Journal Collabits
Vol. 3 No. 1 January 2026 : 16-25
E-ISSN : 3046-6709, P-ISSN : 3062-8601
https://publikasi.mercubuana.ac.id/index.php/collabits

Analysis and Prediction of Customer Churn in the Telecommunications Industry Using Logistic Regression and Random Forest

Celsi Alisa Nabila[1*], Ryno Julian Santoso[2], Sabila Alya Nafisa[3], Yuni Roza[4]

[1,2,3] Computer Science, Universitas Mercu Buana, Indonesia
[4] Computer Science, Institut Teknologi Perusahaan Listrik Negara, Indonesia

*Coressponden Author: celsialisa7@gmail.com

Abstract—Customer churn represents a major challenge for telecommunication companies because of its significant influence on revenue stability and customer retention efforts. Intense competition among service providers has increased the need for reliable predictive models capable of identifying customers with a high probability of terminating their subscriptions. This study focuses on the analysis and prediction of customer churn by applying machine learning techniques to the Telco Customer Churn dataset. The research workflow includes data preprocessing stages such as duplicate removal, treatment of missing values, and transformation of both categorical and numerical features. Exploratory data analysis supported by visualization techniques is employed to examine customer behavior and feature relationships. Subsequently, the dataset is partitioned into training and testing subsets using an 80:20 stratified split. A preprocessing pipeline is applied, incorporating feature scaling for numerical variables and one-hot encoding for categorical variables. Predictive models are developed using Logistic Regression and Random Forest algorithms, and their performance is assessed through accuracy measurements and classification reports. The results indicate that the Random Forest model delivers better predictive performance than Logistic Regression, demonstrating its effectiveness in modeling complex data patterns. Overall, the study confirms that machine learning-based approaches can serve as effective tools for churn prediction and offer meaningful insights to support strategic decision-making in customer retention within the telecommunication sector.

Keywords :
Customer Churn;
Machine Learning;
Logistic Regression;
Random Forest;

Article History:
Received: 05-12-2025
Revised: 26-12-2025
Accepted: 08-01-2025

Article DOI : 10.22441/collabits.v3i1.37599

INTRODUCTION

The rapid development of the telecommunication industry has intensified competition among service providers, making customer retention an increasingly important factor for long-term business sustainability. Customers are able to switch providers easily due to competitive pricing, service quality improvements, and various promotional offers. This condition, known as customer churn, presents a major challenge for telecommunication companies, as retaining existing customers is generally more cost-effective than acquiring new ones.

Consequently, customer churn analysis has become a significant area of research, as it allows companies to identify customers who are likely to discontinue their services and to design appropriate retention strategies. Conventional churn analysis methods often rely on descriptive statistics and manual decision-making processes, which are limited when applied to large and complex datasets. With advances in information technology, machine learning techniques have increasingly been used to address these limitations by enabling data- driven predictive modeling.

Machine learning algorithms can learn patterns from historical customer data and generate accurate predictions of customer behavior. Several studies have shown that classification methods such as Logistic Regression and Random Forest are effective for churn prediction tasks. Logistic Regression is commonly used because of its simplicity, interpretability, and computational efficiency, while Random Forest provides stronger predictive capability by combining multiple decision trees to capture more complex relationships among features.

However, the performance of machine learning models in churn prediction is strongly influenced by data preprocessing and feature transformation steps. Proper data cleaning, handling of missing values, encoding of categorical variables, and scaling of numerical attributes are essential to improve model accuracy. Therefore, an integrated approach that includes data cleaning, exploratory data analysis, visualization, and model comparison is required.

Based on these considerations, this study focuses on predicting customer churn in the telecommunication industry using the Telco Customer Churn dataset. The main objective is to compare the performance of Logistic Regression and Random Forest algorithms in predicting customer churn and to evaluate their effectiveness using accuracy and classification metrics. The results of this study are expected to provide useful insights for telecommunication companies in developing data-driven customer retention strategies.

