JIKO (Jurnal Informatika dan Komputer)
Vol. 7, No. 3, December 2024, pp. 209-216
DOI: 10.33387/jiko.v7i3.8711

Accredited KEMENDIKBUD RISTEK, No.105/E/KPT/2022
p-ISSN: 2614-8897
e-ISSN: 2656-1948

CUSTOMER CHURN PREDICTION USING THE RANDOM FOREST ALGORITHM

Yosep Setiawan[1*], Asep Id Hadiana[2], Fajri Rakhmat Umbara[3]

[1,2,3]Faculty of Science and Informatics, Universitas Jenderal Achmad Yani Jawa Barat, Cimahi, Indonesia
Email: [*1]yosepsetiawan20@if.unjani.ac.id, [2]asep.hadiana@lecture.unjani.ac.id,
[3]fajri.rakhmat@lecture.unjani.ac.id

(Received: 24 August 2024, Revised: 02 September 2024, Accepted: 30 October 2024)

Abstract

Customer churn prediction plays a vital role in modern business, accurately influencing strategic and operational decisions that influence customer loyalty to a service. Customer churn focuses on customer retention being more profitable than attracting new customers because long-term customers provide lower profits and costs while losing customers increases the costs and need to attract new customers. However, customer churn still occurs frequently and cannot be predicted. If customer churn is left unchecked, it will endanger the company or banking industry because it can cause loss of income, damage reputation, and decrease market share. Random Forest, a data mining technique, was used in this research because of its ability to predict and handle many variables. This research aims to predict customer churn using the Random Forest method with datasets from Europe, especially France, Spain, and Germany, hoping to benefit the banking industry by identifying customers at high risk of abandoning services. This research is expected to benefit business people from customer churn predictions. Especially in the banking industry, it can help identify customers at high risk of abandoning service. Thus, companies can take appropriate steps to retain these customers, increase customer retention, strengthen customer loyalty and optimize their business performance. The results of this research are an accurate system for predicting customer churn in the future. The research obtained accuracy results of 87% in predicting customer churn using accuracy testing in the form of a confusion matrix.

Keywords: *Random Forest, Customer Churn, Prediction, Data Mining, Banking Industry*

**Corresponding Author: Yosep Setiawan*

1. INTRODUCTION

In modern business, customer churn data is a vital component that drives strategic and operational decisions. The accuracy of customer churn predictions influences customer interest in remaining loyal to a service in a business. Customer churn prediction is essential in customer relationship management because retaining existing customers is more profitable than attracting new ones. Successful companies tend to have long-term relationships with customers, which results in financial benefits and reduced service costs. Losing customers increases costs and reduces profits, so customer churn prediction is critical in customer retention strategies [1][2].

The discipline of machine learning focuses on how machines can gain knowledge through experience, with learning ability as a key indicator of intelligence. For many scientists, the terms "machine learning" and "artificial intelligence" are often used interchangeably, considering that both relate to the development of computer systems that are able to adapt and learn from experience. Machine learning enables the identification of hidden patterns in data and the adjustment of algorithms to increase the robustness of those patterns. A set of instructions known as a machine learning algorithm allows a computer to automatically learn from historical data and continuously improve its performance without requiring complex programming [3][4].

Data Mining is a machine learning discipline that focuses on the systematic process of finding significant patterns in large data sets by utilizing Machine Learning, Statistics, and Artificial Intelligence. This process includes categorization, integration, transformation, discretization, and assessment of data patterns to uncover hidden relationships that can predict future events. Applied in

areas such as marketing, sales, and scientific discovery, Data Mining enables the exploration of deep insights from big data, supporting better decision making. These advances in technology reflect the need for more complex information than simple transaction data and historical facts [5][6].

In data mining, there are several techniques that are often used to extract insights from data. The techniques that are often used are prediction, regression, clustering, association, anomaly detection, and dimensionality reduction. In customer churn data research, predictions estimate future events by utilizing relevant information from the past, aiming to provide probability estimates regarding the possibility of events that will occur [7].

