Faktor Exacta

Vol. 19, No. 1, March 2026, pp. 25~37
eISSN: 2502-339X, pISSN: 1979-276X, https://doi.org/10.30998/faktorexacta.v19i1.828

Comparative Performance Analysis of Logistic Regression and Random Forest for Telecommunications Customer Churn Prediction Using Recursive Feature Elimination

Anggreita Tiara Putri[1], Aa Kurniawan[2], Nurfiqih[3]
[1,2,3] Study Program of Informatics Engineering, Universitas Pamulang, Indonesia

Article Info

Article history:

Received 19/10/2025
Revised 17/06/2026
Accepted 27/06/2026

Keywords:

churn
Feature Selection
Logistic Regression
Prediksi
Random Forest

ABSTRACT

Customer retention is a critical concern in the telecommunications sector due to high revenue risks and subscriber acquisition costs. While machine learning is widely used for churn prediction, cross-validated studies evaluating the impact of Recursive Feature Elimination (RFE) on Logistic Regression and Random Forest remain limited. This research compares these two algorithms, selected for interpretability and non-linear modeling capabilities respectively, before and after RFE feature selection using the Telco Customer Churn dataset (7,043 records). Performance was assessed via 5-fold cross-validation using accuracy, recall, precision, and F1-score. RFE isolated ten key predictors (SeniorCitizen, Dependents, Tenure, PhoneService, MultipleLines, OnlineSecurity, OnlineBackup, TechSupport, Contract, and PaperlessBilling), significantly reducing the feature space. Experimental results show that RFE increased Logistic Regression accuracy from 73.46% to 79.61% (F1-score: 42.35% to 58.60%). Conversely, Random Forest exhibited a modest accuracy gain from 79.85% to 81.00% (precision: 71.00%, F1-score: 63.85%). Ultimately, RFE yields more substantial improvements for Logistic Regression, though Random Forest delivers the highest overall predictive performance.

Corresponding Author:

Name of Corresponding Author,
Department of Informatic,
Universitas Indraprasta PGRI,
Jl. Nangka No. 58 C, Tanjung Barat, Jagakarsa, Jakarta Selatan.
Email: dosen02384@unpam.ac.id

1. INTRODUCTION

In the modern digital environment, the telecommunications industry has become an essential enabler of advances in information and communication technology, supporting connectivity and digital transformation across various sectors. The increasing demand for internet-based services and digital transformation has intensified competition among telecommunications service providers. Sustainable business growth in the telecommunications industry depends not only on customer acquisition but also on effective customer retention strategies. However, many providers face customer churn, in which subscribers end their service contracts and switch to competitors, potentially reducing company revenue and market share. A high level of customer attrition can adversely affect the financial performance of telecommunications providers by reducing recurring revenue and increasing expenses to acquire new customers to offset subscriber losses. Compared with other industries, the telecommunications sector is known to experience relatively high churn rates, exceeding 30% annually in some highly competitive markets[1] [2]. High churn rates are influenced by intense

competition among service providers, low switching costs, and the increasing variety of alternative services available to customers [3]. Identifying subscribers at high risk of leaving a service provider is increasingly recognized as a crucial capability for optimizing customer retention programs and guiding strategic decisions with analytical insights. The likelihood of customer churn is influenced by a diverse set of factors, including customer demographics, service consumption patterns, perceived service quality, responsiveness to customer complaints, pricing competitiveness, and the attractiveness of competing telecommunications offerings [4]. The complexity of these factors makes churn prediction challenging. Failure to accurately predict customer churn may result in revenue loss, increased customer acquisition expenses, and weakened long-term strategic planning. To address the challenges of customer churn, machine learning-based predictive analytics has become increasingly popular in both academic research and industrial applications. By leveraging historical customer data, machine learning models can uncover complex relationships among variables, identify behavioral patterns, and generate predictions that help organizations formulate effective retention strategies. Numerous classification methods have been explored for churn prediction, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Neural Network-based approaches. The predictive effectiveness of these algorithms often varies with dataset characteristics, feature representations, and preprocessing strategies [5] [6]. Previous studies have reported that ensemble learning techniques, such as Random Forest, AdaBoost, and XGBoost, generally achieve strong predictive performance by reducing model variance and improving generalization. In addition, deep learning and sequential modeling approaches have shown considerable potential for capturing complex customer behavior, although they may require greater computational resources and larger datasets [7], [8]. Given the diversity of available approaches and their varying performance across datasets, conducting a comparative evaluation of multiple algorithms is essential to determine the most suitable model for customer churn prediction [9].

Despite the growing adoption of machine learning methods for customer churn prediction, several important issues remain insufficiently explored in the existing literature. Previous research by Triyafebrianda et al. compared the predictive capabilities of multiple classification algorithms; however, the study did not examine whether feature selection techniques could yield further performance improvements. Similarly, Mandurkar and Khanke reported the positive impact of feature selection on churn prediction models, yet their work did not specifically examine its influence on Logistic Regression and Random Forest using the Telco Customer Churn dataset, which is widely recognized as a benchmark dataset in churn prediction research. Furthermore, studies published during 2024–2025 have predominantly focused on enhancing predictive accuracy or evaluating individual machine learning algorithms. Comparatively, limited attention has been given to assessing how Recursive Feature Elimination (RFE) affects the performance of Logistic Regression and Random Forest models. Another limitation observed in several earlier studies is the reliance on a single validation technique, which may reduce the reliability and generalizability of the evaluation results due to potential validation bias.

