International Journal of Computer Applications (0975 – 8887)
Volume 186 – No.16, April 2024

Performance Comparison Random Forest and Logistic Regression in Predicting Time Deposit Customers with Feature Selection

Reski Noviana
Master Program of Information Technology
Universitas Teknologi Yogyakarta
Sleman, Yogyakarta, Indonesia

Enny Itje Sela
Master Program of Information Technology
Universitas Teknologi Yogyakarta
Sleman, Yogyakarta, Indonesia

ABSTRACT
Machine learning algorithms can be used to analyze data and predict customer behavior. One important aspect in developing machine learning models is feature selection. Proper feature selection can significantly affect model performance. Irrelevant or redundant features can impair the performance of the model and increase its complexity. Therefore, feature selection is an important stage in building an effective prediction model. The main objective of this research is to compare the performance of Random Forest and Logistic Regression in predicting customers' decision to subscribe to time deposits. In addition, this research also includes the use of feature selection using Forward Selection and Recursive Feature Extraction (RFE) to ensure only relevant features are used in the model. The overall results show that the use of Forward Selection and Recursive Feature Elimination (RFE) feature selection also affects the accuracy value. In this study, the best accuracy was obtained by the first scenario, namely Radom Forest and Logistic Regression classification without using selection features but the target class has been balanced using the SMOTE method, resulting in the best accuracy of Random Forest 95.56%, and 96% for precision, recall and f1 score. While Logistic Regression 87.21% and 87% for precision, recall and f1 score. Then when using the feature selection scenario there is a decrease in accuracy for Random Forest by 3.39% when using Forward Selection and 0.33% when using RFE. While Logistic Regression there is a decrease in accuracy of 1.87% when using Forward Selection and 0.22% when using RFE. Further research can deepen the influence of parameters on classification models that can provide further information to improve model performance.

General Terms
Data Mining, Classification, Machine Learning

Keywords
Random Forest, Logistic Regression, Deposit Customers, Feature Selection, Forward Selection, Recursive Feature Elimination

1. INTRODUCTION
One of the key challenges in the banking industry is predicting customer behavior, especially in terms of their decision to subscribe to certain products or services, such as time deposits. Predicting this behavior has a direct impact on marketing strategies, financial planning, and risk management in banks [1]. Machine learning algorithms have become an especially useful tool in analyzing data and predicting customer behavior [2]. One important aspect in the development of machine learning models is feature selection [3]. Proper feature selection can significantly affect model performance. Irrelevant or redundant features can impair the performance of the model and increase its complexity [4]. Therefore, feature selection is an important stage in building an effective prediction model. Previous research on the prediction of time deposit bank customers is by [5] comparing Decision Tree and Random Forest algorithms on Guimarães Bank of Portugal data showed that registered customer features have a meaningful relationship with their future decisions in the banking system, such as opening long-term deposits. Decision tree and Random Forest classification methods were used to predict the opening of long-term deposits based on customer features registered in the bank's database, and both methods achieved an accuracy of 90.73% in Random Forest and 88.7% in Decision Tree.
The main objective of this research is to compare the performance of Random Forest and Logistic Regression in predicting customers' decision to subscribe to a term deposit. In addition, this research also includes the use of feature selection using Forward Selection and Recursive Feature Extraction (RFE) to ensure only relevant features are used in the model. The results of this study are expected to provide valuable insights to banks in designing more effective marketing strategies and improving customer retention.

2. LITERATURE REVIEW
2.1 Previous Study
This research is a comparison of the performance of Random Forest and Logistic Regression in predicting time deposit customers with feature selection using Forward Selection and Recursive Feature Elimination (RFE). There are many studies related to the prediction of time deposit customers using machine learning algorithms, but in previous studies by [6] and [5] only compared algorithms for predicting bank customers, then research by [7] This study only uses one feature selection, namely correlation based for predicting telemarketing bank customers for deposits. In this research, the prediction of time deposit customers will be conducted using feature selection by comparing Forward Selection and Recursive Feature Elimination (RFE), then measuring the accuracy performance of the Random Forest and Logistic Regression algorithms before and after using feature selection.

2.2 Feature Selection
Feature selection is an important and frequently used technique in the data pre-processing stage [8]. The goal of feature selection is to find the most informative subset of a high-dimensional dataset by removing redundant and irrelevant features, to improve the classification and prediction accuracy of machine learning models [9]. So, feature selection is important to find relevant features for classification.

