

Comparison of Naïve Bayes and Random Forest Algorithms for Customer Churn Prediction in Telecommunication Services

Hikmal Maulana^{1*}, Achmad Nurhadi², Aji Sumandito³

^{1,2,3} Informatika, Bina Sarana Informatika University

hikmalmaulana212@gmail.com^{1*}, achmad.ahh@bsi.ac.id², aji.ajs@bsi.ac.id³

Abstract

Customer churn refers to the phenomenon of customers discontinuing a service, posing a serious challenge for telecommunication companies because it directly affects revenue and company reputation. This study compares the performance of the Naïve Bayes and Random Forest algorithms in predicting customer churn in telecommunication services and analyzes the effect of feature selection on their performance. The dataset used is the Telco Customer Churn dataset obtained from Kaggle, consisting of 7,043 customer records with 21 attributes. A supervised learning approach was employed, with preprocessing stages including data cleaning, feature selection using the Gain Ratio method, and data splitting into 80% training and 20% testing data. Modeling and evaluation were conducted using Orange Data Mining software. The results show that Random Forest outperformed Naïve Bayes after feature selection, achieving an accuracy of 0.783, precision of 0.773, recall of 0.783, and F1-score of 0.776, while Naïve Bayes achieved an accuracy of 0.747, precision of 0.801, recall of 0.747, and F1-score of 0.760. Feature selection improved Naïve Bayes performance while slightly reducing Random Forest performance. The Contract, OnlineSecurity, and TechSupport attributes were identified as the most influential features for predicting customer churn.

Keywords: Customer Churn, Feature Selection, Machine Learning, Naïve Bayes, Random Forest

1. Introduction

The rapid development of information and communication technology has driven increasing public demand for telecommunication services, marked by the growing use of digital devices to support communication activities and access to information [1]. This condition has also fueled increasingly intense competition in the telecommunications industry. Fast-changing technology, market liberalization, and the emergence of various service innovations and attractive offers from providers give customers many options, making it easier for them to switch to other service providers. This phenomenon of customers switching providers is known as customer churn [2]. Several factors contribute to churn, including customer dissatisfaction, service pricing, service quality, and the appeal of offers provided by competitors [3].

A previous study at one telecommunication company recorded an average customer churn rate of around 6.48% per month [4]. This high churn rate highlights the importance of companies understanding the factors that drive customers to stop using their services, so that effective strategies can be designed to reduce customer turnover and increase customer loyalty. Besides reducing revenue, poorly managed churn can also damage a company's image and cause it to lose its competitive advantage in the long run [5].

Along with technological developments, Machine Learning has become one of the approaches widely used in data analysis because it enables systems to learn patterns from data and generate predictions automatically. In the context of churn prediction, classification methods are a common technique used to identify customers who are likely to stop using a service [6]. Based on this problem, the present study aims to analyze and compare the performance of the Naive Bayes and Random Forest algorithms in predicting customer churn in telecommunication services. Customer data will be processed using classification methods to identify the patterns that influence customer decisions. The results of this study are expected to provide a recommendation for the most optimal algorithm and help companies make strategic decisions to improve customer retention and reduce churn rates.

2. Literature Review

2.1. Customer Churn

Customer Churn is a term used to describe a condition in which customers stop using a service, resulting in a decline in the number of customers. Customer Churn can also be defined as customers' tendency to discontinue using a service at a certain point in time. In addition, churn occurs when customers decide to switch from one service provider to another for various reasons. At present, the churn phenomenon is a serious concern for companies across various industry sectors because it is difficult to manage. [7]

Several factors cause customers to churn, including the desire to switch to another company offering similar services, as well as changes in customer needs such as upgrading or downgrading the service package used. The impact of Customer Churn can reduce a company's revenue. In addition, companies also need to spend considerable costs to acquire new customers to replace the ones they have lost. [8]