To predict future events by utilizing relevant information from the past, you can use the Decision Tree method as an important tool in decision making which visualizes options and decisions through a tree structure. By analyzing the relationships between variable attributes and target variables, these models are effective in classification and prediction, depicting data patterns in a clear and easy to understand manner. The main focus is on discrimination and classification, enabling the prediction of categorical values of target variables based on existing attributes [8][9].

The Random Forest algorithm takes a decision tree as an effective ensemble method for classification and regression, which consists of a series of decision trees. Each tree is built independently using a random subset of the features and training data, then the prediction results from all the trees are combined to produce a final prediction. This method significantly reduces the risk of overfitting and improves prediction accuracy on complex data. By randomly generating child nodes, Random Forest builds a decision tree that includes root nodes, internal nodes, and leaf nodes, randomly selecting attributes and data to improve overall accuracy results [10][11].

To find out which features have an important role in customer churn data when using Random Forest, Information Gain (IG) is used. Information Gain (IG) measures the effectiveness of a feature in classifying data by assessing how much uncertainty or entropy about the target variable reduces when the feature value is known. In the decision tree algorithm, IG indicates the extent to which the feature provides additional information in predicting the data class [12].

SMOTE-ENN integrates oversampling techniques with Edited Nearest Neighbors (ENN) rules to overcome class imbalance in the dataset, with the aim of eliminating synthetic samples that are irrelevant and likely to be misclassified. This technique improves overall model performance, although it runs the risk of losing valuable information if informative minority samples are close to the majority. Research by Y. Chachoui shows that SMOTE-ENN successfully improves model performance, especially for SVM and Random Forest, with promising results in terms of accuracy, precision, recall, F1-score, and AUC [13].

Because Random Forest handles categorical data, Binning is used to group continuous data into smaller intervals or categories, or "bins," to simplify the data, reduce noise, and simplify analysis by converting continuous data to discrete. By grouping adjacent values, binning reduces data complexity and increases model efficiency, making it easier for algorithms to find patterns [14].

To evaluate the accuracy results of predicting customer churn data using the Random Forest algorithm, a Confusion Matrix is used which functions to assess the performance of the classification model by comparing the model prediction results to the actual values of the observed data. This table includes four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP indicates the amount of data correctly classified as positive, TN indicates the amount of data correctly classified as negative, FP represents the amount of data incorrectly classified as positive, and FN represents the amount of data incorrectly classified as negative. Confusion matrix is used to calculate various classification evaluation metrics, such as accuracy, precision, recall, and F1-score [15][16].

Previous research by Pahul Preet Singh emphasized the importance of understanding and handling customer churn in the banking sector. A high churn rate indicates underlying problems such as a less than satisfactory customer experience. This study uses machine learning to predict churn with algorithms such as XGBoost and Random Forest. The use of data visualization applications such as the Data Visualization RShiny app is also discussed. With this technology, banks can understand customer behavior and take proactive steps to reduce churn. Evaluations show that both XGBoost and Random Forest perform well, with XGBoost tending to be slightly superior in some evaluation metrics [17].

In previous research, it was explained that using the random forest algorithm with smote and not using smote on customer churn data. Those who didn't use a smote produced an accuracy of 0.84 and using a smote produced an accuracy of 0.78. Random Forest can work well on imbalanced datasets, but its performance can be better if the dataset is balanced. In previous research, the accuracy obtained without using a smote on an unbalanced dataset was greater than using a smote on a balanced dataset. In previous research, smote carried out an oversampling technique from the minority class to create synthetic samples to increase the minority class in order to overcome class imbalance, but after smote carried out there was still noise and outliers. So in this research, research was carried out again on customer churn data using the random forest algorithm, and changed smote to smote-enn to clean the data from noise and outliers, then to remove examples that were incorrectly classified based on their k nearest neighbors so that the dataset

became cleaner and strengthened the pattern. the correct one so that it is more easily recognized by the prediction model.