To address the gaps identified in the existing literature, this study focuses on two main research objectives. The first objective is to examine the impact of Recursive Feature Elimination (RFE) on the predictive effectiveness of Logistic Regression and Random Forest models for customer churn prediction. The second objective is to determine which algorithm delivers the most reliable predictive performance following feature selection, as measured by accuracy, recall, precision, and F1-score. A notable gap in the existing literature is the limited number of studies that systematically examine how Recursive Feature Elimination (RFE) affects the predictive performance of Logistic Regression and Random Forest for telecommunications customer churn prediction. Moreover, comprehensive evaluations using the Telco Customer Churn dataset, together with cross-validation techniques, remain relatively uncommon. Therefore, this study seeks to assess the effectiveness of RFE by comparing the performance of Logistic Regression and Random Forest models before and after feature selection. To ensure robust and unbiased evaluation results, a 5-fold cross-validation approach is employed throughout the experimental process. The contribution of this research lies in providing empirical evidence regarding the impact of feature selection on changes in accuracy, recall, precision, and F1-score across two algorithms with different characteristics: an interpretable linear model (Logistic Regression) and a tree-based ensemble model (Random Forest). The findings of this study are expected to serve as a valuable reference for practitioners and researchers by providing empirical evidence regarding the effectiveness of churn prediction models, thereby supporting the adoption of more accurate and efficient customer retention strategies in the telecommunications sector.

The present study investigates the performance of two well-established classification algorithms, namely Logistic Regression and Random Forest. Logistic Regression is frequently used for its interpretability, computational efficiency, and ability to provide clear insights into the relationship between predictor variables and churn outcomes. In contrast, Random Forest leverages an ensemble learning framework in which multiple decision trees are constructed and aggregated to generate final predictions. This approach enhances the model's

ability to represent non-linear relationships, increases prediction stability, and reduces susceptibility to overfitting, making it a popular choice for customer churn prediction applications [10] [11].

To optimize model performance, this research employs Recursive Feature Elimination (RFE) for feature selection. The primary objective of RFE is to identify the most influential predictors by repeatedly training a model, evaluating feature importance, and removing attributes that contribute minimally to the prediction task. By focusing on a smaller yet more informative set of variables, RFE can enhance computational efficiency and improve the generalization capability of classification algorithms [12]. Recent findings in churn prediction research suggest that effective feature selection contributes not only to higher predictive accuracy but also to the construction of more streamlined and interpretable models. Compared with other feature selection methods, RFE is widely recognized for its ability to identify an optimal feature subset, thereby facilitating the development of more accurate predictive models. Furthermore, integrating feature selection into machine learning pipelines has produced more efficient models suitable for large-scale implementation without compromising predictive performance. Several recent studies support the effectiveness of this approach. For example, a study in the banking sector reported that Logistic Regression achieved 76.85% accuracy, whereas Random Forest achieved 83.12% [13]. Likewise, research in the telecommunications sector revealed that applying feature selection and hyperparameter tuning led to a notable improvement in Decision Tree performance, increasing accuracy from 74.09% to 80.05% [14]. Additional evidence was provided by a telecommunications website-based churn prediction study, where Random Forest attained an accuracy of 95%, exceeding the 92% accuracy achieved by Decision Tree [15], These results collectively indicate that optimization strategies and algorithm selection play a critical role in enhancing the reliability and effectiveness of churn prediction models.

The analysis in this study is based on the Telco Customer Churn Dataset, which is publicly available on Kaggle. This dataset contains a diverse range of variables describing customer demographics, service consumption patterns, and subscription-related information, allowing for a comprehensive examination of factors associated with customer churn. Owing to its richness and widespread use in churn prediction research, the dataset serves as an appropriate benchmark for evaluating the performance of classification models in a telecommunications context. Moreover, the dataset's characteristics closely reflect operational conditions commonly observed in real-world telecommunications environments.

All analytical processes were conducted using RapidMiner, a widely recognized data science platform that offers a visual interface and supports various machine learning algorithms. RapidMiner serves not only as a platform for developing predictive models but also as an environment that supports optimization and comprehensive performance evaluation. These capabilities enable researchers to build models that achieve improved predictive accuracy while maintaining efficient analytical workflows. From an academic perspective, this research is expected to contribute to the ongoing development of data mining and machine learning literature by providing a comprehensive assessment of feature selection optimization in customer churn prediction. The comparative analysis of Logistic Regression and Random Forest offers valuable insights into the strengths and limitations of these algorithms, thereby supporting future research and practical applications in predictive analytics. From a practical perspective, the findings are expected to help telecommunications companies more accurately identify customers at risk of churn. Such insights can support the development of retention strategies, including loyalty programs, personalized services, and service quality improvements, ultimately enhancing customer loyalty, operational efficiency, and long-term profitability.

Therefore, this study is important from both academic and practical perspectives. Academically, it enriches the existing literature on churn prediction by comparing two widely used classification algorithms within the telecommunications domain. Beyond its academic contribution, this research offers practical implications for telecommunications service providers by supporting the development of more effective customer retention strategies in an increasingly competitive and dynamic business environment. To address the identified gaps in the literature, this study investigates the comparative performance of Logistic Regression and Random Forest models for telecommunications customer churn prediction, incorporating feature selection optimization to enhance predictive effectiveness.

2. METHOD

The study was conducted using a comparative research design involving a series of systematic stages, including requirements analysis, data preparation, classification model comparison, and performance assessment. The overall workflow adopted in this research is depicted in Figure 1.