2.3 Forward Selection

The step-by-step selection approach starts with no features in the model. At each step, features that have the most impact in improving the model are added, followed by the inclusion of new variables that do not improve model performance [9].

Model: $y = \beta 0 + \beta 1 * x1$ (1)

Where, y is the response variable or dependent variable, x1 is the predictor variable or independent variable, β0 is the intercept, which is the value of y when x1 is equal to zero, β1 is the regression coefficient, which describes the expected change in y when x1 increases by one unit.

2.4 Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a wrapper feature selection method used to reduce the number of features in a dataset by selecting the features that contribute most to improving the performance of the learning model [9]. Unlike Forward Selection and Backward Selection, the feature selection process in RFE is done recursively by starting with all features and iteratively removing the features with the lowest weights. After each iteration, the model is re-evaluated and the feature with the lowest contribution is removed. This process continues until the desired number of features is reached. The RFE equation uses simple linear regression:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 \quad (2)$$

Where, y is the response variable or dependent variable, x1, and x2 are the predictor variables or independent variables, xa is the predictor variable to a. β0 is the intercept, which is the value of y when x1, and x2 are equal to zero, β1, and β2 are the regression coefficients of x1, and x2 which describe the change in y expected when each predictor variable increases by one unit. β3 is the regression coefficient for xa to a.

2.5 Random Forest (RF)

According to [10] Random Forest (RF) is a classification model that is an extension of a single classification tree by applying bootstrap aggregating (bagging) and random feature selection techniques. The way it works involves creating several classification trees in parallel, then prediction results are taken based on most votes. The Random Forest modeling process involves several steps, including a bootstrap process that involves random sampling of the training data with returns, construction of a single classification tree with the training data resulting from the bootstrap process, and random feature selection. In addition, tree building involves random feature selection, where some features are randomly selected and used as separators. This process continues until reaching the minimum size of observations at the nodes. The last step is to repeat this process k times to build k classification trees, and the final classification result is determined based on the majority vote of the k trees.

According to [11] in deciding Tree, it is necessary to calculate entropy and information gain. The equation below is the formula for entropy and information gain:

$$Entropy(Y) = -\sum i\, p(c|Y) log2p(c|Y) \quad (3)$$

Where Y is the set of cases, and $p(c|Y)$ *is* the proportion of Y values to class c.

$$Information(Y) = Entropy(Y) - \sum ve\, Values \frac{|Yv|}{|Ya|} Entropy(Yv) \quad (4)$$

Where Values(a) are all values in the set of cases a, Yv is a subclass of Y with class v corresponding to class a. Yes, are all values corresponding to a.

2.6 Logistic Regression (LR)

Logistic Regression (LR) is a classification algorithm that integrates target variables and prediction variables to produce certain output probabilities [12]. Logistic Regression is a variant of regression specifically designed to classify data with two prediction groups, true and false. The dependent variable (Y) in this model depends on the independent variables (X1, X2, etc.) that influence its change. Thus, the Logistic Regression model can be explained as follows [13]:

$$Y = X1 + X2 + X3 + \cdots + Xn \quad (5)$$

Where Y is the response or dependent variable, $X1, X2, X3$ are independent variables and Xn is the nth independent variable.

The mathematical formula for the logistic regression model, with $\sigma(t)$ as the logistic function that describes the adoption of sigmoid activation.

$$\sigma(t) = \frac{1}{1 + e^{-t}} \quad (6)$$

Where σ(t) is the logistic function, t is the input or argument of the sigmoid function, e is the Euler number, which is a mathematical constant with a value of about 2.71828.

3. Research Methods

The research stages include the steps that will be taken in this research. This research was carried out in a structured manner in accordance with the compiled research stages. The following figure is the stage of this research:

