
IMPROVING CUSTOMER CHURN PREDICTION WITH GRADIENT BOOSTING MACHINE: A HYBRID APPROACH OF PARAMETER OPTIMIZATION AND DATA BALANCING

Devita Maulidya, Vivi Aida Fitria

Institut Teknologi dan Bisnis Asia, Malang

*Corresponding author: dvtmaulidya@gmail.com

Abstract

Customer churn prediction in the banking industry plays a crucial role in reducing the potential for significant financial losses. By applying machine learning methods, banks can identify patterns and analyze customers who have the potential to churn, thus enabling preventive actions to increase customer loyalty. This research focuses on applying the Gradient Boosting Machine (GBM) algorithm for customer churn prediction, modified with Synthetic Minority Oversampling Technique (SMOTE) in the data preprocessing stage and Randomized Search for parameter optimization. The purpose is to evaluate the effectiveness of the combination of SMOTE and Randomized Search methods in improving the performance of the prediction model. This research uses the CRISP-DM framework as the research framework. The parameter optimization process is performed using Randomized Search, which offers a randomized parameter search approach to optimize accuracy. In the data preprocessing stage, SMOTE was used to handle significant data imbalances to ensure more optimal data quality for the model. The GBM algorithm was tested in two scenarios, with and without the implementation of SMOTE and Randomized Search. Model performance was evaluated using various metrics, including confusion matrix, accuracy, precision, recall, F1-score, and AUC-ROC. The results showed that GBM without Randomized Search and SMOTE resulted in an accuracy of 87.17%. In contrast, the implementation of GBM with Randomized Search and SMOTE increased the accuracy to 90.85%. This study concludes that the implementation of SMOTE and Randomized Search significantly improves the performance of the customer churn prediction model.

Keywords: Customer churn prediction, Gradient boosting machine, Machine learning, Randomized search CV, SMOTE.

INTRODUCTION

Customer churn is the percentage of customers who stop using a business product or service within a certain period of time. In the banking industry, customer churn refers to a condition where customers leave or stop using the services provided by the bank. Addressing customer churn is crucial for a company's success, as it indicates a potential issue with the product or service offered (Irmanda *et al.*, 2019). In addition, technological developments and shifting customer preferences can also be some of the causes of customer churn, making banks have to adapt quickly to maintain their customer base. The phenomenon of customer churn can cause significant financial losses, so an effective strategy is needed to predict customer churn. In an effort to predict customer churn, data analysis methods can help banks in analyzing customers who have the potential to churn. One of them is by utilizing machine learning methods. By using machine learning methods that produce output in the form of prediction results, banks can be helped in analyzing which customers have the potential to churn, and classification methods with output in the form of labels are an appropriate type of machine learning (Farid Naufal *et al.*, 2023). In addition, machine learning also allows banks

to extract information from historical data and identify certain patterns that may indicate potential customer churn. Accurate predictions allow banks to take preventive actions, such as providing special offers or improving services, to increase customer loyalty.

In machine learning, Gradient Boosting Machine (GBM) is a decision tree-based machine learning algorithm used for supervised learning, such as classification and regression, which in this case includes classification. In addition, the GBM algorithm belongs to the ensemble learning technique, which utilizes multiple decision trees to improve prediction performance. Ensemble learning is one of the machine learning paradigms where multiple models (base models) are trained to solve the same problem and combined to get better results (Maretva Cendani & Wibowo, 2022). GBM is an ensemble method that combines many weak learners through an iterative process of each decision tree created sequentially, with each decision tree trying to minimize the function loss.

Various previous studies have been conducted, exploring various methods and algorithms to predict customer churn. In this research, the Cross Industry Standard Process for Data Mining (CRISP-DM) method will be used as an analyst framework. The CRISP-DM method is used to provide a more structured and comprehensive approach. In addition, this research will focus more on the application of the Gradient Boosting Machine (GBM) algorithm for customer churn prediction, which is modified by using the Synthetic Minority Oversampling Technique (SMOTE) in the pre-processing stage and Randomized Search in the parameter optimization process. With a hybrid approach that combines data balancing methods using SMOTE and parameter optimization through randomized search, this research is expected to make a significant contribution in improving the accuracy of customer churn prediction, especially in the context of using the GBM algorithm.