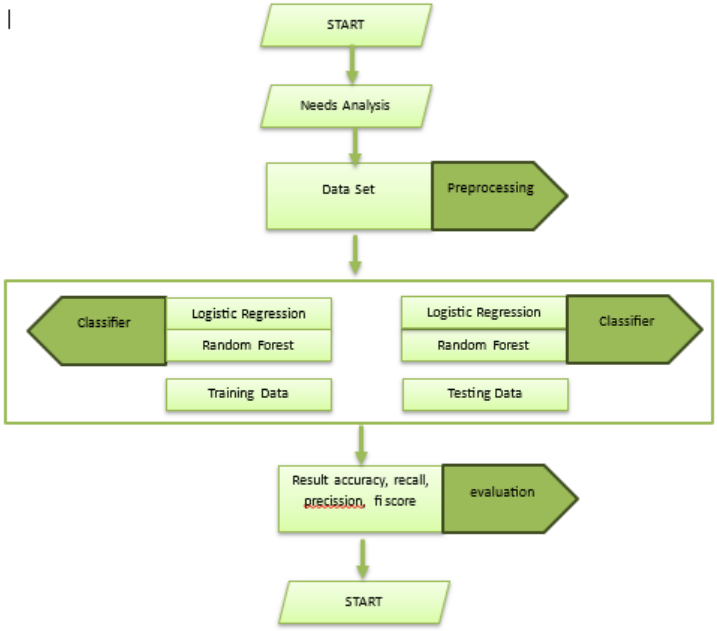

Figure 1. alur tahapan penelitian

2.1 Requirements Analysis

The first phase of the research focused on investigating the challenges of customer churn prediction in the telecommunications domain and on establishing an appropriate methodological framework. To address these challenges, machine learning techniques enhanced with feature-selection optimization were identified as the primary analytical approach for this study. This stage serves as the foundation for designing an accurate, efficient, and practically relevant predictive system that addresses both industrial and academic objectives.

In response to the challenges highlighted in the literature review and problem formulation, this study introduces a customer churn prediction framework that combines Logistic Regression and Random Forest classification models with Recursive Feature Elimination (RFE). The inclusion of RFE is intended to optimize feature representation by selecting the most informative attributes, thereby improving model efficiency and predictive performance. This approach is expected to improve both predictive accuracy and computational efficiency by retaining only the most relevant features during the modeling process.

An exploratory analysis of the Telco Customer Churn dataset revealed 7,043 customer instances described by 21 attributes, encompassing both numerical and categorical data types. This dataset was selected because it is publicly accessible, widely recognized as a benchmark in churn prediction research, and representative of operational conditions commonly encountered in the telecommunications sector. Furthermore, its extensive adoption in prior studies facilitates meaningful comparisons between this research's results and existing findings in the literature. The complete list of attributes utilized in this study is presented in Table 1.

Table 1. Dataset Attributes Description

No.	Atribut	Deskripsi
1.	custumerID	Unique customer identification number.
2.	gender	Indicates the customer's gender category (Male or Female).
3.	SeniorCitizen	Indicates whether a customer is a senior citizen (1 = Yes, 0 = No).
4.	Partner	Indicates whether the customer has a spouse or partner.
5.	Dependents	Whether the customer has dependents (Yes or No)
6.	tenure	The duration of the customer's subscription measured in months.
7.	PhoneService	Indicates whether telephone service is available in the customer's subscription plan.
8.	MultipleLines	Specifies whether the customer subscribes to more than one telephone line.
9.	InternetService	The type of internet connection the customer is subscribed to, such as DSL, Fiber Optic, or no internet service.
10.	OnlineSecurity	Indicates whether the customer subscribes to the online security service
11.	OnlineBackup	Represents the customer’s enrollment in the online backup service
12.	DeviceProtection	Indicates whether device protection service is included in the customer's subscription.
13.	TechSupport	Specifies the availability of technical support services for the customer.

14.	StreamingTV	Indicates whether the customer uses the television streaming service.
15.	StreamingMovies	Represents the customer's subscription status for movie streaming services.
16.	Contract	Describes the contract arrangement selected by the customer, including month to month, one year, or two-year plans
17.	PaperlessBilling	Indicates whether the customer utilizes an electronic billing system instead of paper invoices.
18.	PaymentMethod	Specifies the customer's payment method, including Electronic Check, Mailed Check, Automatic Bank Transfer, or Automatic Credit Card Payment.
19.	MonthlyCharges	The amount charged to the customer each month for subscribed services.
20.	TotalCharges	The cumulative amount paid by the customer throughout the subscription period.
21.	Churn	Target variable indicating whether the customer discontinued the service (Yes) or remained subscribed (No).

Preliminary exploratory analysis highlighted several notable patterns within the dataset. The tenure variable indicated substantial variation in customer retention periods, with an average subscription duration of approximately 32 months and a maximum value of 72 months. The average monthly service charge was USD 64.76, reflecting the typical cost customers incur for telecommunications services. Furthermore, demographic analysis showed that senior citizens represented roughly 16% of the overall customer base.

Analysis of the class distribution revealed a moderate imbalance in the target variable. The non-churn category comprised approximately 73.5% of customer records, while the churn category accounted for the remaining 26.5%. This distribution indicates that customers who maintained their subscriptions substantially outnumbered those who discontinued the service. This imbalance is an important consideration because it may influence classification performance and evaluation outcomes.