This research aims to predict customer churn using the Random Forest method. This research will use customer churn datasets in Europe (France, Spain, Germany) to evaluate how algorithms adapt to data anomalies and what impact they have on customer churn predictions.

This research is expected to provide benefits from customer churn predictions for business people. especially in the banking industry, it can help identify customers who are at high risk of abandoning service. Thus, companies can take appropriate steps to retain these customers, increase customer retention, and ultimately strengthen customer loyalty and optimize their business performance.

2. RESEARCH METHOD

This research will be carried out in several stages, namely data acquisition, pre-processing, training data, testing data, predictions, method evaluation, and reporting and scientific publications, as explained in Figure 1.

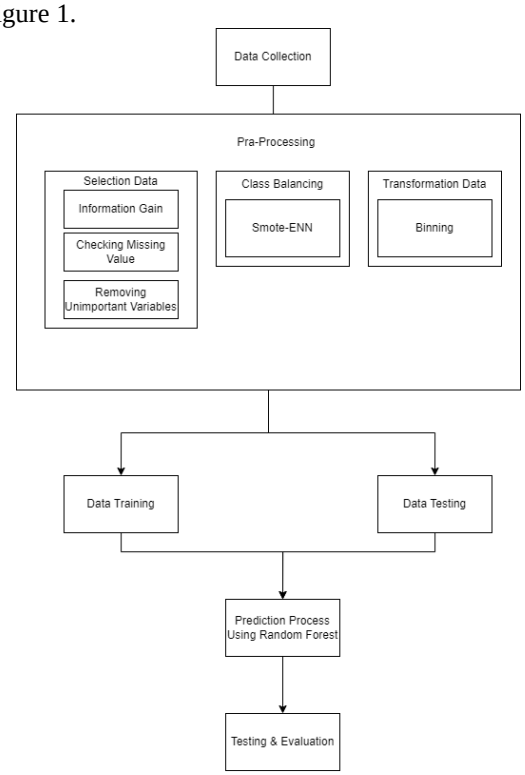

Figure 1. Research Method

2.1 Data Collection

The first stage is an important stage in obtaining information and data related to the problem to be researched. The data used for this research is a dataset obtained from the Kaggle website (https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn/data) with the title "Customer-Churn-Records" in CSV form. This dataset contains 18 attributes and 10,000 data, of these 18 attributes there are 10 attributes that have numerical values and 8 attributes that are categorical.

Literature study is one of the steps that will be taken to search for and review information that is relevant to the research report. By looking for additional information about the research subject which may be found in various sources, including books, guidelines, journals, and literature collected by researchers to fill in the data needed in the research, literature studies seek to be used as research material.

2.2 Preprocessing

The second stage is the data processing stage or pre-processing stage. Pre-processing is an important step in building a prediction model because real-time database data is often insufficient, non-uniform, and inconsistent, leading to incorrect prediction results. In this research, the pre-processing stage is divided into three parts, namely data selection, class balancing and data transformation.

1. Selection Data

In this research, in the pre-processing stage, data selection carried out information gain, checked for missing values, deleted variables that were not important. Information gain is a measure used in information theory and machine learning to determine how well a feature (or attribute) separates classes in a dataset. Specifically, IG measures the reduction in uncertainty (entropy) about a target class obtained by knowing the value of a particular feature. The higher the Information Gain, the better the feature is at providing information about the target class. Missing values in this research look for each data to see if there is empty data in each attribute in the dataset. In this study, unimportant variables were deleted because these variables had nothing to do with the prediction process using random forest.