2.2. Naïve Bayes

Naïve Bayes is a relatively simple probability-based classification method, in which the process involves calculating a number of probabilities through the accumulation of frequencies and combinations of values within a dataset. This method is based on the concepts of probability and statistics first put forward by an English scientist named Thomas Bayes, with the aim of estimating the likelihood of a future event based on prior data or experience. In addition, Naïve Bayes relies on a basic assumption that every attribute in the data is mutually independent of the others, meaning that the value of one attribute is not influenced by the value of another output attribute. [9]

2.3. Random Forest

Random Forest is a development of the Decision Tree method, built by combining many decision trees into a single model. Each tree is formed from a different data sample, and attributes are selected randomly during the tree-building process. This approach offers several advantages, such as the ability to improve accuracy even when data is incomplete, greater resistance to outliers, and greater efficiency in data processing and storage. In addition, Random Forest is also able to perform automatic feature selection to determine the most relevant features, allowing the model's performance in the classification process to become more optimal. [10]

2.4. Machine Learning

Machine Learning (ML) is an approach within artificial intelligence (AI) used to imitate human behavior in decision-making and apply it automatically. Machine Learning works by completing a task without explicit programming, relying instead on experience. Therefore, the application of machine learning focuses on developing systems capable of learning independently to make decisions, without needing to be reprogrammed, and capable of adapting to changes that occur. In the process, Machine Learning uses large collections of data as input, which are then analyzed to discover certain patterns. [11]

2.5. Classification

Classification is the process of grouping data based on certain characteristics it possesses. This process involves a training stage to build a function with a specific target, which is then used to map each object into one of the predetermined class labels. Classification is a type of analysis that can be used to help determine the class label of a number of samples to be categorized. [12]

2.6. Related Research

Several previous studies have applied the Random Forest and Naïve Bayes algorithms, either on their own or compared against other algorithms, to predict customer churn. A summary of the related research used as reference material and a comparison basis for this study is presented in Table 1.

Table 1: Summary of Related Research

Ref.	Title	Research Focus	Key Result
[13]	Evaluation of Telecommunication Customer Churn Classification with SMOTE Using Random Forest and XGBoost Algorithms	Random Forest vs XGBoost with SMOTE on telecom churn records	XGBoost slightly edged ahead with 82% accuracy versus 81% for Random Forest.
[14]	Perbandingan Decision Tree dan Naive Bayes untuk Prediksi Customer Churn	Decision Tree vs Naive Bayes on customer churn prediction	Decision Tree topped the field at 65.04% accuracy; Naive Bayes trailed at 62.68%.
[6]	Perbandingan Optimasi Algoritma Random Forest Menggunakan Teknik Boosting Terhadap Kasus Klasifikasi Churn Pelanggan Di Industri Telekomunikasi	Boosting-optimized Random Forest for telecom churn classification	Random Forest with boosting reached 99.13% accuracy.
[15]	Churn Prediction Analysis of Telecom Customers Using SVM, Random Forest and Logistic Regression Models Using Orange Data Mining Tools	SVM, Random Forest, and Logistic Regression compared for telecom churn using Orange Data Mining	Logistic Regression led at 79% accuracy, ahead of Random Forest (76%) and SVM (74%).

[16]	Prediksi Customer Churn Pada Bisnis Retail Menggunakan Algoritma Naïve Bayes	Naïve Bayes for retail customer churn prediction	Naïve Bayes achieved 80% accuracy.
[17]	Penerapan Naïve Bayes untuk Prediksi Customer Churn (Studi Kasus: PT Hutchison 3 Indonesia)	Naïve Bayes churn prediction case study at a telecommunication provider	Naïve Bayes achieved 91.3% accuracy.

Based on Table 1, previous studies generally compared Random Forest or Naïve Bayes with other algorithms such as XGBoost, Decision Tree, SVM, or Logistic Regression, but few have specifically combined a comparison of these two algorithms with an analysis of the effect of feature selection using the Gain Ratio method on the Telco Customer Churn dataset through the Orange Data Mining software. This study seeks to fill that gap by comprehensively comparing the performance of Naïve Bayes and Random Forest before and after feature selection using four evaluation metrics.