Correlation analysis among numerical variables revealed weak to moderate relationships. The exploratory analysis identified a positive relationship between customer tenure and monthly charges, indicating that prolonged subscription periods are generally associated with higher service expenditures. Furthermore, customers classified as senior citizens had slightly higher average monthly charges. In contrast, the association between senior citizen status and tenure was negligible, suggesting that age category had little influence on the length of customer subscriptions.

Outlier detection showed no significant anomalies in either the tenure or MonthlyCharges variables, suggesting a relatively consistent dataset. The only apparent anomaly was observed in the SeniorCitizen variable; however, this was attributed to its binary nature (0 or 1), making Interquartile Range (IQR)-based outlier detection less applicable.

Overall, the EDA results indicated that the dataset was relatively clean and contained no missing values. Despite a moderate imbalance in the class distribution, the proportions of churn and non-churn observations were considered sufficiently balanced to avoid resampling techniques such as SMOTE, undersampling, or oversampling. To ensure a comprehensive assessment of model effectiveness, the performance evaluation was not limited to accuracy but also included precision, recall, and F1-score, which are widely recognized as more informative measures for imbalanced classification tasks. In addition, a 5-fold cross-validation procedure was implemented to improve the reliability of the evaluation and minimize potential bias from data partitioning.

2.2 Data Preprocessing

Prior to the classification stage, data preprocessing was performed to ensure that the dataset possessed an appropriate format, structure, and scale for optimal machine learning performance [16]. This stage aimed to improve data quality and minimize potential errors during model development.

Several preprocessing steps were conducted in this study. First, the dataset was examined using the Filter Examples operator in RapidMiner to identify and handle potential missing values, although the EDA results indicated that the dataset was relatively clean. Second, the dataset was prepared for model training and testing.

Subsequently, cross-validation was employed to obtain a more stable performance estimate, reduce evaluation bias, and improve the reliability of model assessment. Consequently, the study utilized a Cross Validation operator instead of a conventional data-splitting approachSelanjutnya, digunakan metode *Cross validation* [17]. The selection of 5-fold cross-validation was based on its balance between evaluation quality and computational efficiency, as recommended in the machine learning literature [18].

2.3 Classification Model Comparison

Within the supervised learning paradigm, classification techniques are designed to categorize observations into predefined classes by learning patterns from historical data. To investigate the effectiveness of different predictive approaches, this study compares the performance of Logistic Regression and Random Forest models. The evaluation was conducted both with and without feature-selection optimization, enabling a comprehensive assessment of how feature reduction affects classification performance in customer churn prediction.

a. Logistic Regression

Among various classification methods, Logistic Regression remains one of the most frequently utilized approaches for customer churn prediction due to its interpretability and probabilistic modeling capability. By estimating the likelihood of churn occurrence based on a set of predictor variables, the algorithm provides valuable insights into the factors influencing customer behavior. Furthermore, its relatively simple structure makes it suitable for both predictive analysis and result interpretation [20]. The Logistic Regression model is mathematically defined as follows:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X 1 + \cdots + \beta n X n}}$$

b. Random Forest

Random Forest is a widely adopted ensemble learning method derived from the decision-tree paradigm. The algorithm operates by constructing multiple decision trees from randomly selected subsets of observations and predictor variables, then combining their outputs to produce a final classification result. This aggregation mechanism enhances model stability and generalization capability, making Random Forest less susceptible to overfitting than a single decision tree. Consequently, the algorithm is particularly effective for handling complex datasets characterized by non-linear relationships and interactions among variables.

c. Feature Selection

In machine learning and data mining applications, feature selection plays a vital role in enhancing the effectiveness of predictive models. The main purpose of this process is to identify a subset of highly relevant features while excluding variables that offer limited predictive value or introduce redundancy. By reducing the presence of non-informative attributes, feature selection can improve model interpretability, decrease computational complexity, and support better predictive performance. Through this process, classification models can achieve improved accuracy, greater computational efficiency, and enhanced interpretability

The classification models were implemented in RapidMiner under the default parameter configuration provided by the software. Logistic Regression was trained with L2 regularization to control model complexity, and parameter optimization was performed using the solver integrated into RapidMiner. In contrast, the Random Forest algorithm was configured to construct 100 decision trees, each generated from bootstrap-resampled training data to enhance model diversity and predictive robustness.

Default parameters were intentionally maintained to ensure a fair comparison between algorithms and to focus the study on evaluating the impact of feature selection optimization using Recursive Feature Elimination (RFE). Consequently, any observed performance improvements are more likely attributable to feature selection rather than hyperparameter tuning.

2.4 Model Evaluation

To assess the effectiveness of the proposed classification models, a comprehensive evaluation was conducted using several performance metrics. Among these measures, accuracy served as a fundamental indicator of model performance by quantifying the proportion of correctly predicted observations relative to the overall number of records evaluated.

Despite its widespread use, accuracy does not always provide a complete representation of model performance, particularly when dealing with datasets characterized by unequal class distributions. Under such conditions, a model may achieve high accuracy while performing poorly in identifying minority-class instances. Therefore, this study complemented accuracy with precision, recall, and F1-score metrics to ensure a more thorough and reliable evaluation of the classification models.

Accuracy is calculated as follows:

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \; x \; 100$$

Dimana :
True Positive (TP) = Positive instances correctly classified as positive.
False Positive (FP) = Negative instances incorrectly classified as positive.
True Negative (TN) = Negative instances correctly classified as negative.
False Negative (FN) = Positive instances incorrectly classified as negative.