LITERATURE REVIEW

Various previous studies have been conducted, exploring various methods and algorithms for predicting customer churn. One of them is by Namira *et al.* (2024), about predicting customer churn using the Decision Tree, Random Forest, and Support Vector Machine algorithms with a dataset division ratio of 80:20, and producing an accuracy value of 80.78% Decision Tree, 83.86% Random Forest, and 83.32% Support Vector Machine. Another research was conducted by Hafidatus Sholeha *et al.* (2024), in which they predicted customer churn in online stores with the Tree-Based Gradient Boosted Models method and produced an accuracy value of 80.032% XGBoost, 80.032% LightGBM, and 80.026% Catboost. (Clementine Aksama & Wahyuniati, 2022) in their research predicted bank customer churn using Naive Bayes and ID3 methods, and applied the SEMMA stage, with a 6:4 ratio dataset division and produced an accuracy value of 85.17% Naive Bayes, while for ID3 79.17%. This research uses CRISP-DM as the analysis framework. The GBM model was modified to improve its performance by applying SMOTE in the pre-processing stage to handle data imbalance and Randomized Search in the parameter optimization process to find the best combination of parameters.

RESEARCH METHOD

Cross Industry Standard Process for Data Mining (CRISP-DM)

This research uses the Cross Industry Standard Process for Data Mining (CRISP-DM) as the analysis framework. CRISP-DM is a de facto standard and industry-independent process model for implementing data mining projects (Schröer *et al.*, 2021). CRISP-DM is a highly structured and

systematic methodology for data mining, which covers all important aspects, from business understanding to model implementation (Rianti *et al.*, 2023).

Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) is an algorithm for overcoming imbalanced data by generating synthetic minority samples through random linear interpolation between minority samples and their nearest neighbors, thus improving classification effectiveness on such datasets (Wang *et al.*, 2021). SMOTE creates synthetic data by adding values calculated from the distance between a minority class sample and its neighbors, using random percentages to generate variation (Dablain *et al.*, 2023). SMOTE works by generating synthetic data for the minority class, not by simply duplicating existing data but by creating new data that falls between the minority samples using interpolation. The implementation of this method can be done using SMOTE from the imbalanced-learn library, which allows users to increase the proportion of minority classes in the dataset.

Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) was first introduced by Friedman (2001) in a series of papers entitled Greedy Function Approximation: A Gradient Boosting Machine. GBM is an ensemble learning algorithm designed to provide an alternative solution to the problem by building models incrementally (Zhang & Haghani, 2015). Each new model added serves to correct the prediction error of the previous model, thus creating a stronger overall model. GBM combines weak learners, which are models that perform slightly better than random guesses, into a strong learner through an iterative process (Bentéjac *et al.*, 2021). This algorithm utilizes simple decision trees (weak learners) as its basic element, where each tree has limited prediction capabilities. By combining predictions from many weak learners through a linear combination approach, GBM is able to gradually minimize the empirical risk function and improve overall model performance (Biau *et al.*, 2019). According to Friedman (2001), the main formula of GBM refers to an iterative approach that optimizes the loss function through gradient descent in function space. This approach builds the model incrementally by additive expansion:

$$Fm(x) = Fm - 1(x) + v \cdot \rho m \cdot h(x; \theta m)$$

Randomized Search

Randomized search is a hyperparameter search method used to find the best combination of parameters in a machine learning model. Randomized search creates a population of architectures randomly and evaluates the performance of each architecture to select the best-performing architecture as the final result (Liu *et al.*, 2017). This allows the algorithm to explore various combinations without covering the entire search space, which is particularly useful when the parameter space is very large. By performing a randomized search, Randomized Search can save computational time and reduce the number of iterations required to find the optimal model. The implementation of this method can be done using RandomizedSearchCV from Scikit-learn, which allows the user to specify parameters and value distributions for the hyperparameters to be tested.

RESULTS AND DISCUSSION

Business Understanding

Churn prediction can help banks identify customers who have the potential to stop using their services and take proactive preventive measures. This approach not only helps retain existing customers but also reduces the cost incurred in acquiring new customers, which is usually much more expensive. In predicting customer churn, there are often two challenges, namely data imbalance and the need for optimal parameters for the algorithm. Therefore, this study aims to evaluate the effectiveness of the combination of the Synthetic Minority Oversampling Technique (SMOTE) in the data preprocessing stage and Randomized Search in the parameter optimization process to improve the performance of the Gradient Boosting Machine (GBM) model.