3. Research Method

3.1. Research Stages

This study uses a quantitative approach consisting of five main stages: data collection, data preprocessing, machine learning modeling, model evaluation, and analysis of results. The entire process was carried out using the Orange Data Mining software, which enables the analysis to be performed visually through a drag-and-drop scheme between widgets.

3.2. Research Dataset

The dataset used is the Telco Customer Churn dataset obtained from the Kaggle platform, consisting of 7,043 customer records with 21 attributes covering demographic information, services used, as well as contract and payment data. The target variable of this study is Churn, which indicates whether a customer has stopped (Yes) or remains subscribed (No).

3.3. Data Preprocessing

The data cleaning stage was carried out by handling missing values using the average method so that the amount of data remained intact. Next, feature selection was performed using the Gain Ratio method to select the attributes most relevant to churn prediction, reducing the number of attributes used from 21 to 11. The dataset was then split into training and testing data with an 80:20 ratio using the Data Sampler widget so that model performance could be objectively evaluated on data it had never seen before.

3.4. Modeling and Evaluation

Modeling was carried out using two algorithms, Random Forest and Naïve Bayes, each of which was trained on the dataset before and after feature selection to observe the effect of feature selection on model performance. Evaluation was carried out using the Test and Score and Confusion Matrix widgets in Orange Data Mining to obtain the accuracy, precision, recall, and F1-Score values for each model.

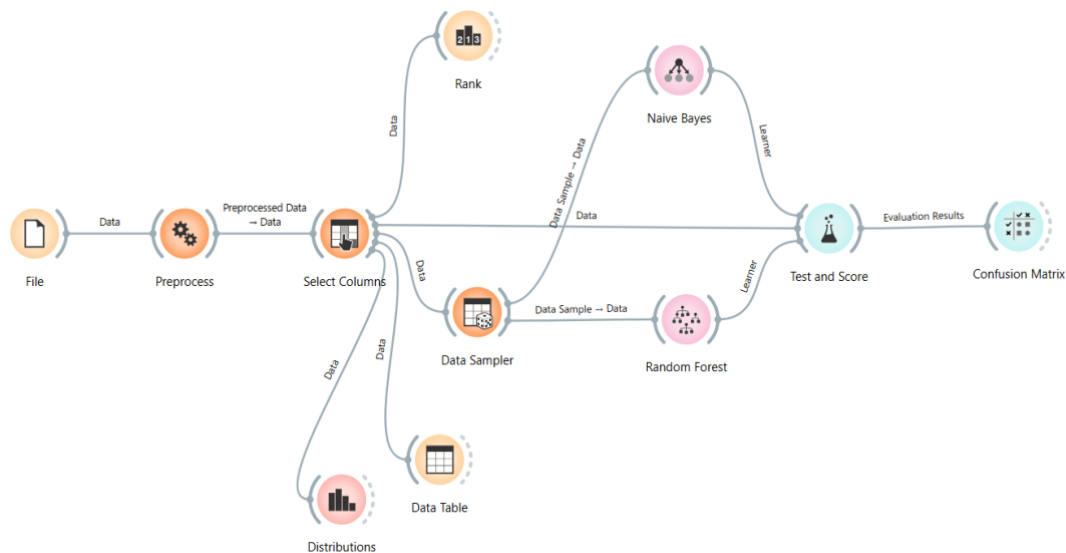


Fig. 1: Orange Data Mining Workflow for Customer Churn Prediction

Fig. 1 shows the visual programming workflow built in Orange Data Mining for this study. The File widget imports the dataset, which is then passed to Preprocess for missing-value handling and to Select Columns for attribute selection. The processed data branches into the Rank widget, which computes Gain Ratio scores for feature selection, into Data Table and Distributions for exploratory inspection, and into the Data Sampler widget, which splits the data into 80% training and 20% testing subsets. The resulting training and testing samples are fed into the Naïve Bayes and Random Forest learners, whose predictions are evaluated jointly through the Test and Score widget and further examined using the Confusion Matrix widget to obtain the accuracy, precision, recall, and F1-Score metrics reported in this study.