Precision is a performance metric that quantifies the reliability of positive classifications produced by a predictive model. Specifically, it measures the proportion of correctly predicted positive cases relative to all instances assigned to the positive class:

$$Precision = \frac{TP}{TP + FP}$$

Recall, also known as the true positive rate or sensitivity, measures the extent to which a predictive model successfully recognizes positive cases. The metric is determined by comparing the number of correctly identified positive instances with the total number of actual positive observations:

$$Recall = \frac{TP}{TP + FN}$$

Because precision and recall often exhibit a trade-off relationship, the F1-score was employed as a balanced measure through the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

A comprehensive performance assessment was carried out for each classification model prior to and following the optimization process. The objective was to determine which algorithm provided the most effective solution for customer churn prediction. Evaluation was based on a combination of performance indicators, emphasizing not only predictive accuracy but also the ability of the models to recognize meaningful data patterns and achieve balanced results across different evaluation metrics.

A Wilcoxon Signed-Rank Test was employed to examine the statistical significance of the performance differences observed between the Logistic Regression and Random Forest models. The test was conducted using evaluation results collected from the five folds generated during the cross-validation procedure, allowing for a robust comparison of model performance across multiple data partitions. The statistical analysis was conducted at a significance level of 5% (α=0,05).

3. RESULT AND DISCUSSION

Model evaluation was conducted for each algorithm before and after the optimization stage to identify the classification method that achieved the best predictive performance. The performance assessment was not limited to overall predictive accuracy. Particular attention was also given to the effectiveness of each model in detecting churn cases, as the accurate identification of customers at risk of leaving represents a critical objective in customer churn prediction. Therefore, the evaluation emphasized metrics capable of reflecting the model's performance in recognizing the churn class alongside its general classification capability.

3.1. Research Results

Model validation in this research was performed using the cross-validation technique, which is widely recognized for assessing predictive performance and model generalization. The dataset was partitioned into five equally sized folds, with each fold serving as the testing set once while the remaining folds were used for model training. This iterative procedure ensures that every observation contributes to both the training and testing phases, thereby reducing the likelihood of biased evaluation results. To obtain a more stable and representative assessment of model effectiveness, a 5-fold cross-validation framework was employed, and the performance metrics from all iterations were subsequently aggregated.

Table 2 presents the classification performance achieved by the telecommunications customer churn prediction models before the implementation of feature selection optimization. The reported results serve as a baseline for evaluating the impact of the subsequent optimization process:

Table 2. Classification Performance Before Optimization

No.	Algoritma	Before Optimization			
		Akurasi	Recall	Precision	F1-Score
1.	Logistic Regression	73,46%	50%	36,73%	42,35%
2.	Random Forest	79.85%	69.75%	74,83%	72,15%

Before conducting the classification experiments, feature selection was performed using Recursive Feature Elimination (RFE) to identify the attributes with the strongest predictive relevance to the churn variable. The objective of this process was to reduce feature redundancy and retain only the most informative variables for subsequent model development. The original dataset consisted of 20 predictor attributes and one target attribute, namely Churn. Through an iterative elimination process, RFE evaluated the contribution of each attribute to model performance and removed features with the lowest predictive importance.

The results of the feature selection stage revealed that only 10 attributes were retained as the optimal feature set for subsequent classification experiments. This reduction in the number of predictor variables was designed to simplify the model structure, enhance computational efficiency, and improve generalization performance by focusing on the most informative features while minimizing the influence of redundant or less relevant attributes.

Table 3. Features Selected by Recursive Feature Elimination (RFE)

No.	Fitur Yang Terpilih Oleh RFE
1.	Senior Citizen
2.	Dependents
3.	Tenure
4.	Phone Service
5.	Multiple Lines
6.	Online Security
7.	Online Backup
8.	Tech Support
9.	Contract
10.	Paperless Billing

After removing the CustomerID attribute, which does not contribute to the classification process, the dataset consisted of 19 predictor variables. The RFE analysis selected ten attributes as the optimal predictors for model construction: SeniorCitizen, Dependents, Tenure, PhoneService, MultipleLines, OnlineSecurity, OnlineBackup, TechSupport, Contract, and PaperlessBilling. Through this process, the original set of predictor variables was reduced by 47.37%, resulting in a more streamlined feature space. Despite the reduction, the selected attributes retained the essential information required by the classification algorithms, ensuring that predictive performance was maintained while model complexity was minimized.

These findings indicate that attributes related to customer tenure, contract type, security services, technical support services, and customer characteristics contribute substantially to determining the likelihood of customer churn. Conversely, attributes such as TotalCharges, StreamingTV, StreamingMovies, and Gender demonstrated relatively lower contributions and were therefore excluded during the feature selection process.

The implementation of RFE was intended to reduce model complexity, improve computational efficiency, and eliminate irrelevant attributes, enabling the classification models to focus on features that have the greatest influence on customer churn prediction.

The classification results after feature selection optimization are presented in Table 4.

Table 4. Classification Performance After Optimization

No.	Algoritma	Setelah Optimasi			
		Akurasi	Recall	Precision	F1-Score
1.	Logistic Regression	79,61%	54,36%	63,58%	58,6%
2.	Random Forest	81%	58%	71%	63,85%

Based on the results presented in Table 4, both Logistic Regression and Random Forest demonstrated different levels of effectiveness in classifying customers who were likely to churn. Performance evaluation was conducted using four primary metrics: accuracy, recall, precision, and F1-score, which collectively provide a comprehensive assessment of the models' ability to identify customers at risk of churn.