LITERATURE REVIEW

Customer churn is widely acknowledged as one of the most serious challenges faced by the telecommunication sector because of its substantial impact on company revenue, customer lifetime value, and long-term organizational stability. In competitive service environments, customers are no longer bound to a single provider and can easily migrate to alternative services that offer better pricing schemes, improved service quality, or more attractive incentives. Consequently, churn prediction has become a strategic necessity for telecommunication companies seeking to design effective customer retention programs. Numerous studies confirm that retaining existing customers requires considerably lower costs than acquiring new ones, which has encouraged the adoption of analytical and predictive techniques to better understand customer behavior and churn dynamics [1], [8], [21], [25].

The increasing availability of large-scale customer data has shifted churn analysis from traditional statistical approaches toward data mining and machine learning methodologies. Earlier studies relied heavily on descriptive and rule-based analysis; however, such methods proved inadequate for processing complex, high-dimensional datasets. Recent research increasingly employs supervised learning models to distinguish churn and non-churn customers by leveraging demographic characteristics, service usage patterns, and subscription-related variables [3], [10], [31]. Among these methods, Logistic Regression continues to be widely used as a baseline classifier due to its transparent mathematical structure, low computational cost, and ability to explain the influence of individual predictors on churn probability [2], [11], [20], [39].

In contrast to linear models, tree-based and ensemble learning techniques have demonstrated stronger capabilities in modeling non-linear interactions among customer attributes. Decision Tree-based models and Random Forest classifiers are frequently adopted in churn prediction studies, with Random Forest consistently reported to deliver superior performance. This advantage is primarily attributed to its ensemble mechanism, which reduces overfitting and improves generalization across diverse feature sets [1], [4], [7], [35], [40]. Comparative investigations further indicate that Random Forest often outperforms conventional classifiers such as Naive Bayes, K-Nearest Neighbors, and Support Vector Machine, particularly when applied to complex and heterogeneous customer datasets [6], [14], [18], [36].

Recent research also emphasizes that predictive accuracy is not solely determined by algorithm selection, but is strongly influenced by data preprocessing and feature engineering strategies. Processes such as duplicate removal, missing value treatment, feature scaling, categorical encoding, and feature selection have been shown to significantly affect model outcomes [17], [31], [33], [37]. Additionally, customer churn datasets commonly suffer from class imbalance, where churned customers represent a

minority class. To mitigate this issue, resampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been widely applied to enhance classification performance and reduce predictive bias [12], [13], [38].

From a managerial perspective, churn behavior is closely linked to factors such as customer satisfaction, pricing policies, contract duration, and service utilization patterns. Empirical findings suggest that variables including customer tenure, billing methods, subscription types, and service features play a critical role in influencing churn decisions [8], [21], [26], [30]. Consequently, churn prediction models are increasingly integrated into customer relationship management systems to support strategic planning, customer segmentation, and personalized retention initiatives [22], [23], [28], [29].

Despite the extensive body of research on churn prediction, several limitations remain evident. Many studies prioritize maximizing predictive accuracy through complex models, while offering limited discussion on interpretability and real-world deployment. Moreover, comparative studies often lack standardized preprocessing pipelines or restrict their analysis to a limited set of algorithms, making cross-study performance evaluation challenging [7], [16], [34]. Although ensemble learning models demonstrate strong predictive potential, simpler and more interpretable approaches such as Logistic Regression remain highly relevant, particularly for organizations that require transparency and explainability in decision-making processes.

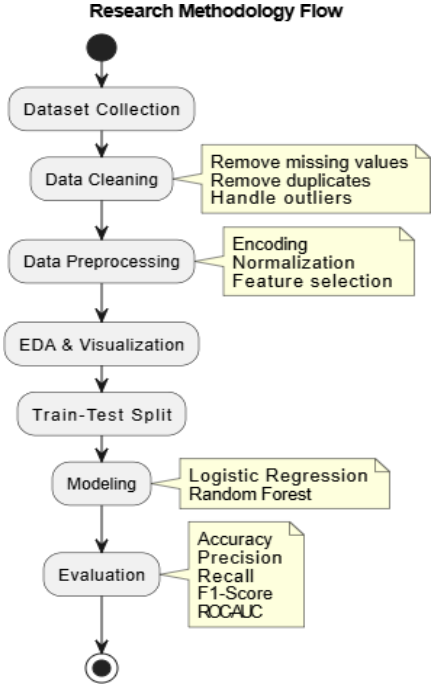
Addressing these gaps, the present study conducts a structured comparison between Logistic Regression and Random Forest models for customer churn prediction in the telecommunication domain. By implementing a comprehensive preprocessing framework, exploratory data analysis, and standardized evaluation metrics, this research aims to provide empirical evidence regarding the trade-offs between predictive performance and interpretability in both classical and ensemble machine learning models.