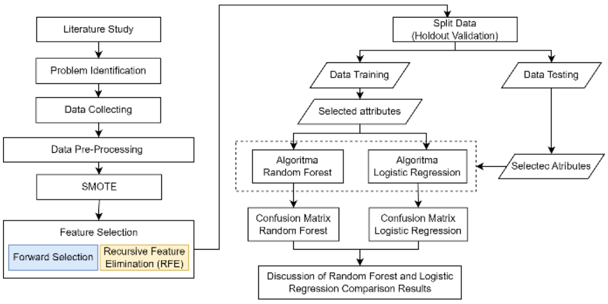

Fig 1: Research Stages

The first stage in this research is a literature study, which involves searching and analyzing relevant literature related to predicting deposit customers using machine learning algorithms, before conducting research. The second step is the identification of the problem to be solved or researched in the study. This helps formulate the research objectives and focus of the analysis. This research will identify problems in the classification of deposit customers. The third step is data collection. The data used in this study is public secondary data sourced from the UCI Machine learning Repository "Bank Marketing" from Portuguese Banks, which can be accessed through the link: https://archive.ics.uci.edu/dataset/222/bank+marketing [14]. This data relates to the direct marketing efforts of financial institutions in Portugal conducted through phone calls. In product marketing or campaigns, it is often necessary to have multiple contacts with the same client to determine whether the client is willing to subscribe to a banking product, such as a time deposit ('yes') or not ('no'). The amount of data used in this study is 41188 data with 21 features or attributes. Next, data preprocessing is performed, which includes cleaning, organizing, and preparing the data before entering it into the model. This includes checking or removing missing data, encoding and normalization. After preprocessing, the clean data is ready to be used for the classification process. Since the classes in the dataset are not balanced, class balancing is performed using the SMOTE method. Next, the feature selection process will be conducted using Featured Selection and Recursive Feature Elimination (RFE). The feature selection test results are evaluated using confusion matrix and the test results will be discussed, analyzed, and evaluated.

4. RESULT AND DISCUSSION

4.1 Preprocessing Result

The preprocessing process conducted in this study is the process of checking missing values, removing duplicate data, encoding and normalization. In the results of checking the missing value in the dataset, it is found that there is no missing value, or the data is not empty. In the results of checking duplicate data, there are twelve duplicate data. The duplicate data will be deleted. The initial data before checking the duplicate data is 41188 data, then after the duplicate data is deleted, it becomes 41176 data. The following is the source code and the results of checking duplicate data. The encoding process is a process needed to convert categorical data into numerical data so that it can be processed into the system. In this research using the Ordinal Encoder method. Data normalization needs to be done because in many cases, the variables in the dataset can have different scales or value ranges. Normalization helps to equalize the scale of these variables, which can improve the performance of some machine learning models and statistical analysis. In this research, Standard Scaler is used for the normalization process.

4.2 SMOTE Implementation

The target class in this dataset is unbalanced where the y class has a No class of 36537 and a Yes class of 4639. The dataset imbalance in this research is resolved using the SMOTE method. In handling dataset imbalance using the SMOTE method, this process produces a new dataset that is twice the amount of the previous dataset. Before using SMOTE the dataset amounted to 41176, then after applying SMOTE it became 73074 data. Table 1 Comparison of Data Split Before and After Using SMOTE

Table 1. Comparison of Data Split Before and After Using SMOTE

Split Data (%)	Before SMOTE= 41176			After SMOTE= 73074,				
	Training		Testing	Training			Testing	
	No (0)	Yes (1)	No (0)	Yes (1)	No (0)	Yes (1)	No (0)	Yes (1)
60:40	21957	2748	14580	1891	21894	21950	14643	14587
70:30	25608	3215	10929	1424	25508	25643	11029	10894
75:25	27443	3439	9094	1200	27334	27471	9203	9066
80:20	29272	3668	7265	971	29189	29270	7348	7267
90:10	32892	4166	3645	473	32886	32880	3651	3657

4.3 Result Without Featured Selection

In the first scheme, a classification trial using Random Forest and Logistic Regression without using feature selection is conducted. The following are the results of the classification implementation without using Random Forest and Logistic Regression. In Table 2, are the classification results without using feature selection.

Table 2. Classification Results Without Using Feature Selection.

split data	*Random Forest*				*Logistic Regression*			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
60:40	94.79	95	95	95	87.00	87	87	87
70:30	95.06	95	95	95	81.19	87	87	87
75:25	94.95	95	95	95	86.94	87	87	87
80:20	95.30	95	95	95	**87.21**	**87**	**87**	**87**
90:10	**95.56**	**96**	**96**	**96**	86.86	87	87	87

In the first scenario of the study, the results of the implementation without using feature selection (forward selection and Recursive Feature Elimination (RFE)) on the Random Forest (RF) and Logistic Regression (LR) algorithms. The results in the first scenario are, Random Forest method produces the highest accuracy of 95.65% with precision, recall and f1 score 96%, and is in the 90%:10% data split scheme. Then, Logistic Regression gets the best result of 87.21%, with precision, recall and f1 score 87%, and is in the 75%:25% data split scheme.