4. Results and Discussion

4.1 Data Distribution Results

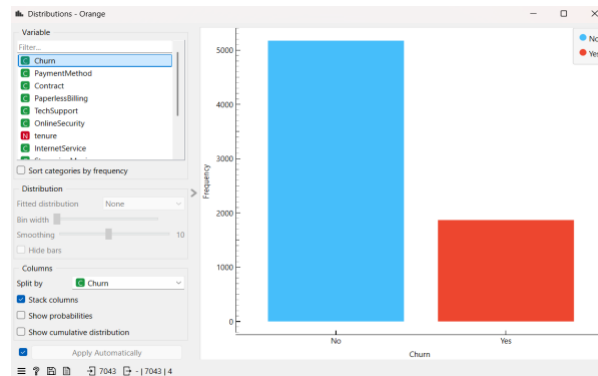


Fig. 2: Distribution of the Churn Target Variable

Fig. 2 presents the distribution of the Churn target variable, generated with the Distributions widget in the Orange workflow shown in Fig. 1. Of the 7,043 customer records, 5,174 customers (73.46%) were classified as not churned, while 1,869 customers (26.54%) were classified as churned, indicating that the dataset is imbalanced (class imbalance). This condition is the reason a combination of four evaluation metrics was used instead of accuracy alone, so that the model's performance in detecting the minority class could be assessed more objectively.

4.2 Feature Selection Results

Feature selection was performed using the Gain Ratio method to identify the attributes with the strongest influence on churn prediction, reducing the number of attributes used for modeling from 21 to 11. The resulting Gain Ratio scores are presented in Table 2.

Table 2: Gain Ratio Scores of the Selected Attributes

Attribute	Gain Ratio
Contract	0.099
OnlineSecurity	0.062
TechSupport	0.061
InternetService	0.052
Tenure	0.047
PaymentMethod	0.033
StreamingMovies	0.030
StreamingTV	0.030
PaperlessBilling	0.028
TotalCharges	0.020
MonthlyCharges	0.020

Contract obtained the highest score (0.099), followed by OnlineSecurity (0.062) and TechSupport (0.061), indicating that contract type, online security, and technical support are the strongest predictors of customer churn, while MonthlyCharges and TotalCharges contributed the least among the selected attributes.

4.3 Model Performance Comparison Results

To more comprehensively determine the effect of applying feature selection on the performance of both algorithms, a comparison of accuracy (CA), precision, recall, and F1-Score values was carried out before and after the feature selection process. The results of this comparison are presented in Table 3.

Table 3: Comparison of Model Performance Before and After Feature Selection

Algorithm	Condition	CA	Precision	Recall	F1-Score
Random Forest	Before	0,787	0,776	0,787	0,779
Naïve Bayes	Before	0,732	0,800	0,732	0,747
Random Forest	After	0,783	0,773	0,783	0,776
Naïve Bayes	After	0,747	0,801	0,747	0,760

Based on Table 3, the application of feature selection had a different impact on each algorithm. For the Random Forest algorithm, all metrics decreased only very slightly after feature selection was applied: accuracy dropped from 0.787 to 0.783, precision from 0.776 to 0.773, recall from 0.787 to 0.783, and F1-Score from 0.779 to 0.776. This very small decrease indicates that Random Forest's performance

is relatively stable and not significantly affected by the reduction in the number of attributes from 21 to 11, consistent with its nature as an ensemble method that builds many decision trees randomly and can therefore still capture data patterns even when the number of attributes is reduced.

In contrast to Random Forest, the Naïve Bayes algorithm actually showed an improvement across all metrics after feature selection was applied. Accuracy increased from 0.732 to 0.747, precision from 0.800 to 0.801, recall from 0.732 to 0.747, and F1-Score from 0.747 to 0.760. This shows that Naïve Bayes benefits more from the feature selection process than Random Forest, because eliminating irrelevant attributes helps the model estimate probabilities more accurately. Naïve Bayes is highly sensitive to irrelevant attributes because the model assumes that every attribute is independent of the others, so the presence of uninformative attributes can disrupt probability estimation and reduce prediction accuracy.