The references used in this study primarily consist of recent literature published within the last decade and are drawn from reputable national and international journals as well as international conference proceedings [2], [3]. The research problems, objectives, and contributions are presented narratively without the use of specific subheadings, in accordance with academic writing standards. Operational definitions, where necessary, are also described narratively to ensure conceptual clarity [4].

METHODOLOGY

This research adopts a quantitative methodological framework that applies machine learning techniques to predict customer churn within the telecommunication sector. The overall research process is organized into a sequence of structured stages, comprising data acquisition, data preprocessing, exploratory analysis, model construction, and performance assessment. These stages are implemented to ensure that the dataset undergoes appropriate cleaning, transformation, and examination prior to the modeling phase. In addition, the methodology incorporates standardized preprocessing workf lows and evaluation criteria to support an objective comparison between the selected classification models. Through this systematic procedure, the resulting predictive models are expected to deliver accurate, consistent, and interpretable outcomes for customer churn prediction.

Figure 1. Research Framework for Customer Churn Prediction



Dataset Description

This study utilizes the Telco Customer Churn dataset, which provides comprehensive customer records covering demographic characteristics, service subscriptions, billing information, and service termination status. The dataset is composed of a mixture of numerical and categorical variables, making it appropriate for supervised classification approaches in churn prediction. In this research, the churn indicator serves as the response variable, representing whether a customer has ceased using the telecommunication service.

Table 1. Dataset Feature Description

Feature Type	Attributes
Numerical	tenure, MonthlyCharges, TotalCharges
Categorical	gender, SeniorCitizen, Partner, Contract, PaymentMethod, etc.
Target	Churn

Data Cleaning and Preprocessing

Data preparation was conducted to enhance the robustness and credibility of the predictive models. Redundant observations were eliminated to minimize potential bias during the learning process. For numerical variables containing incomplete values, median-based imputation was applied to retain the original data characteristics. Attributes that function solely as identifiers and do not contribute meaningful information were removed to prevent unnecessary interference in model learning. In addition, the churn outcome was converted into a binary numerical variable to ensure compatibility with supervised classification methods.

Categorical features were transformed into numerical form through one-hot encoding, whereas numerical features were normalized using standardization techniques. This approach allows all variables to be represented on a comparable scale, ensuring balanced contribution of features during the model training phase.

Data Splitting

Following preprocessing, the dataset was partitioned into two subsets consisting of training and testing data using an 80:20 proportion. A stratified sampling strategy was employed to preserve the original class distribution of churn and non-churn instances across both subsets. The training subset was

utilized for model development, while the testing subset was used solely to evaluate predictive performance.

Model Development

Table 2. Model Parameters

Model	Parameter	Value
Logistic Regression	Max Iteration	1000
Random Forest	Number of Trees	200

This research applies two supervised learning techniques for classification, namely Logistic Regression and Random Forest. Logistic Regression was employed as a reference model because of its straightforward formulation, interpretability, and common use as a baseline classifier. In contrast, Random Forest, which is an ensemble-based approach constructed from multiple decision trees, was utilized for its strength in modeling non- linear patterns and complex interactions among input features.

To ensure objective evaluation, both classifiers were trained on an identical preprocessed training dataset. The same preprocessing pipeline was uniformly applied to each model, allowing feature scaling and categorical variable encoding to be seamlessly incorporated into the learning process.

Model Evaluation

Model evaluation was conducted using widely adopted classification performance measures, namely accuracy, precision, recall, and the F1-score. These indicators were selected to provide a thorough evaluation of each model's capability in identifying customer churn. Accuracy reflects the overall prediction correctness, whereas precision and recall focus on the model's effectiveness in correctly detecting churned customers. The F1-score serves as a balanced metric that integrates precision and recall, which is particularly relevant in churn prediction tasks where the data distribution between classes is often uneven.