4.4 Results of Forward Selection

In the second scenario in this study, the scheme using feature selection using Forward Selection was evaluated. The following in table 3, are the results of the forward selection implementation.

Table 3. Forward Selection Implementation Result

split data	*Random Forest*				*Logistic Regression*			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
60:40	**92.17**	**92**	**92**	**92**	85.27	85	85	85
70:30	90.69	91	91	91	**85.34**	**85**	**85**	**85**
75:25	92.09	92	92	92	85.15	85	85	85
80:20	92.08	92	92	92	85.30	85	85	85
90:10	92.15	92	92	92	85.19	85	85	85

In the second scenario, which uses the Forward Selection feature, the best accuracy of Random Forest is 92.17% with precision, recall and f1 score 92%. While Logistic Regression gets the best accuracy of 85.34% with precision, recall and f1 score 85%. It can be concluded that in this scenario, the accuracy of the Random Forest algorithm is superior to Logistic Regression.

4.5 Results of Recursive Featured Elimination (RFE)

In the third scenario in this study, the scheme using feature selection using RFE was evaluated. The following in table 4, is the result of RFE implementation.

Table 4. RFE Implementation Result

split data	*Random Forest*				*Logistic Regression*			
	accuracy	*precision*	*recall*	*F1*	*accuracy*	*precision*	*recall*	*F1*
60:40	94.49	95	94	94	86.76	87	87	87
70:30	94.70	95	95	95	86.94	87	87	87
75:25	94.64	95	95	95	86.77	87	87	87
80:20	94.78	95	95	95	**86.99**	**87**	**87**	**87**
90:10	**95.23**	**95**	**95**	**95**	86.72	87	87	87

In the third scenario, namely using the RFE selection feature, the best Random Forest accuracy is 95.23% with precision, recall and f1 score 95%. While Logistic Regression gets the best accuracy result of 86.99% with precision, recall and f1 score 87%. It can be concluded that in this scenario, the accuracy of the Random Forest algorithm is superior to Logistic Regression.

4.6 Comparison of Feature Selection Using Forward Selection and RFE and Classification of Random Forest and Logistic Regression

Based on the test results, the best results are obtained when classification using Random Forest with the RFE method has an accuracy of 95.23%, and a value of 95% for precision and recall and F1 score. with a 90%: 10% data split scheme. As for the Forward Selection method, Random Forest has the best accuracy of 92.17% and a value of 92% for precision and recall and F1 score. Then the Logistic Regression method gets the best results when using RFE selection features with the best accuracy of 86.99% and a value of 87% for precision and recall and F1 score. While Logistic Regression using Forward Selection has the best accuracy of 85.34% and a value of 85% for precision and recall and F1 score. The following in Figure 2, is a comparison of the best Random Forest and Logistic Regression when using the RFE selection feature.

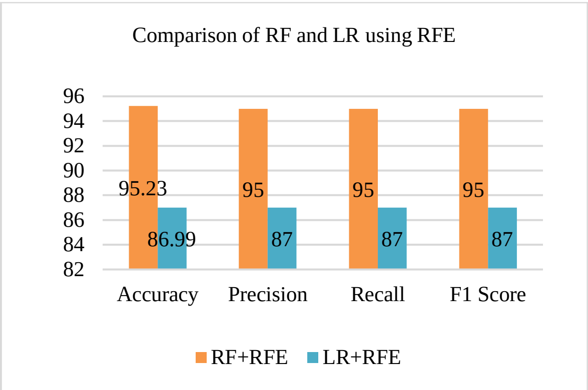

Fig 2. Comparison of Random Forest and Logistic Regression Using RFE

In Figure 2, it shows that Random Forest has accuracy, precision, recall and F1 score values that are superior to the Logistic Regression method for deposit customer classification with the best accuracy result of 95.23%.