For Logistic Regression, a substantial improvement in performance was observed after optimization. Accuracy increased from 73.46% to 79.61%, indicating that the model became more effective in correctly classifying both churn and non-churn customers. Recall increased from 50.00% to 54.36%, suggesting enhanced sensitivity in identifying customers who actually churned. This improvement is particularly important in the telecommunications industry, where failing to identify churn-prone customers may result in direct revenue losses.

Furthermore, precision increased significantly from 36.73% to 63.58%, indicating that a greater proportion of customers predicted as churners were indeed actual churners. In other words, the optimized model generated fewer false positives, allowing retention resources to be allocated more effectively toward genuinely at-risk customers. This improvement contributed to an increase in the F1-score from 42.35% to 58.60%, reflecting a better balance between recall and precision. Overall, these results suggest that the optimized Logistic Regression model became considerably more reliable for churn prediction.

In contrast, Random Forest exhibited relatively stable performance but did not experience substantial improvement after optimization. Although accuracy increased slightly from 79.85% to 81.00%, recall decreased from 69.75% to 58.00%, and precision declined from 74.83% to 71.00%. Consequently, the F1-score decreased from 72.15% to 63.85%. These results indicate that while Random Forest maintained a high overall accuracy, its sensitivity in identifying actual churn customers declined after optimization. From a business perspective, this may result in a larger number of high-risk customers remaining undetected, thereby reducing opportunities for effective retention interventions.

Overall, the evaluation results demonstrate that the implementation of RFE produced different effects on the two classification algorithms. Logistic Regression experienced significant performance improvements across all evaluation metrics, particularly in precision and F1-score. Meanwhile, Random Forest continued to achieve the highest overall predictive performance, with an accuracy of 81.00%, recall of 58.00%, precision of 71.00%, and F1-score of 63.85%. These findings suggest that RFE provided greater benefits for Logistic Regression, whereas Random Forest remained the best-performing model on the Telco Customer Churn dataset.

The confusion matrix results are presented in Table 5.

Table 5. Hasil Confussion Matrix

No.	Algoritma	Before Optimization				After Optimization			
		TN	FP	FN	TP	TN	FP	FN	TP
1.	Logistic Regression	65,77 %	7,70%	11,93 %	14,61 %	66,02 %	7,44 %	13,87 %	12,67 %
2.	Random Forest	79.85 %	69.75 %	74,83 %	72,15 %	81%	58%	71%	63,85 %

Based on the confusion matrix analysis, the optimized Random Forest model demonstrated a higher success rate in classifying non-churn customers than Logistic Regression. However, identifying churn customers remained challenging, as indicated by the relatively large proportion of false negatives. Therefore, the confusion matrix complements the accuracy, precision, recall, and F1-score metrics by providing a more comprehensive understanding of model performance.

In the context of telecommunications customer churn analysis, recall and F1-score are often considered more critical than accuracy alone because failing to identify churn customers (false negatives) may result in the loss of valuable customers without appropriate retention actions. Accordingly, these findings suggest that optimized Logistic Regression offers a more effective approach for early churn detection systems in the telecommunications industry.

3.2. Discussion

This study aimed to analyze and compare the performance of Logistic Regression and Random Forest in predicting customer churn within the telecommunications industry. Model evaluation was conducted using accuracy, recall, precision, and F1-score before and after optimization. These metrics provide a comprehensive perspective on model performance, not only in terms of overall prediction accuracy but also in identifying customers who are genuinely at risk of churn. The RFE results identified ten attributes that contributed most significantly to churn prediction: SeniorCitizen, Dependents, Tenure, PhoneService, MultipleLines, OnlineSecurity, OnlineBackup, TechSupport, Contract, and PaperlessBilling. These findings indicate that customer churn behavior is influenced by a combination of customer characteristics, subscription duration, and the extent of service utilization.

Among these variables, Tenure and Contract emerged as particularly important indicators because they are directly associated with customer loyalty and commitment. Customers with longer subscription

periods and long-term contracts generally exhibit lower churn tendencies. Similarly, OnlineSecurity, OnlineBackup, TechSupport, PhoneService, and MultipleLines represent the degree of customer engagement with the company's service ecosystem. The more services customers utilize, the higher the switching cost becomes, thereby reducing the likelihood of churn. Furthermore, SeniorCitizen, Dependents, and PaperlessBilling reflect customer demographic characteristics and interaction patterns with telecommunications services. The inclusion of these variables suggests that demographic and behavioral factors also influence customers' decisions to remain subscribed or switch to competing providers

Overall, the feature selection results demonstrate that RFE successfully identified the most relevant attributes for churn prediction, reducing the number of features from 21 to 10 without sacrificing essential predictive information. The experimental results revealed that Logistic Regression experienced substantial performance improvements following optimization. Accuracy increased from 73.46% to 79.61%, recall improved from 50.00% to 54.36%, precision rose from 36.73% to 63.58%, and F1-score increased from 42.35% to 58.60%. These improvements indicate that the optimized model became more effective in identifying churn customers while maintaining greater prediction reliability.