The performance outcomes generated from the testing dataset were subsequently examined and compared in order to identify the classification model that demonstrates the most reliable performance for customer churn prediction within the telecommunication domain.

RESULTS AND DISCUSSION

Table 3. Training and Testing Data Distribution

Data Type	Number of Records	Percentage
Training Data	5,634	80%
Testing Data	1,409	20%
Total	**7,043**	**100%**

This chapter outlines the outcomes of exploratory analysis, data visualization, and machine learning model assessment applied to customer churn prediction using the Telco Customer Churn dataset. The dataset contains 7,043 customer instances described by 21 attributes, comprising 16 categorical features, three numerical variables (tenure, MonthlyCharges, and TotalCharges), and a single target variable representing churn status. Following the data cleaning and preprocessing stages, the dataset was partitioned into a training set consisting of 5,634 records (80%) and a testing set of 1,409 records (20%) to support an objective and reliable evaluation of the developed models.

1.1 Exploratory Data Analysis and Visualization

Exploratory data analysis was performed to examine patterns and relationships among the main numerical variables and their influence on customer churn behavior.

Figure 1. Scatter Plot of MonthlyCharges and TotalCharges

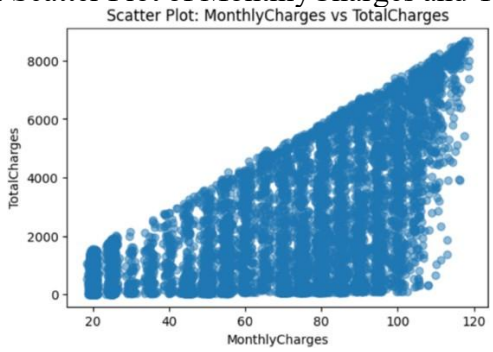


Figure 2. Correlation Heatmap of Numerical Features

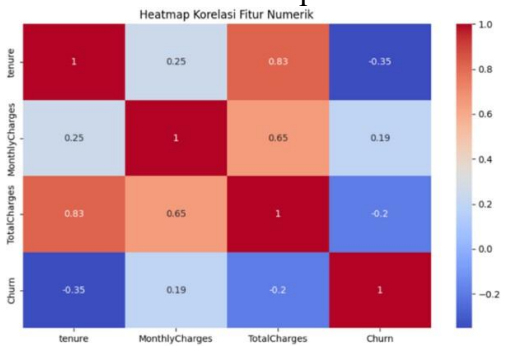
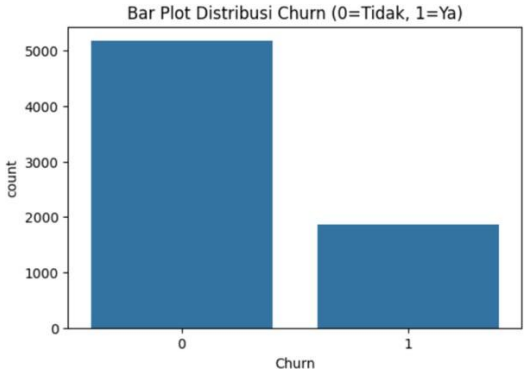


Figure 3. Distribution of Customer Churn



The scatter diagram illustrating the relationship between MonthlyCharges and TotalCharges indicates a clear upward trend, suggesting that customers who incur higher monthly fees generally accumulate greater total costs throughout their subscription period. The spread of data points also highlights the effect of customer tenure, where longer service durations are associated with increased total billing amounts.

Additionally, the correlation heatmap of numerical variables demonstrates a strong positive association between tenure and TotalCharges, with a correlation coefficient of 0.83. MonthlyCharges also shows a moderate positive relationship with TotalCharges (0.65). Conversely, the churn variable exhibits a negative correlation with tenure (-0.35), implying that customers with shorter subscription durations have a higher likelihood of terminating their services. This observation is consistent with previous research identifying tenure as a key indicator of customer retention.

Moreover, the bar chart representing churn distribution reveals that non-churn customers constitute the majority of the dataset, confirming the existence of class imbalance. The relatively smaller proportion of churned customers is a critical consideration, as it can influence model learning behavior and reduce predictive performance for the minority class.