4.7 Discussion

In this study, three test scenarios were carried out for the classification of deposit customers, namely by using Random Forest and Logistic Regression classification without using feature selection, Random Forest and Logistic Regression classification using Forward Selection, and Random Forest and Logistic Regression classification using Recursive Feature Elimination (RFE). This research was conducted using the python programming language using Google Collab tools to process the data. In this research, preprocessing is done by checking missing values, removing duplicate data, encoding, and normalization. After that, class imbalance handling is conducted using the SMOTE method, and then split data is divided into training data and test data according to the specified scheme and continued the classification process according to the predetermined scenario.

The research results in the first scenario, namely the results of implementation without using feature selection (forward selection and Recursive Feature Elimination (RFE)) on the Random Forest (RF) and Logistic Regression (LR) algorithms. The results in the first scenario are, Random Forest method produces the highest accuracy of 95.65% with precision, recall and f1 score 96%, and is in the 90%:10% data split scheme. Then, Logistic Regression gets the best result of 87.21%, with precision, recall and f1 score 87%, and is in the 75%:25% data split scheme. In the second scenario, which uses the Forward Selection feature, the best accuracy of Random Forest is 92.17% with precision, recall and f1 score 92%. While Logistic Regression gets the best accuracy of 85.34% with precision, recall and f1 score 85%. It can be concluded that in this scenario, the accuracy of the Random Forest algorithm is superior to Logistic Regression. In the third scenario, namely using the RFE selection feature, the best accuracy of Random Forest is 95.23% with precision, recall and f1 score 95%. While Logistic Regression gets the best accuracy result of 86.99% with precision, recall and f1 score 87%. It can be concluded that in this scenario, the accuracy of the Random Forest algorithm is superior to Logistic Regression. In the third scenario, namely using the RFE selection feature, the best accuracy of Random Forest is 95.23% with precision, recall and f1 score 95%. While Logistic Regression gets the best accuracy result of 86.99% with precision, recall and f1 score 87%. It can be concluded that in this scenario, the accuracy of the Random Forest algorithm is superior to Logistic Regression.

This research is also compared with previous research using the same dataset, namely research by [15] and [16]. The research used the Wrapped Subset Equal method for feature selection and got the best results of 94.39% for Fuzzy and Decision Tree algorithms. Then research by using feature selection also produces the best accuracy for Naïve Bayes and Logistic Regression 91.14%, while the accuracy in the proposed research when using the RFE method with the best accuracy of Random Forest is 95.23%, which distinguishes it from previous research, namely in this study adding class imbalance handling using the SMOTE method which can improve classification accuracy.

The overall research results show that the use of Forward Selection and Recursive Feature Elimination (RFE) selection features also affects the accuracy value. In this study, the best accuracy was obtained by the first scenario, namely Radom Forest and Logistic Regression classification without using selection features but the target class has been balanced using the SMOTE method, resulting in the best accuracy of Random Forest 95.56%, and 96% for precision, recall and f1 score. While Logistic Regression 87.21% and 87% for precision, recall and f1 score. Then when using the feature selection scenario there is a decrease in accuracy for Random Forest by 3.39% when using Forward Selection and 0.33% when using RFE. While Logistic Regression there is a decrease in accuracy of 1.87% when using Forward Selection and 0.22% when using RFE. The comparison of feature selection between Forward Selection and RFE in this study shows that the RFE method is superior to the Forward Selection method. Figures 3 and 4 are a comparison of the test scenario results.

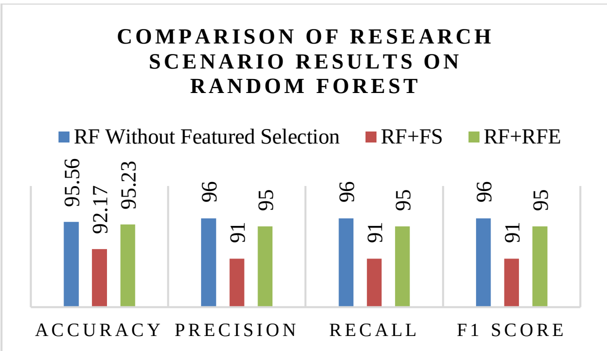

Fig 3. Comparison of Research Scenario Results on Random Forest

In Figure 3, the best accuracy results on Random Forest when using the first scenario, which is without using selection features only Random Forest and SMOTE with an accuracy result of 95.56%.