2. Class Balancing (Smote-Enn)

In this research, in the pre-processing stage, class balancing uses smote-enn. SMOTE-ENN is a combination of two techniques: SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors). This technique is used to handle class imbalance in a dataset by oversampling and undersampling simultaneously. The SMOTE-ENN process begins by applying SMOTE to create synthetic samples from minority classes, as described previously. After that, the ENN technique is applied to remove ambiguous samples or those that are considered noise from the SMOTE dataset. The ENN technique checks each sample in the SMOTE dataset and deletes it if the majority of its k-nearest neighbors are from the majority class. In this way, this technique helps clean the dataset from synthetic samples that are useless or that can introduce noise. SMOTE-ENN aims to reduce the effects of generating synthetic samples that may be irrelevant or noisy, in order to

maintain diversity in minority classes. This can help improve model quality and prevent overfitting.

3. Transformation Data (Binning)

In this research, in the pre-processing stage, data transformation uses Binning. Binning is a technique in data mining that is used to group numerical data into intervals or "bins." This technique helps simplify data by reducing the number of categories or unique values that must be analyzed, as well as reducing the impact of noise or small variations in the data. Binning helps reduce the influence of noise or outliers in data by grouping data into larger groups. Binning converts continuous data into categorical data, which can be easier to analyze in some cases, especially when working with certain algorithms that require categorical data. In some cases, machine learning models can perform better with data that has been grouped into bins, as this can reduce model complexity and overfitting.

2.3 Prediction Process Using Random Forest

The third stage is the stage of the data sharing process and prediction process, where in data sharing the data will be divided from data that has been pre-processed, where the data will be divided into 2, namely testing data by 20% and training data by 80%. Stages of the prediction process, where predictions will be made using the Random Forest algorithm from test data and training data. This prediction application with the random forest model uses the Python programming language and library as well as the Flask framework.

2.4 Testing and Evaluation

The fourth stage is the testing and evaluation stage, which is the result of evaluating the prediction results using the random forest method to measure the level of accuracy using a confusion matrix.

3. RESULT AND DISCUSSION

3.1 Data Collection

In this research, the data that will be used is customer churn data located in Europe (Spain, Germany, France) with a total of 10,000 records.

1. Data Attribute Description

Before starting the data mining process stage, steps are needed to explain each attribute contained in the dataset that has been obtained. The goal is to identify relevant attributes in the context of customer churn data. Detailed information about each attribute can be seen in Table 1.

Table 1. Attributes Description

No	Data Attributes	Description
1	CreditScore	Customer credibility
2	Geography	Customer location name
3	Gender	Customer gender
4	Age	Customer age
5	Tenure	Time period, refers to the number of years a customer has been a bank customer.
6	Balance	Customer account balance
7	NumOfProducts	Number of products purchased by customers through the bank
8	HasCrCard	Whether the customer has a credit card or not
9	IsActiveMember	Is the customer a customer who actively uses the service?
10	EstimatedSalary	Estimated customer salary
11	Exited	Whether customers remain loyal or not to the service
12	Satisfaction Score	Score given by customers for complaint resolution
13	Card Type	The type of card held by the customer
14	Point Earned	Points earned by customers for using credit cards

The description of the data attributes in the table above explains the information contained in the dataset attributes, which has 14 data attributes that will be used in the research.

3.2 Preprocessing

The pre-processing stage is the stage for changing data into data that is ready to be tested. In the pre-processing stage, the data is processed through the data selection stage, class balancing using smote-enn, and data transformation using binning.

1. Selection Data (Information Gain)

Information Gain (IG) is a metric used to measure how much information is obtained from an attribute in separating data based on targets or labels. This is an important concept in decision making, especially in algorithms such as decision trees. The higher the IG value, the better the attribute is at separating data into different groups.

In the data selection process, Information Gain is used to select the most informative attributes to use in the model. This process involves calculating the entropy of the data as a whole, Calculating the entropy for each attribute, Subtracting the entropy of each attribute from the overall entropy to get the Information Gain, Selecting the attribute with the highest Information Gain to use in the model. Detailed information about the information gain results obtained from this research can be seen in Table 2.