Data Understanding

This research is quantitative research with a computational experiment approach that aims to evaluate and compare the performance of machine learning-based customer churn prediction models. The type of data used is secondary data, in the form of historical customer datasets from the banking industry. This dataset is taken from Kaggle, a public dataset provider platform that is widely used for data mining and machine learning-based research. The dataset used has 10,000 data points with 18 features covering customer information from the relevant banking industry, as in Table 1.

Table 1. Dataset feature details

Feature Name	Data Type	Feature Description
Identification		
RowNumber	int	Line number
CustomerId	int	Customer ID
Surname	object	Bank customer name
Demographics		
Age	int	Age
Geography	object	Location of customer's residence
Gender	object	Gender
Finance and Activity		
CreditScore	int	Credit score
Tenure	int	Period of employment
Balance	float	Balance
NumOfProducts	int	Number of products purchased by customers through the bank
HasCrCard	int	Indicates whether the customer has a credit card

		Value 0 = No Value 1 = Yes
IsActiveMember	int	Active member Score 0 = No Value 1 = Yes
EstimatedSalary	float	Estimated salary
Satisfaction and Loyalty		
Points Earned	int	Points earned by customers from using credit cards
Complaint	int	Whether the customer has a complaint or not
Satisfaction Score	int	Satisfaction score, the score given by customers for the resolution of their complaints
Card Type	object	Type of card owned by the customer
Target or Label		
Exited	int	Whether the customer has left the bank or not Value 0 = No Churn Value 1 = Churn

Table 2. Dataset descriptive statistics

Atribut	Mean	Std Dev	Min	Q1	Q3
CreditScore	650.53	96.65	350.00	584.00	718.00
Age	38.92	10.49	18.00	32.00	44.00
Tenure	5.01	2.89	0.00	3.00	7.00
Balance	76,485.89	62,397.41	0.00	0.00	127,644.24
NumOfProducts	1.53	0.58	1.00	1.00	2.00
HasCrCard	0.71	0.46	0.00	0.00	1.00
IsActiveMember	0.52	0.50	0.00	0.00	1.00
EstimatedSalary	100,090.24	57,510.49	11.58	51,002.11	149,388.25
Exited	0.20	0.40	0.00	0.00	0.00
Complain	0.20	0.40	0.00	0.00	0.00
Satisfaction Score	3.01	1.41	1.00	2.00	4.00
Point Earned	606.52	225.92	119.00	410.00	801.00

The target class distribution shows a significant imbalance, with around 20% of customers churning, this is shown in Table 2 of the Exited section.

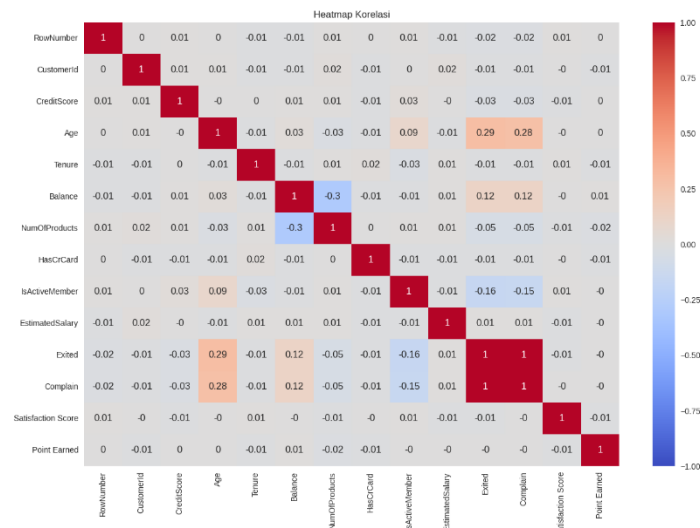


Figure 2. Correlation heatmap

Figure 2 shows the correlation heatmap, which serves to visualize the linear relationship between attributes in the dataset, including independent features (age, balance, or credit score) and the target feature (Exited). In the heatmap, correlation is measured with values of -1 to 1. A value of 1 indicates a perfect positive correlation (when one feature increases, the other feature also increases). A value of -1 indicates a perfect negative correlation (when one feature increases, another feature decreases). A value of 0 indicates no linear relationship between features. Based on the correlation heatmap, it can be analyzed that Complain and Exited have a correlation of 1 indicating a perfect positive relationship between these features. Every customer who makes a complaint (Complain = 1) is always directly related to churn (Exited = 1). In other words, all customers who complain must have stopped being customers, and no complaints are filed by customers who stay. In addition, features with high correlation can cause overfitting, so to prevent this, the Complain feature was removed. In this study, the RowNumber, CustomerId, Surname features were removed because they had no correlation with the research target. In addition, the one-hot encoding technique was used to perform processing on categorical variables such as Geography, Gender, and Card Type, after which the total features increased to 19 features.