This improvement can be explained by the characteristics of Logistic Regression, which is particularly effective in modeling linear relationships between predictor variables and churn probability. By removing irrelevant and redundant features, RFE enabled the model to focus on the most informative predictors, thereby reducing noise and improving classification performance. In contrast, Random Forest exhibited relatively stable performance but did not benefit substantially from feature selection. Although accuracy increased slightly, recall, precision, and F1-score declined. One possible explanation is that Random Forest, as an ensemble-based algorithm, inherently handles high-dimensional datasets effectively and often benefits from a larger set of predictor variables. Consequently, removing certain features may have reduced the diversity of information available to the model, leading to lower performance on minority-class detection. Overall, the findings indicate that RFE provided greater performance improvements for Logistic Regression than for Random Forest. Logistic Regression achieved increases of 6.15% in accuracy, 4.36% in recall, 26.85% in precision, and 16.25% in F1-score. Nevertheless, when considering the final performance after optimization, Random Forest remained the strongest overall model, achieving higher values across all evaluation metrics. Therefore, while Logistic Regression benefited more substantially from feature selection, Random Forest maintained superior predictive performance on the Telco Customer Churn dataset.

From a practical perspective, these findings have important implications for customer retention strategies. By utilizing more accurate churn prediction models, telecommunications companies can proactively identify customers at high risk of discontinuing their subscriptions. This information can support targeted interventions, such as personalized offers, service quality improvements, and loyalty programs designed to increase customer retention and reduce customer acquisition costs. Moreover, the improvement in precision indicates that the models became more effective at minimizing false positives. As a result, retention campaigns can be focused on customers who are genuinely at risk of churn, thereby improving resource allocation efficiency and enhancing the effectiveness of data-driven marketing initiatives.

Although Random Forest achieved higher accuracy, precision, recall, and F1-score values than Logistic Regression, statistical testing was required to determine whether these differences were significant. Therefore, a Wilcoxon Signed-Rank Test was performed using the evaluation results obtained from each fold of the 5-fold cross-validation procedure. The test produced a p-value of 0.0625. At the 5% significance level ($\alpha = 0.05$), this value exceeds the significance threshold ($0.0625 > 0.05$), indicating that the performance difference between Logistic Regression and Random Forest is not statistically significant. Consequently, although Random Forest demonstrated superior empirical performance, its advantage over Logistic Regression cannot be considered statistically significant. This finding suggests that both algorithms provide comparable predictive capabilities for telecommunications customer churn prediction, with the choice of algorithm depending on organizational priorities such as interpretability, computational efficiency, and implementation requirements.

4. CLOSING

This study evaluated and compared the performance of two classification algorithms, namely Logistic Regression and Random Forest, in predicting customer churn in the telecommunications industry through the application of feature selection optimization. Based on the results presented in the Results and Discussion section, the findings demonstrate that feature selection optimization can improve classification performance, particularly for the Logistic Regression algorithm.

The consistency between the research objectives outlined in the Introduction and the findings reported in the Results and Discussion section is reflected in the improvements observed in accuracy, recall, precision, and F1-score following optimization. The optimized Logistic Regression model achieved the most balanced performance in terms of sensitivity and predictive precision, making it effective for identifying customers with

a high risk of churn. In contrast, although Random Forest maintained relatively high predictive accuracy, its recall performance declined after optimization. These findings support the research hypothesis that the effectiveness of customer churn prediction models is strongly influenced by both the feature selection technique employed and the characteristics of the underlying classification algorithm.

From a practical perspective, the findings provide valuable insights for telecommunications companies in designing more effective customer retention strategies. By utilizing the optimized Logistic Regression model, organizations can proactively identify customers who are likely to discontinue their subscriptions and implement targeted interventions, such as personalized offers, service quality improvements, and customer loyalty programs. Such initiatives can help reduce churn rates while improving the efficiency of resource allocation. Preventive retention measures may include personalized retention strategies, such as offering customized service packages at competitive prices, improving network quality in areas with high-risk customers, or providing incentives through loyalty points and bonus data quotas. These personalized approaches not only increase customer retention opportunities but also strengthen customer engagement with the brand. In the long term, such strategies may transform at-risk customers into loyal customers and potential brand advocates.

Beyond churn prevention, the predictive model may also support customer acquisition initiatives. Analyzing the characteristics of customers who are more likely to remain loyal can provide valuable insights into the profiles of high-value prospective customers. For example, if the model indicates that customers with high data usage and longer subscription periods are less likely to churn, telecommunications companies can target similar customer segments through more focused marketing campaigns. This approach enables organizations to prioritize customer quality over acquisition volume by attracting customers with greater long-term value and stronger loyalty potential. From a managerial perspective, the findings also contribute to data-driven decision-making processes. Compared with more complex algorithms such as Random Forest, Logistic Regression offers greater interpretability and transparency, making it easier for decision-makers to understand the factors influencing customer churn. Consequently, management can more effectively formulate policies related to customer service quality, network performance, pricing strategies, and customer relationship management initiatives.

Within the telecommunications industry, these findings have important implications for Customer Relationship Management (CRM) strategies. Models with strong recall and F1-score performance are particularly valuable because organizations must identify as many potential churn customers as possible to implement timely retention actions. Failing to identify actual churn customers (false negatives) may result in the loss of valuable customers without intervention, whereas incorrectly identifying non-churn customers as churners (false positives) primarily leads to inefficient allocation of promotional resources. Therefore, the optimized Logistic Regression model provides a favorable balance between effectiveness and operational efficiency in supporting business decision-making.