Table 2. Information Gain

Feature	Information Gain
NumOfProducts	0.078426
Age	0.073962
Geography	0.025767
IsActiveMember	0.015017
Balance	0.007451
Point Earned	0.004271
Gender	0.003634
Estimated Salary	0.002710
Card type	0.000706
HasCrCard	0.000420
CreditScore	0.000000
Tenure	0.000000
Satisfaction Score	0.000000

2. Selection Data (Checking Missing Value)

Missing values are data that is missing or not available in the dataset. This missing data can be caused by various reasons, such as recording errors, technical errors, or the irrelevance of questions to respondents in the survey.

To check for missing values in a dataset, several general steps can be taken: identifying the total number of missing values in the dataset, determining whether the missing values are random or there is a certain pattern, deleting rows or columns that have missing values, and filling in the missing values in a certain way such as, mean, median, mode, or using a predictive model. The results of checking missing values in this study were that there were no missing values.

3. Selection Data (Removing Unimportant Variable)

Removing unimportant variables is an important step in data pre-processing. The goal is to increase model efficiency and avoid overfitting. The steps usually taken are: deleting variables that do not have a significant contribution, identifying and deleting variables that are highly correlated with other variables, because they provide redundant information. Detailed information about the results of deleting unimportant variables obtained from this research can be seen in Table 3.

Table 3. Variables Used

No	Variabel
1	CreditScore
2	Geography
3	Gender
4	Age
5	Tenure
6	Balance
7	NumOfProducts
8	HasCrCard
9	IsActiveMember
10	EstimatedSalary
11	Exited
12	Satisfaction Score
13	Card Type
14	Point Earned

4. Class Balancing

At this stage, class balancing is carried out to balance exited data. This class balancing stage uses smote-enn by balancing the exited data from the majority and minority classes so that the data is not too far apart.

The use of SMOTE-ENN aims to improve the performance of the classification model by handling class imbalance and also reducing noise in the data. This is especially useful in domains where class imbalance is a significant problem, such as in fraud detection, medical diagnosis, and other pattern recognition especially customer churn. The results of Smote-Enn can be seen in Table 4.

Table 4. Class Balancing Dataset

Credit Score	Geography	Gender	Card Type	Point Earned
822	0	1	3	206
528	0	1	1	264
476	0	0	3	119
587	2	1	2	732
726	0	0	1	477

From the data above are the results of class balancing for classes using smote enn. This method aims to overcome class imbalance between the majority and minority classes, as well as reduce noise in the data.

Next in this research, a comparative analysis will be presented between using smote enn and not using smote enn. The performance results between using the Enn smote and not using the Enn smote can be seen in Figure 2 and Figure 3.

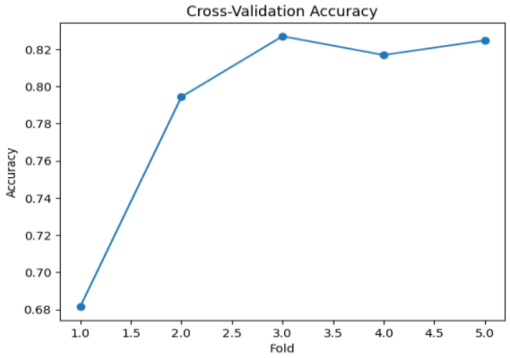

Figure 2. Performance Using Smote Enn

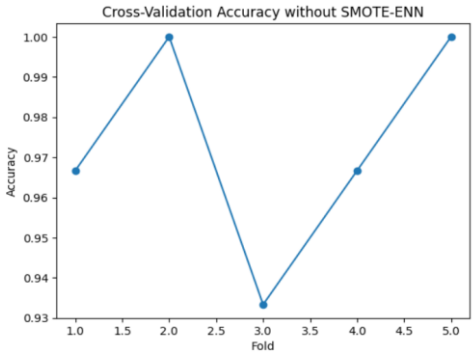

Figure 3. Performance Without Using Smote Enn

In Figure 3 it can be seen that there are significant fluctuations in accuracy values between folds, which shows that the model has difficulty generalizing in certain subsets of data during cross-validation. this can be caused by unbalanced dataset characteristics, outliers, or variations in distribution between folds. in figure 3 the model performance without using smote enn becomes less consistent, as can be seen from the large difference between accuracy on the second and third folds. This indicates that imbalanced data can cause the model to perform poorly on certain subsets of the data.