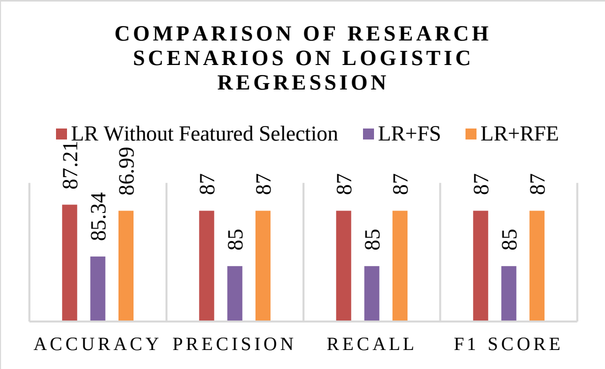

Fig 4. Comparison of Research Scenario Results on Logistic Regression

In Figure 4, the best accuracy results on Logistic Regression when using the first scenario, which is without using selection features only Logistic Regression and SMOTE with an accuracy result of 87.21%.

5. CONCLUSION

Based on tests that have been carried out for the classification of term deposit bank customers, with three test scenarios that have been carried out, the results obtained in the first scenario, namely the implementation results without using feature selection (forward selection and Recursive Feature Elimination (RFE)) on the Random Forest (RF) and Logistic Regression (LR) algorithms. The results in the first scenario are, Random Forest method produces the highest accuracy of 95.65% with precision, recall and f1 score 96%, and is in the 90%:10% data split scheme. Then, Logistic Regression gets the best result of 87.21%, with precision, recall and f1 score 87%, and is in the 75%:25% data split scheme. In the second scenario, which uses the Forward Selection feature, the best accuracy of Random Forest is 92.17% with precision, recall and f1 score 92%. While Logistic Regression gets the best accuracy of 85.34% with precision, recall and f1 score 85%. In the third scenario, namely using the RFE selection feature, the best accuracy of Random Forest is 95.23% with precision, recall and f1 score 95%. While Logistic Regression gets the best accuracy of 86.99% with precision, recall and f1 score 87%. In the third scenario, namely using the RFE selection feature, the best Random Forest accuracy is 95.23% with precision, recall and f1 score 95%. While Logistic Regression gets the best accuracy of 86.99% with precision, recall and f1 score 87%. It can be concluded that in this scenario, the accuracy of the Random Forest algorithm is superior to Logistic Regression. The overall research results show that the use of Forward Selection and Recursive Feature Elimination (RFE) selection features also affects the accuracy value. In this study, the best accuracy was obtained by the first scenario, namely Radom Forest and Logistic Regression classification without using selection features but the target class has been balanced using the SMOTE method, resulting in the best accuracy of Random Forest 95.56%, and 96% for precision, recall and f1 score. While Logistic Regression 87.21% and 87% for precision, recall and f1 score. Then when using the feature selection scenario there is a decrease in accuracy for Random Forest by 3.39% when using Forward Selection and 0.33% when using RFE. While Logistic Regression there is a decrease in accuracy of 1.87% when using Forward Selection and 0.22% when using RFE.

Based on the findings of this study, there are suggestions for further research, including that further research can explore more deeply the optimization of model parameters. Improving the performance of Random Forest and Logistic Regression models can be achieved by optimizing parameters, so further research can deepen the effect of parameters on classification models can provide further information to improve model performance.

6. REFERENCES

[1] R. I. T. Linggadjaya, B. Sitio, and P. Situmorang, "Transformasi Digital Pt Bank Jago Tbk dari Bank Konvensional menjadi Bank Digital," International Journal of Digital Entrepreneurship and Business, vol. 3, no. 1, pp. 9–22, Feb. 2022, doi: 10.52238/ideb.v3i1.76.

[2] R. Pratama, M. I. Herdiansyah, D. Syamsuar, and A. Syazili, "Prediksi Customer Retention Perusahaan Asuransi Menggunakan Machine Learning," Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 12, no. 1, pp. 96–104, Mar. 2023, doi: 10.32736/sisfokom.v12i1.1507.

[3] I. Sulistiani, "Systematic Literature Review: Bankruptcy Prediction Menggunakan Teknik Machine Learning dan Deep Learning," INTECH, vol. 2, no. 1, pp. 13–18, Jun. 2021, doi: 10.54895/intech.v2i1.824.

[4] F. H. Rachman and I. Imamah, "Pendekatan Data Science untuk Mengukur Empati Masyarakat terhadap Pandemi Menggunakan Analisis Sentimen dan Seleksi Fitur," Jurnal Edukasi dan Penelitian Informatika (JEPIN), vol. 8, no. 3, p. 492, Dec. 2022, doi: 10.26418/jp.v8i3.56655.