Data Preparation

At this stage, several important steps are taken to prepare the data for optimal use by the Gradient Boosting Machine (GBM) model. The first step is to remove duplicate data to ensure that each entity in the dataset is unique and there is no redundant information that could affect the model training. After that, the next step is to handle class imbalance by using SMOTE. SMOTE generates synthetic data for the minority class, in this case churning customers, by creating new examples similar to the existing examples around the original data points. After the application of SMOTE, the amount of data used for training increased to 15,924 data points, with a more balanced class distribution between churn and non-churn customers. This helps the model to be more sensitive to the minority class, which is often overlooked in datasets with large imbalances. The class distribution before and after the application of SMOTE can be seen in Figure 3, which illustrates a visual comparison between the two conditions.

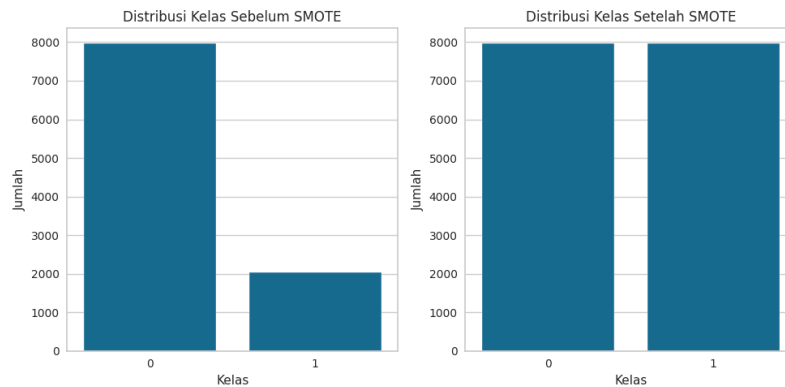


Figure 3. SMOTE Implementation

After applying the SMOTE technique, the next step is data normalization using Robust Scaler. This method was chosen for its ability to handle outliers that may be present in the data so that the distribution of the data is maintained and not distorted by extreme values. This normalization is applied to all features to ensure the scale of the data remains consistent in order to learn patterns more effectively. Once the data is processed, the dataset is split in a 7:3 ratio, where 70% of the data is used for training and 30% is used for model testing. This division is done to ensure that the model can be tested with data that has never been seen before, so that the evaluation of model performance can be done more objectively.

Modeling

In the initial stage of the research, the Gradient Boosting Machine (GBM) model was applied with the default parameter configuration, i.e., using `random_state=42` to ensure the experiments can provide consistent results. At this stage, the model did not use SMOTE techniques for data balancing or Randomized Search for parameter optimization. In addition, the initial evaluation is done without using a cross-validation scheme. The initial evaluation is done by directly dividing the dataset into training data and testing data using a proportion of 7:3. This approach was chosen to provide an initial assessment of the model performance without additional complex validation processes.

Furthermore, parameter optimization is performed using Randomized Search. Randomized Search randomly selects a combination of parameters from a predetermined parameter space, allowing for broader exploration with lower computation time, resulting in optimal parameters as shown in Table 3.

Table 3. Randomized search

Parameters	Optimization Parameters
criterion: friedman_mse, squared_error	Best criterion: squared_error
learning_rate: 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8	Best cross-validation score: 0.9111786470619448
n_estimators: 50, 100, 150, 200, 250, 300, 350	Best n_estimators: 250
max_depth: None, 3, 5, 7, 10, 15, 20, 30	Best max_depth: 15
min_samples_split: 2, 5, 10, 20, 30	Learning Rate: 0.4
min_samples_leaf: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Best min_samples_split: 30
max_features: None, sqrt, log2	Best min_samples_leaf: 5
subsample: 0.5, 0.6, 0.7, 0.8, 0.9, 1.0	Best max_features: None

Evaluation

In this study, two scenarios were conducted to evaluate the performance of the Gradient Boosting Machine (GBM) model in predicting customer churn. Table 4 shows the model evaluation results in scenario 1, namely without the application of SMOTE and Randomized Search, while Table 5 shows the model evaluation results in scenario 2, namely with the application of SMOTE and Randomized Search.