5. Transformation Data

At this stage, data transformation is carried out to change data variables into categorical data forms. This

transformation stage uses binning to transform data variables into categories using the Equal-width binning technique. Equal-width binning is a type of binning that divides data evenly into bins of a certain width.

In the transformation process, data binning is used to group continuous data into smaller intervals or "bins." The goal is to simplify the data and make it easier to analyze, as well as to reduce the impact of noise or small fluctuations in the data. Detailed information about the results of data binning result from this research can be seen in Table 5.

Table 5. Data Binning Results

Variabel	Number of Bins	Label	Bins Range
CreditScore	10	Bin1, Bin2, Bin3, Bin4, Bin5, Bin6, Bin7, Bin8, Bin9, Bin10	Bin1 (0% - 10%), Bin2 (11% - 20%), Bin3 (21% - 30%), Bin4 (31% - 40%), Bin5 (41% - 50%), Bin6 (51% - 60%), Bin7 (61% - 70%), Bin8 (71% - 80%), Bin9 (81% - 90%), Bin10 (91% - 100%)
Geography	3	France, Germany, Spain	Bin1 (France): interval from -1 to 0.5, Bin2 (Germany): interval from 0.6 to 1.5, Bin3 (Spain): interval from 1.6 to 3
Gender	2	Female, Male	Bin1 (Female): interval from -1 to 0.5, Bin2 (Male): interval from 0.6 to 1.5
Age	10	Bin1, Bin2, Bin3, Bin4, Bin5, Bin6, Bin7, Bin8, Bin9, Bin10	Bin1 (0% - 10%), Bin2 (11% - 20%), Bin3 (21% - 30%), Bin4 (31% - 40%), Bin5 (41% - 50%), Bin6 (51% - 60%), Bin7 (61% - 70%), Bin8 (71% - 80%), Bin9 (81% - 90%), Bin10 (91% - 100%)
Tenure	5	Bin1, Bin2, Bin3, Bin4, Bin5	Bin1 (0% - 20%), Bin2 (21% - 40%), Bin3 (41% - 60%), Bin4 (61% - 80%), Bin5 (81% - 100%)
Balance	10	Bin1, Bin2, Bin3, Bin4, Bin5, Bin6, Bin7, Bin8, Bin9, Bin10	Bin1 (0% - 10%), Bin2 (11% - 20%), Bin3 (21% - 30%), Bin4 (31% - 40%), Bin5 (41% - 50%), Bin6 (51% - 60%), Bin7 (61% - 70%), Bin8 (71% - 80%), Bin9 (81% - 90%), Bin10 (91% - 100%)
NumOf Products	4	One, Two, Three, Four	Bin1 (One): interval from 0.5 to 1.5, Bin2 (Two): interval from 1.6 to 2.5, Bin3 (Three): interval from 2.6 to 3.5, Bin4 (Four): interval from 3.6 to 4.5
HasCrCard	2	No, Yes	Bin1 (No): interval from -1 to 0.5, Bin2 (Yes): interval from 0.6 to 1.5
IsActive Member	2	No, Yes	Bin1 (No): interval from -1 to 0.5, Bin2 (Yes): interval from 0.6 to 1.5
Estimated Salary	10	Bin1, Bin2, Bin3, Bin4, Bin5, Bin6, Bin7, Bin8, Bin9, Bin10	Bin1 (0% - 10%), Bin2 (11% - 20%), Bin3 (21% - 30%), Bin4 (31% - 40%), Bin5 (41% - 50%), Bin6 (51% - 60%), Bin7 (61% - 70%), Bin8 (71% - 80%), Bin9 (81% - 90%), Bin10 (91% - 100%)
Exited	2	No, Yes	Bin1 (No): interval from -1 to 0.5, Bin2 (Yes): interval from 0.6 to 1.5
Satisfaction Score	5	One, Two, Three, Four, Five	Bin1 (One): interval from 0.5 to 1.5, Bin2 (Two): interval from 1.6 to 2.5, Bin3 (Three): interval from 2.6 to 3.5, Bin4 (Four): interval from 3.6 to 4.5,