[5] M. Shahriari and M. H. Asoodeh, "Predicting Long-Term Deposit Openings of Bank Customers Using Decision Tree and Random Forest Classification," KEPES, vol. 19, no. 3, pp. 70–81, 2021, doi: 10.5281/zenodo.7936583#44.

[6] A. A. Aqham and K. D. Hartomo, "Data Mining untuk Nasabah Bank Telemarketing Menggunakan kombinasi Algoritm Naïve Bayes Dan Algoritma Genetik," InfoTekJar (Jurnal Nasional Informatika dan Teknologi Jaringan), vol. 4, no. 1, pp. 47–56, Sep. 2019, doi: 10.30743/infotekjar.v4i1.1574.

[7] A. N. Puteri, A. Arizal, and A. D. Achmad, "Feature Selection Correlation-Based pada Prediksi Nasabah Bank Telemarketing untuk Deposito," MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 20, no. 2, pp. 335–342, May 2021, doi: 10.30812/matrik.v20i2.1183.

[8] D. N. Aini, B. Oktavianti, M. J. Husain, D. A. Sabillah, S. T. Rizaldi, and M. Mustakim, "Seleksi Fitur untuk Prediksi Hasil Produksi Agrikultur pada Algoritma K-Nearest Neighbor (KNN)," Jurnal Sistem Komputer dan Informatika (JSON), vol. 4, no. 1, p. 140, Sep. 2022, doi: 10.30865/json.v4i1.4813.

[9] M. Zivkovic, C. Stoean, A. Chhabra, N. Budimirovic, A. Petrovic, and N. Bacanin, "Novel Improved Salp Swarm Algorithm: An Application for Feature Selection," Sensors, vol. 22, no. 5, p. 1711, Feb. 2022, doi: 10.3390/s22051711.

[10] I. A. Rahmi, F. M. Afendi, and A. Kurnia, "Metode AdaBoost dan Random Forest untuk Prediksi Peserta JKN-KIS yang Menunggak," Jambura Journal of Mathematics, vol. 5, no. 1, pp. 83–94, Jan. 2023, doi: 10.34312/jjom.v5i1.15869.

[11] A. Nugroho, I. Asror, and Y. F. A. Wibowo, “Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma Random Forest,” in e-Proceeding of Engineering, Bandung: Telkom University, Mar. 2023, pp. 1824–1832. Accessed: Nov. 10, 2023. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20030/19395

[12] B. T. P. Briandy, E. Yulianingsih, Fatmasari, and Ferdiansyah, “Analisis Tingkat Akurasi Prediksi Gejala COVID - 19 Dengan Menggunakan Metode Logistic Regression dan Support Vector Machine,” JURNAL FASILKOM, vol. 13, no. 02, pp. 269–278, Aug. 2023, doi: 10.37859/jf.v13i02.5629.

[13] Elina, J. Cristian, V. Louise, S. Koka, and Christnatalis, “Prediksi Keberhasilan Lamaran Pekerjaan Dengan Count Vectorizer dan Logistic Regression,” in Prosiding Seminar Nasional Riset dan Information Science (SENARIS), Pematangsiantar: STIKOM Tunas Bangsa, Apr. 2022, pp. 16–25. Accessed: Oct. 27, 2023. [Online]. Available: http://tunasbangsa.ac.id/seminar/index.php/senaris/article/view/204

[14] S. Moro, P. Rita, and P. Cortez, “Bank Marketing,” UCI Machine Learning, vol. 1, no. 1, Jan. 2012, doi: https://doi.org/10.24432/C5K306.

[15] F. Safarkhani and S. Moro, “Improving the Accuracy of Predicting Bank Depositor’s Behavior Using a Decision Tree,” Applied Sciences, vol. 11, no. 19, p. 9016, Sep. 2021, doi: 10.3390/app11199016.

[16] F. K. Fikriah, “Feature Selection dengan Decision Tree untuk Prediksi Telemarketing Bank,” Jurnal Ilmu Komputer, vol. 15, no. 1, pp. 1–7, Apr. 2022, Accessed: Oct. 27, 2023. [Online]. Available: https://ojs.unud.ac.id/index.php/jik/article/view/74903