A Machine Learning Framework towards Bank Telemarketing Prediction,” *J. Risk Financ. Manag.*, vol. 15, no. 6, pp. 1–19, 2022, doi: 10.3390/jrfm15060269.

[10] R. Patel and H. Mazumdar, “Prediction of Bank Investors using Neural Network in Direct Marketing,” *Int. J. Eng. Appl. Sci.*, vol. 5, no. 2, p. 257283, 2018.

[11] H. Homayouni, S. Hashemi, and A. Hamzeh, “A Lazy Ensemble Learning Method to Classification,” *Int. J. Comput. Sci. Issues*, vol. 7, no. 5, pp.

344–349, 2010.
[12] Y. Cai, D. Ji, and D. Cai, "A KNN Research Paper Classification Method Based on Shared Nearest Neighbor," *Int. J. Comput. Sci. Issues*, pp. 336–340, 2010.
[13] J. D. De Guia, R. S. Concepcion, A. A. Bandala, and E. P. Dadios, "Performance Comparison of Classification Algorithms for Diagnosing Chronic Kidney Disease," *2019 IEEE 11th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag. HNICEM 2019*, 2019, doi: 10.1109/HNICEM48295.2019.9073568.
[14] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019, doi: 10.30591/jpit.v4i1.1253.
[15] Y. Zhang, J. Wen, G. Yang, Z. He, and J. Wang, "Path loss prediction based on machine learning: Principle, method, and data expansion," *Appl. Sci.*, vol. 9, no. 9, 2019, doi: 10.3390/app9091908.
[16] K. H. Kim, C. S. Lee, S. M. Jo, and S. B. Cho, "Predicting the success of bank telemarketing using deep convolutional neural network," *Proc. 2015 7th Int. Conf. Soft Comput. Pattern Recognition, SoCPaR 2015*, pp. 314–317, 2016, doi: 10.1109/SOCPAR.2015.7492828.
[17] F. Ma, Y. Li, S. Ni, S. Huang, and L. Zhang, "Data Augmentation for Audio–Visual Emotion Recognition with an Efficient Multimodal Conditional GAN," *Appl. Sci.*, vol. 12, no. 1, 2022, doi: 10.3390/app12010527.
[18] S. G. Alonso *et al.*, "Data Mining Algorithms and Techniques in Mental Health: A Systematic Review," *J. Med. Syst.*, vol. 42, no. 9, 2018, doi: 10.1007/s10916-018-1018-2.
[19] A. Jiménez-Cordero and S. Maldonado, "Automatic feature scaling and selection for support vector machine classification with functional data," *Appl. Intell.*, vol. 51, no. 1, pp. 161–184, 2021, doi: 10.1007/s10489-020-01765-6.
[20] S. Lukman, Y. Y. Nazaruddin, B. Ai, and E. Joelianto, "Path loss measurement validation using naïve bayes classification for high-speed rail," *Int. J. Eng. Appl.*, vol. 9, no. 4, pp. 226–233, 2021, doi: 10.15866/irea.v9i4.20414.
[21] S. Visa, B. Ramsay, A. Ralescu, and E. van deer Knaap, "Confusion Matrix-based Feature Selection Sofia Visa," *ConfusionMatrix-based Featur. Sel. Sofia*, vol. 710, no. January, p. 8, 2011.