Card Type	4	Diamond, Gold, Silver, Platinum	Bin5 (Five): interval from 4.6 to 5.5 Bin1 (Diamond): interval from -1 to 0.5, Bin2 (Gold): interval from 0.6 to 1.5, Bin3 (Silver): interval from 1.6 to 2.5, Bin4 (Platinum): interval from 2.6 to 3.5
Point Earned	10	Bin1, Bin2, Bin3, Bin4, Bin5, Bin6, Bin7, Bin8, Bin9, Bin10	Bin1 (0% - 10%), Bin2 (11% - 20%), Bin3 (21% - 30%), Bin4 (31% - 40%), Bin5 (41% - 50%), Bin6 (51% - 60%), Bin7 (61% - 70%), Bin8 (71% - 80%), Bin9 (81% - 90%), Bin10 (91% - 100%)

3.3 Prediction Process Using Random Forest

In this process, data that has gone through the previous process will be separated, and the data will be converted into two subsets, namely training data and test data. The percentage for training data is 80%, while for test data it is 20%.

3.4 Testing and Evaluation

Accuracy testing is a stage for testing the accuracy results obtained on the model used. This test uses the confusion matrix method. Testing of prediction results carried out by the Random Forest model was tested using the confusion matrix method. The following are the confusion matrix values which can be seen in Figure 4.

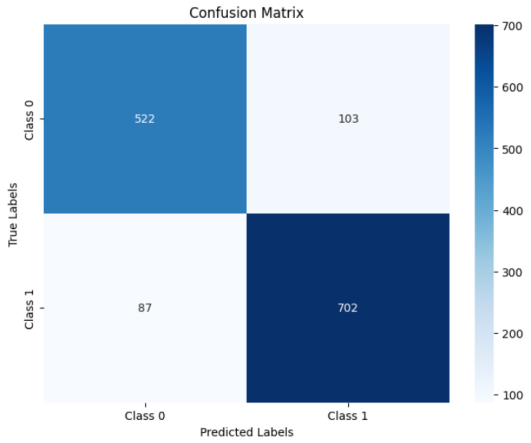

Figure 4. Confusion Matrix Graph

The following is the formula and manual calculation of the confusion matrix which can be seen in Table 6.

Table 6. Confusion Matrix Value

True Positive	False Positive	False Negative	True Negative
522	103	87	702

Formula:

$$Accuracy = \left(\frac{(TP+TN)}{TP+FP+FN+TN}\right) \times 100\% \quad (1)$$

Result:

$$Accuracy = \left(\frac{(522+702)}{522+103+87+702}\right) \times 100\% = 86\%$$

4. CONCLUSION

Based on this research, it was created using the random forest and smote-enn algorithms as a class balancing process. In this research, 80% of the data was used as training data, 20% as test data, and the accuracy of the results in predicting customer churn based on the attributes in the data were obtained at 87%. This research concludes that the random forest method is a prediction method that can be applied in the banking sector, especially in predicting customer churn, where customer loyalty is categorized as difficult to detect. Using Information Gain to select attributes, the author can ensure that only the attributes that contribute most to separating the data are selected. Information Gain increases the relevance and effectiveness of the model in making accurate decisions. Attributes with low or zero Information Gain can be ignored, reducing data dimensionality and model complexity. Information Gain speeds up model training and reduces the risk of overfitting, where the model becomes too specific to the training data and less able to generalize to new data. Focusing on attributes that provide high Information Gain makes the resulting model more efficient and easier to interpret. Key attributes selected based on Information Gain enable a better understanding of the critical factors influencing decisions in the data, thereby facilitating better data-based decision-making.