1.2 Model Performance Comparison

This study evaluates the performance of two supervised classification models, namely Logistic Regression and Random Forest, using standard evaluation metrics including accuracy, precision, recall, and F1-score.

Table 4. Accuracy Comparison of Classification Models

Model	Accuracy
Logistic Regression	80,5 %
Random Forest	78,1 %

The evaluation results indicate that Logistic Regression achieved a higher accuracy rate of 80.55%, surpassing the Random Forest model, which recorded an accuracy of 78.14%. Further analysis of the classification metrics shows that Logistic Regression provides a more balanced performance in distinguishing churn and non-churn customers, particularly reflected in its higher recall value of 0.56 for churn cases, compared to 0.47 obtained by Random Forest. This suggests that Logistic Regression is more effective in identifying customers at risk of churn, which is a crucial objective in churn prediction applications.

Although Random Forest is widely known for its strength in capturing complex and non-linear patterns, its slightly lower performance in this study may be attributed to factors such as class imbalance, hyperparameter selection, and the predominance of categorical features. These aspects can limit the effectiveness of ensemble-based models when optimization is not fully tailored to the dataset characteristics.

Discussion

The findings demonstrate that Logistic Regression shows strong performance and, in this experimental context, exceeds the results obtained by the Random Forest model. This outcome suggests that, after preprocessing, the underlying churn characteristics within the dataset tend to follow a near-linear pattern, enabling Logistic Regression to establish an effective decision boundary. In addition to its predictive capability, Logistic Regression offers superior model transparency, allowing decision-makers to clearly interpret how individual features influence the probability of customer churn.

Table 5. Performance Metrics of Logistic Regression

Class	Precision	Recall	F1-Score
Non-Churn (0)	0.85	0.89	0.87
Churn (1)	0.66	0.56	0.60
Overall Accuracy			**0.81**

The analysis further reveals that several attributes substantially affect churn behavior, including customer tenure, contract category, monthly billing amount, and payment method. Customers with shorter subscription periods, higher recurring charges, and non-long-term contract arrangements exhibit a greater tendency to terminate their services. These observations align with prior studies that highlight longer service duration and stable contract commitments as key drivers of customer retention. In summary, integrating exploratory data analysis, visual exploration, and systematic model comparison yields valuable insights into churn dynamics. The results also demonstrate the practical value of machine learning-based approaches in assisting telecommunication companies to design more informed and targeted customer retention strategies.

CONCLUSION

This research confirms that machine learning approaches are well suited for predicting customer churn in the telecommunications sector using the Telco Customer Churn dataset. The dataset contains 7,043 customer observations comprising both categorical and numerical features, which were effectively processed through systematic data cleaning, transformation, and encoding stages. The exploratory analysis identified clear associations between customer characteristics and churn behavior, particularly for variables such as tenure, MonthlyCharges, and TotalCharges.

The comparative evaluation shows that Logistic Regression outperformed Random Forest in this study, achieving an accuracy of 80.55%, while Random Forest attained 78.14%. In addition, Logistic Regression produced higher recall and F1-score values for the churn class, indicating a stronger capability to detect customers who are at risk of terminating their subscriptions. These findings suggest that Logistic Regression is not only competitive in performance but also well suited as an interpretable baseline model, especially in scenarios where model transparency and explainability are essential.

Although Random Forest is recognized for its strength in modeling nonlinear patterns and complex feature interactions, its results in this experiment were relatively lower, particularly in identifying churned customers. This limitation may be attributed to class imbalance in the dataset and the absence of advanced sampling strategies during model training. Nonetheless, Random Forest remains an important analytical tool due to its stability and ability to capture relationships that may not be adequately represented by linear models.

From a managerial standpoint, the results indicate that customers with shorter service duration, higher monthly fees, and specific contract arrangements are more prone to churn. These insights can support telecommunication providers in designing proactive retention initiatives, including customized pricing schemes, contract adjustments, and enhanced customer service programs.