Table 4. GBM without SMOTE and Randomized Search

Actual Value	Predicted Value	
	No Churn	No Churn
No Churn	TN = 2336	No Churn
Churn	FN = 305	Churn

Table 5. GBM with SMOTE and Randomized search

Actual Value	Predicted Value	
	No Churn	No Churn
No Churn	TN= 2284	No Churn
Churn	FN = 258	Churn

Evaluation is done using metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC as shown in Table 6.

Table 6. Evaluation

	Scenario 1	Scenario 2
Accuracy	0.8717	0.9085
Precision	0.78	0.92
Recall	0.48	0.89
F1-Score	0.59	0.90

AUC-ROC indicates the model's ability to distinguish between positive and negative classes. Based on the evaluation results:

1. In scenario 1, the AUC-ROC value is 0.86 as shown in Figure 4.

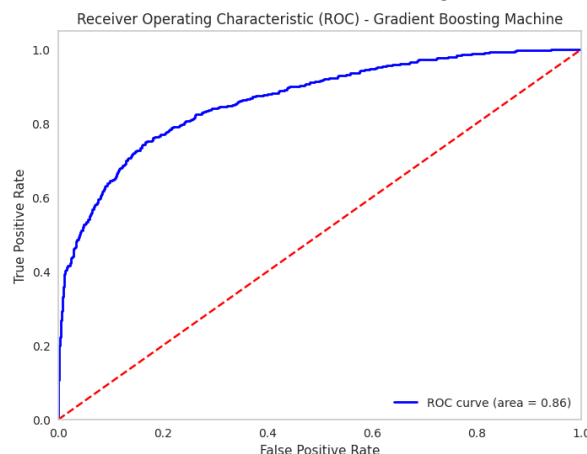


Figure 4. Grafik ROC skenario 1

2. In scenario 2, the AUC-ROC value increased to 0.96 as shown in Figure 5.

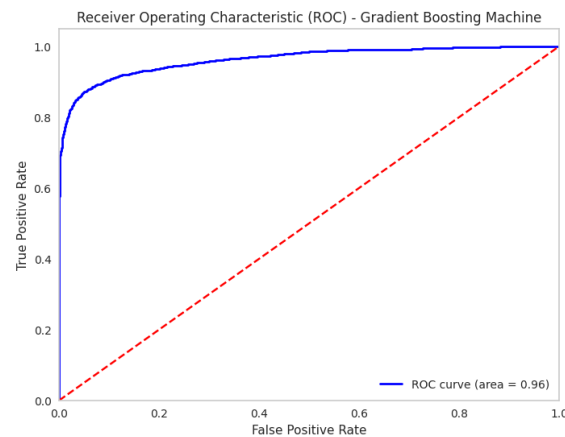


Figure 5. ROC Skenario 2 chart

Based on the calculation results, it can be seen that the application of SMOTE and Randomized Search in scenario 2 has a significant positive impact. The increase in Precision, Recall, and F1-Score values indicates that the model is more effective in recognizing and predicting potentially churning customers, despite the initial data imbalance. The 3.45% increase in accuracy has significant implications in the context of the banking industry, as it can help banks identify customers who require special attention for retention strategies. In addition, the increase in AUC-ROC value from 0.86 to 0.96 reflects the model's ability to provide more accurate predictions overall.

In scenario 1, without the application of SMOTE, the GBM model struggles to detect churn customers, as evidenced by the low Recall value (48%) indicating a high number of false negatives. SMOTE addresses the data imbalance by generating synthetic examples for the churn class, which increases the proportion of churn data in the training dataset. In scenario 2, the application of SMOTE increased Recall to 89%, making the model more sensitive to churned customers. Although the precision decreases slightly, SMOTE still helps the model recognize churned customers better and maintain a balanced prediction quality between the two classes. In scenario 2, the use of Randomized Search for hyperparameter optimization was shown to significantly improve model performance. Randomized Search selects parameter combinations randomly, allowing the model to explore more possibilities in less time. As a result, the model accuracy increased from 87.17% to 90.85%, thanks to parameters such as `n_estimators`, `max_depth`, and `learning_rate` being optimized. In addition, the precision, recall, and F1-score also improved, while the AUC-ROC increased from 0.86 to 0.96, indicating the model's improved ability to distinguish between churn and non-churn customers.