Nevertheless, the Random Forest algorithm still shows superior performance compared to Naïve Bayes on the accuracy, recall, and F1-Score metrics both before and after feature selection. Meanwhile, Naïve Bayes consistently recorded a higher precision value in both testing conditions. The high precision value of Naïve Bayes indicates that this model is more conservative in predicting customers as churn, so the positive predictions it produces tend to be more accurate. Conversely, Random Forest's superiority in recall shows a better ability to detect customers who actually churn, which is more relevant from a business perspective, since telecommunication companies suffer greater losses when churning customers go undetected by the model.

The confusion matrix results for both algorithms, before and after feature selection, are summarized in Table 4.

Table 4: Confusion Matrix Summary of Random Forest and Naïve Bayes

Algorithm	Condition	TN	FP	FN	TP
Random Forest	Before	4608	566	934	935
Naïve Bayes	Before	3646	1528	362	1507
Random Forest	After	4581	593	934	935
Naïve Bayes	After	3795	1379	404	1465

Random Forest consistently produced a higher number of False Negatives (934) than False Positives, meaning it more often failed to detect customers who actually churned. Naïve Bayes showed the opposite pattern, with a considerably higher number of False Positives (1,528 before and 1,379 after feature selection) but far fewer False Negatives, confirming its tendency toward higher recall for Random Forest and higher precision for Naïve Bayes discussed above.

5. Conclusion

This study compared the Naïve Bayes and Random Forest algorithms for customer churn prediction in telecommunication services using the Telco Customer Churn dataset. After feature selection, Random Forest achieved an accuracy of 0.783, precision of 0.773, recall of 0.783, and F1-score of 0.776, outperforming Naïve Bayes overall, which achieved an accuracy of 0.747, precision of 0.801, recall of 0.747, and F1-score of 0.760. Naïve Bayes, however, consistently achieved a higher precision than Random Forest both before and after feature selection. The Gain Ratio feature selection method identified Contract, OnlineSecurity, and TechSupport as the most influential attributes for churn prediction. Feature selection improved every performance metric of Naïve Bayes but only marginally reduced the performance of Random Forest, indicating that Random Forest is more robust to a reduction in the number of attributes. Overall, Random Forest is more suitable when the priority is detecting as many churn customers as possible (higher recall), while Naïve Bayes is preferable when reliable, conservative churn predictions (higher precision) are required. A limitation of this study is the use of a single train-test split without cross-validation and an imbalanced dataset that was not corrected prior to modeling, which may affect the generalizability of these results. Future work is recommended to apply class-imbalance handling techniques such as SMOTE and to compare additional algorithms, such as XGBoost, Support Vector Machine, or Gradient Boosting, to obtain a more optimal churn prediction model.

Acknowledgement

The author wishes to express gratitude to the supervisor and Bina Sarana Informatika University for their guidance and support throughout this research.