For future work, it is recommended to incorporate imbalance-handling techniques such as SMOTE, evaluate additional algorithms such as gradient boosting and deep learning models, and apply feature selection methods to further improve model performance. Moreover, developing real-time churn prediction systems and incorporating temporal customer behavior data could significantly enhance the applicability of churn prediction models in rapidly changing business environments.

REFERENCES

[1] D. A. Kusuma, A. R. Dewi, and A. R. Wijaya, “Prediksi Customer Churn Menggunakan Algoritma Random Forest pada Data Pelanggan Telekomunikasi,” Jurnal Sistem Informasi, vol. 10, no. 2, pp. 186–194, 2025.

[2] P. Dewi, R. N. Aulia, R. Taufiqillah, and J. Heikal, “Customer Churn Prediction for Life Insurance Using Binary Logistic Regression,” Economic Reviews Journal, vol. 3, no. 3, pp. 2289–2299, 2024, doi: 10.56709/mrj.v3i3.353.

[3] Y. Yudiana, A. Y. Agustina, and N. Khofifah, “Prediksi Customer Churn Menggunakan Metode CRISP-DM pada Industri Telekomunikasi sebagai Implementasi Mempertahankan Pelanggan,” IJIEB: Indonesian Journal of Islamic Economics and Business, vol. 8, no. 1, pp. 1–20, Jun. 2023.

[4] D. Putriani, A. P. A. Prayogi, A. I. Shofyana, A. Ristyawan, and E. Daniati, “Prediksi Customer Churn Menggunakan Algoritma Decision Tree,” INOTEK, vol. 8, Aug. 2024.

[5] A. R. K. Maranto, L. Damayanti, and I. R. Ramadika, “Perbandingan Algoritma C4.5 dan Naïve Bayes dalam Prediksi Loyalitas Pelanggan,” Bit-Tech (Binary Digital–Technology), vol. 7, no. 2, Dec. 2024.

[6] N. Namira, I. Slamet, and I. Susanto, “Prediksi Nasabah Churn dengan Algoritma Decision Tree, Random Forest dan Support Vector Machine,” in Proc. 3rd ESCAF, 2024, pp. xx–xx.

[7] M. F. Naufal et al., “Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Potensi Hilangnya Nasabah Bank,” Techno.COM, vol. 22, no. 1, pp. 1–11, Feb. 2023.

[8] N. A. Khafsoh and Suhairi, “Pemahaman Mahasiswa Terhadap Kekerasan Seksual di Kampus,”