From an academic perspective, this study provides empirical evidence regarding the effectiveness of integrating classification algorithms with feature selection optimization in customer churn prediction. The findings contribute to the growing body of knowledge in predictive analytics and customer behavior modeling and may serve as a foundation for future research in these fields. Several opportunities for future research can be identified. First, future studies may investigate more advanced machine learning algorithms, such as Gradient Boosting, XGBoost, or Deep Neural Networks, to develop models that are more adaptable to complex nonlinear patterns. Second, incorporating comprehensive hyperparameter optimization techniques, including Grid Search and Bayesian Optimization, may further improve predictive performance. Third, the proposed churn prediction framework can be applied to other domains, such as banking, e-commerce, and digital services, to evaluate the consistency and generalizability of the findings across different industries.

In conclusion, this study not only achieved its primary research objectives but also provides broader opportunities for the advancement of machine learning methodologies and their practical applications in supporting data-driven strategic decision-making within highly competitive industries.

DAFTAR PUSTAKA

[1] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra and F. Benson, "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models," *Algorithms*, p. 231, 2024.

[2] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard and J. Qadir, "Customer churn management in the telecommunications sector: A predictive analytics approach," *J. Bus. Analytics*, vol. 7, no. 2, p. 112–129, 2024.

[3] H. Ribeiro, B. Barbosa, A. C. Moreira and R. G. Rodrigues, "Determinants of churn in telecommunication services: a systematic literature review," vol. 2024, no. 74, pp. 1327-1364, 2023.

[4] H. A. Triyafebrianda and N. A. Windasari, "Factors Influence Customer Churn on Internet Service Provider in Indonesia," *TIJAB*, vol. 6, no. 2, 2022.

[5] A. Mandrurkar and S. Khanke, "Churn Prediction Using Various Machine Learning Algorithms and Feature Selection," *International Conference on Emerging Trends in Smart Technologies (ICETST)*, 2022.

[6] R. Patil, "Customer Churn Prediction using Machine Learning in E-commerce," *International Journal of Innovative Research in Computer Science & Technology*, 2024.

[7] P. Raj and S. Gupta, "The Efficacy of Customer Churn Prediction System Using Different Machine Learning Models," *International Journal of Scientific & Technology Research*, 2024.

[8] R. Dodda and S. Raghavendra, "A Comparative Study of Machine Learning Algorithms for Custom-er Churn Prediction," *International Conference on Artificial Intelligence and Data Engineering (AIDE)*, 2024.

[9] S. Pulkundwar and A. Rudani, "A Comparison of Machine Learning Algorithms for Customer Churn Prediction," *International Journal of Computer Science and Mobile Computing*, vol. 12, no. 4, pp. 45-54, 2023.

[10] N. Waningsih, A. S. Akbar, S. Ariska, R. F. Husnayaini, E. Nurrin, R. Ridho and F. T. P. Situmorang, "Classification-Based Supervised Learning Algorithms for Accurate Prediction of Customer Churn in Banking," *IJATIS*, vol. 3, no. 1, pp. 41-49, 2026.

[11] F. Zulfikar and S. Anggoro, "Prediksi Churn Pelanggan Telekomunikasi Menggunakan Algoritma Random Forest dengan Penerapan Tree Prunning," *JISBT*, 2026.

[12] A. M. Prayitno and T. Widiyaningtyas, "A SYSTEMATIC LITERATURE REVIEW: RECURSIVE FEATURE ELIMINATION ALGORITHMS," *JITK*, vol. 9, no. 2, pp. 196-207, 2024.

[13] E. Mufida, D. Andriansyah and H. Hertyana, "Customer churn prediction pada sektor perbankan dengan model Logistic Regression dan Random Forest," *Comput. Sci. (CO-SCIENCE)*, vol. 5, no. 1, 2025.

[14] S. Antoh, I. Herteno, I. Budiman, D. Kartini and M. Itqan, "Prediksi churn pelanggan telekomu-nikasi dengan optimalisasi seleksi fitur dan tuning hyperparameter pada algoritma klasifikasi C4.5," *J. Syst. Inf. Bus*, vol. 15, no. 1, 2025.

[15] Holilurrahman and M. Imron, "Implementasi Model Prediksi Churn Pelanggan menggunakan Algoritma Random Forest Pada Website Industri Telekomunikasi," *Jurnal Teknologi Informasi (AJTI)*, vol. 1, no. 1, 2025.

[16] F. Admojo and S. Jabir, "Analisis Performa Metode Naive Bayes Clasifier pada Eectronic Nose Dalam Identifikasi Formalin Pada Tahu," *Indones.J. Data SCI*, vol. 4, pp. 1-16, 2023.

[17] R. Sudriyanto and M. Hariri, "Implementasi Algoritma Decision Tree (C4.5) dengan Optimize Wheights (PSO) Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu," *Jurnal Informatika Universitas Pamulang*, vol. 6, pp. 252-257, 2021.

[18] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statical Learning (2nd ed), Springer, 2021.

[19] A .Fauzi, "Analisis Data Bank Direct Marketing dengan Perbandingan Klasifikasi Data Mining Berbasis Optimize Selection (Evolutionary)," *Jurnal Informatika Universitas Pamulang*, vol. 6, p. 102, 2021.

[20] A. H. Yunial, "Analisis Optimasi Algortima Klasifikasi Support Vector Machine, Decision Trees, dan Neural Network Menggunakan Adaboost dan Bagging," *Jurnal Informatika Universitas Pamulang*, vol. 5, p. 247, 2020.