References

- [1] M. A. Iskandar, U. Latifa, F. Teknik, U. S. Karawang, and C. Churn, "WEBSITE PREDIKSI CUSTOMER CHURN UNTUK MEMPERTAHANKAN," vol. 7, no. 2, pp. 1308–1316, 2023, doi: <https://doi.org/10.36040/jati.v7i2.6639>.
- [2] P. N. Sabrina *et al.*, "PREDIKSI CUSTOMER CHURN PADA PERUSAHAAN TELEKOMUNIKASI MENGGUNAKAN ALGORITMA C4.5 BERBASIS PARTICLE SWARM OPTIMIZATION," vol. 7, no. 5, pp. 3369–3375, 2023, doi: <https://doi.org/10.36040/jati.v7i5.7476>.
- [3] M. N. Faladiba and R. Haryastuti, "CHURN ANALISIS PADA DATA PELANGGAN TELEKOMUNIKASI MENGGUNAKAN ENSEMBLE LEARNING Muthia Nadhira Faladiba 1*, Rizqi Haryastuti 2 1," vol. 5, no. 1, pp. 45–54, 2023.
- [4] N. Salma and A. Aprianingsih, "Customer Churn Analysis : Analyzing Customer Churn Determinants on an ISP Company in Indonesia," vol. 19, no. 1, pp. 29–40, 2021, doi: [10.17933/bpostel.2021.190103](https://doi.org/10.17933/bpostel.2021.190103).
- [5] A. T. Hartanto, "Prediksi Pelanggan Churn pada Perusahaan Telekomunikasi Menggunakan Logistic Regression di RapidMiner," vol. 4, no. 2, pp. 49–54, 2025.
- [6] N. Agian, S. Dinata, G. Abdurrahman, and N. Q. Fitriyah, "Perbandingan Optimasi Algoritma Random Forest Menggunakan Teknik Boosting Terhadap Kasus Klasifikasi Churn Pelanggan Di Industri Telekomunikasi," vol. 5, no. 1, pp. 28–37, 2023.
- [7] H. Aulia and N. Armelia, "TIJAB (The International Journal of Applied Business) Factors Influence Customer Churn on Internet Service Providers in Indonesia," vol. 6, no. 2, pp. 134–144, 2022.
- [8] D. H. Tisantri, R. C. Wihandika, and S. Adinugroho, "Prediksi Keputusan Pelanggan Menggunakan Extreme Learning Machine Pada Data Telco

- Customer Churn,” vol. 3, no. 11, pp. 10516–10523, 2021.
- [9] M. A. Mokoagow and A. S. Purnomo, “Penerapan Metode Naïve Bayes Pada Sistem Pakar Untuk Mendiagnosis,” vol. 4, no. 2, 2024, doi: <https://doi.org/10.58794/jekin.v4i2.803>.
- [10] O. Manullang, C. Prianto, and N. H. Harani, “Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest,” no. 54, 2023, doi: <https://doi.org/10.33884/jif.v11i02.7987>.
- [11] M. N. Fahmi, S. Tinggi, A. Islam, and N. Islam, “Implementasi Mechine Learning menggunakan Python Library : Scikit-Learn (Supervised dan Unsupervised Learning),” vol. 2, pp. 87–96, 2023, doi: <https://doi.org/10.52620/sainsdata.v1i2.31>.
- [12] A. Srirahayu and L. S. Pribadie, “Review Paper Data Mining Klasifikasi Data Mining,” vol. 14, no. April, 2023, doi: <https://doi.org/10.36982/jiig.v14i1.2981>.
- [13] L. N. Wakhidah, A. K. Zyen, and B. B. Wahono, “Evaluation of Telecommunication Customer Churn Classification with SMOTE Using Random Forest and XGBoost Algorithms,” vol. 9, no. 1, pp. 89–95, 2025.
- [14] F. Firdaus and L. P. Pane, “Perbandingan Decision Tree dan Naive Bayes untuk Prediksi Customer Churn,” vol. 4, no. 1, pp. 195–200, 2026, doi: [10.26905/jisad.v4i1.16837](https://doi.org/10.26905/jisad.v4i1.16837).
- [15] A. Nurtriana, D. D. Rachmawati, M. Artiyasa, D. Syahrudin, and Z. Sidiq, “Churn prediction analysis of telecom customers using svm , random forest and logistic regression models using orange data mining tools,” vol. 02012, 2024.
- [16] A. Yulianto, “Prediksi Customer Churn Pada Bisnis Retail Menggunakan Algoritma Naïve Bayes,” vol. 6, pp. 41–47, 2021.
- [17] R. Alfarez and V. Purwayoga, “PENERAPAN NAÏVE BAYES UNTUK PREDIKSI CUSTOMER CHURN (STUDI KASUS : PT HUTCHISON 3 INDONESIA),” vol. 05, no. 02, pp. 301–307, 2024.