Marwah: Jurnal Perempuan, Agama dan Jender, vol. 20, no. 1, pp. 61–75, 2021.
[9] F. Sinata et al., "Klasifikasi Pelanggan pada Customer Churn Prediction Models Menggunakan Decision Tree," Jurnal Algoritma, Logika dan Komputasi, vol. 8, no. 2, pp. 839–846, 2025.
[10] R. Govindaraju, T. Simatupang, and T. M. A. Samadhi, "Perancangan Sistem Prediksi Churn Pelanggan PT. Telekomunikasi Seluler dengan Memanfaatkan Proses Data Mining," Jurnal Informatika, vol. 9, no. 1, pp. 33–42, 2008.
[11] L. N. Wakhidah, A. K. Zyen, and B. B. Wahono, "Evaluation of Telecommunication Customer Churn Classification with SMOTE Using Random Forest and XGBoost Algorithms," Journal of Applied Informatics and Computing, vol. 9, no. 1, pp. 89–95, Feb. 2025.
[12] F. S. Pratiwi, M. A. Barata, and A. D. Ardianti, "Implementasi Metode SMOTE dan Random Over- Sampling pada Algoritma Machine Learning untuk Prediksi Customer Churn di Sektor Perbankan," Jurnal Sistem Informasi dan Informatika (Simika), vol. 8, no. 1, 2025.
[13] N. Suryana, Pratiwi, and R. T. Prasetio, "Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting," IJCIT, vol. 6, no. 1, pp. 31– 37, 2021.
[14] A. K. Harahap et al., "Perbandingan Akurasi Algoritma K-Nearest Neighbor dan Logistic Regression untuk Prediksi Customer Churn," Jurnal Sentinel, vol. 5, no. 1, pp. 575–581, 2024.
[15] A. R. P. Astawa, G. H. Martono, and Mayadi, "Penerapan Ensemble Learning dengan Hard Voting untuk Klasifikasi Customer Churn," in Seminar Nasional CORISINDO, 2025.
[16] Holilurrahman and M. Imron, "Implementasi Model Prediksi Churn Pelanggan Menggunakan Algoritma Random Forest pada Website Industri Telekomunikasi," Alghoritm: Jurnal Teknologi Informasi, vol. 1, no. 1, 2025.
[17] A. R. Y. Siregar and M. Iqbal, "Prediksi Customer Churn pada Layanan IndiHome Menggunakan Algoritma Decision Tree," Journal of Science and Social Research, vol. 8, no. 1, pp. 204–211, 2025.
[18] A. N. Rachmi, "Implementasi Metode Random Forest dan XGBoost pada Klasifikasi Customer Churn," Tugas Akhir, Universitas Islam Indonesia, 2020.
[19] R. N. S. Hakim and Asmunin, "Optimasi Algoritma Random Forest dengan Teknik Boosting dalam Prediksi Churn Pelanggan di Industri Telekomunikasi," Universitas Negeri Surabaya.
[20] S. T. Utomo, "Analisis Pengaruh Ketidakpuasan dan Perilaku Mencari Variasi dalam Minat Churn," Tesis, Universitas Diponegoro.
[21] I. M. Latief, A. Subekti, and W. Gata, "Analisis Data Pelanggan Menggunakan Pendekatan Machine Learning," Jurnal Informatika, vol. 21, no. 1, 2021.
[22] M. R. Zulman et al., "Temporal Pattern Recognition: A BiLSTM-Based Framework for Churn Prediction," Journal of Artificial Intelligence and Software Engineering, vol. 5, no. 2, pp. 651–659,2025.
[23] S. D. Lukitasari, "Service Innovation for Customer Satisfaction of Telecommunication Companies," ITEJ, vol. 5, no. 1, pp. 14–24, 2020.
[24] R. W. Pertiwi, "Analisis Faktor-Faktor yang Mempengaruhi Ketidakpuasan Pelanggan dan Implikasinya Terhadap Minat Churn Indosat," Skripsi, Universitas Diponegoro, 2015.
[25] M. G. Saputra and B. J. Santoso, "Implementation of Feature Selection Using Boruta to Improve the Accuracy of the Lapser Prediction Model," MALCOM, vol. 5, no. 3, pp. 886–895, 2025.
[26] Yulianti, "Metode Data Mining untuk Prediksi Churn Pelanggan," Jurnal ICT Akademi Telkom Jakarta, vol. 17, 2018.
[27] A. Hermawan et al., "Membangun Model Prediksi Churn Pelanggan yang Akurat," Merkurius, vol. 2, no. 6, pp. 67–81, 2024.
[28] D. Maheswari et al., "Implementasi Algoritma C4.5 untuk Klasifikasi Dampak Pola Penggunaan Media Sosial," JATI, vol. 9, no. 2, 2025.
[29] N. P. N. Fauzi et al., "Penerapan Feature Engineering dan Hyperparameter Tuning untuk Meningkatkan Akurasi Model Random Forest," JTIIK, vol. 12, no. 2, pp. 251–262, 2025.
[30] R. R. Aryanto et al., "Studi Komparasi Model Klasifikasi Berbasis Pembelajaran Mesin," Jurnal RESTI, vol. 5, no. 5, pp. 853–862, 2021.

[31] I. P. Putri et al., "Comparative Analysis of Machine Learning Algorithms for Predicting Child Stunting," MALCOM, vol. 4, no. 1, pp. 257–265, 2024.
[32] T. A. Tutupoly and I. Alfarobi, "Identifikasi Keakuratan Data Pelanggan Menggunakan C4.5 dan Naïve Bayes," Jurnal Teknik Informatika STMIK Antar Bangsa, vol. 4, no. 2, 2018.
[33] J. Jeffry, S. Usman, and F. Aziz, "Analisis Perilaku Pelanggan Menggunakan Metode Ensemble Logistic Regression," Jurnal Penelitian Teknik Informatika, vol. 6, no. 2, 2023.