Overall, the application of SMOTE and Randomized Search in this study shows that they can work synergistically to improve the performance of the GBM model in predicting customer churn. SMOTE helps the model in handling data imbalance, while Randomized Search improves the model's ability to find the optimal parameter combination. Both together provide better results in terms of accuracy, precision, recall, and AUC-ROC, which are crucial in improving the effectiveness of customer retention strategies in the banking industry.

CONCLUSION

This research shows that the combination of the Synthetic Minority Over-sampling Technique (SMOTE) and Randomized Search for hyperparameter optimization can significantly improve the performance of the Gradient Boosting Machine (GBM) model in predicting customer churn in the banking industry. The application of SMOTE helps to overcome the data imbalance problem by increasing the sensitivity of the model to the minority class (churned customers), which is reflected in the improvement of recall and quality of churn prediction. Meanwhile, Randomized Search successfully found the optimal parameter combination, which contributed to the increase in accuracy from 87.17% to 90.85%, as well as a significant increase in precision, recall, F1-score, and AUC-ROC values.

Overall, the results of this study show that by using SMOTE and Randomized Search, the GBM model can provide more accurate and balanced predictions, which is crucial for a more effective customer retention strategy. The implementation of these two techniques can assist banks in better identifying customers at risk of churn, enabling more targeted countermeasures to increase customer loyalty and reduce potential financial losses. However, this study has a limitation, which is that it does not include a deployment process to apply the model in a real banking operational system, so the impact of implementing this model in the real world cannot be evaluated.

REFERENCES

- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937–1967. <https://doi.org/10.1007/s10462-020-09896-5>
- Biau, G., Cadre, B., & Rouvière, L. (2019). Accelerated gradient boosting. *Machine Learning*, 108(6), 971–992. <https://doi.org/10.1007/s10994-019-05787-1>
- Clementine Aksama, M., & Wahyuniati, A. (2022). Prediksi Churn Nasabah Bank Menggunakan Klasifikasi Naive Bayes dan ID3. *Jurnal Processor*, 17(1), 9–18. <https://doi.org/10.33998/processor.2022.17.1.1170>
- Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9), 6390–6404. <https://doi.org/10.1109/TNNLS.2021.3136503>
- Farid Naufal, M., Fernando Susanto, A., Nathaneil Kansil, C., & Huda, S. (2023). Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Potensi Hilangnya Nasabah Bank Application of Machine Learning to Predict Potential Loss of Bank Customer. In *Februari* (Vol. 22, Issue 1).
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. In *Source: The Annals of Statistics* (Vol. 29, Issue 5).
- Hafidatus Sholeha, S., Faid, M., & Yaqin, M. A. (2024). Prediksi Perpindahan Pelanggan Pada Toko Online Menggunakan Metode Tree-Based Gradient Boosted Models. *Journal of Computer System and Informatics*, 5(3), 605–614. <https://doi.org/10.47065/josyc.v5i3.5215>
- Irmanda, H. N., Astriratma, R., & Afrizal, S. (2019). Perbandingan Metode Jaringan Syaraf Tiruan Dan Pohon Keputusan Untuk Prediksi Churn. *Jurnal Sistem Informasi*, 11, 1817–1825.
- Liu, H., Simonyan, K., Vinyals, O., Fernando, C., & Kavukcuoglu, K. (2017). *Hierarchical Representations for Efficient Architecture Search*. <http://arxiv.org/abs/1711.00436>

- Maretva Cendani, L., & Wibowo, A. (2022). *Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes* (Vol. 13, Issue 1).
- Namira, Slamet, I., & Susanto, I. (2024). Prediksi Nasabah Churn Dengan Algoritma Decision Tree, Random Forest Dan Support Vector Machine. *ESCAF*, 1045–1053. <https://help.sap.com>
- Rianti, A., Majid, N. W. A., & Fauzi, A. (2023). *CRISP-DM: Metodologi Proyek Data Science*.
- Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181, 526–534. <https://doi.org/10.1016/j.procs.2021.01.199>
- Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-021-03430-5>
- Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308–324. <https://doi.org/10.1016/j.trc.2015.02.019>