5. REFERENCE

[1] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *Eur. J. Oper. Res.*, vol. 269, no. 2, pp. 760–772, 2018, doi: 10.1016/j.ejor.2018.02.009.

[2] V. Geetha, A. Punitha, A. Nandhini, T. Nandhini, S. Shakila, and R. Sushmitha, "Customer Churn Prediction in Telecommunication Industry Using Random Forest Classifier," *2020 Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2020*, 2020, doi: 10.1109/ICSCAN49426.2020.9262288.

[3] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan,

V. K. Menon, and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 1643–1647, 2017, doi: 10.1109/ICACCI.2017.8126078.

[4] A. Yosipof, R. C. Guedes, and A. T. García-Sosa, "Data mining and machine learning models for predicting drug likeness and their disease or organ category," *Front. Chem.*, vol. 6, no. MAY, pp. 1–11, 2018, doi: 10.3389/fchem.2018.00162.

[5] D. Papakyriakou and I. S. Barbounakis, "Data Mining Methods: A Review," *Int. J. Comput. Appl.*, vol. 183, no. 48, pp. 5–19, 2022, doi: 10.5120/ijca2022921884.

[6] L. P. Muri, B. Pramono, and J. Y. Sari, "Prediksi tingkat penyakit demam berdarah di kota kendari menggunakan metode," *semanTIK*, vol. 4, no. 1, pp. 103–112, 2018.

[7] W. A. P. Wanto Anjar, "Analisis Prediksi Indeks Harga Konsumen Berdasarkan Kelompok Kesehatan Dengan Menggunakan MWanto, A. (2019). Analisis Prediksi Indeks Harga Konsumen Berdasarkan Kelompok Kesehatan Dengan Menggunakan Metode Backpropagation. Jurnal & Penelitian Teknik Infor," *J. Penelit. Tek. Inform.*, vol. 2, no. 2, pp. 37–44, 2017, [Online]. Available: https://zenodo.org/record/1009223#.Wd7norlTbhQ

[8] A. I. Sang, E. Sutoyo, and I. Darmawan, "Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma Decision Tree Dan Support Vector Machine Data Minning Analysis for Classification of Air Quality Data Dki Jakarta Using Decision Tree Algorthm and Support Vector," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 8954–8963, 2021.

[9] R. F. Ramadhani, S. S. Prasetiyowati, and Y. Sibaroni, "Performance Analysis of Air Pollution Classification Prediction Map with Decision Tree and ANN," *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 536–543, 2022, doi: 10.47065/josyc.v3i4.2117.

[10] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[11] K. Schouten, F. Frasincar, and R. Dekker, "An information gain-driven feature study for aspect-based sentiment analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9612, pp. 48–59, 2016, doi: 10.1007/978-3-319-41754-7_5.

[12] M. Chen and Z. Liu, "Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 12, p. e32570, 2024, doi: 10.1016/j.heliyon.2024.e32570.

[13] Y. Chachoui, N. Azizi, R. Hotte, and T. Bensebaa, "Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning," *Comput. Educ. Artif. Intell.*, vol. 6, no. April, p. 100222, 2024, doi: 10.1016/j.caeai.2024.100222.

[14] T. Imbeault-Nepton, J. Maitre, K. Bouchard, and S. Gaboury, "Filtering Data Bins of UWB Radars for Activity Recognition with Random Forest," *Procedia Comput. Sci.*, vol. 201, no. C, pp. 48–55, 2022, doi: 10.1016/j.procs.2022.03.009.

[15] T. V. Ramana, "A Deep Learning Model for Detection Cancer in Breast," *J. Nurs. Res. Saf. Pract.*, no. 23, pp. 1–7, 2022, doi: 10.55529/jnrpsp.23.1.7.

[16] U. Shah, S. Garg, N. Sisodiya, N. Dube, and S. Sharma, "Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques," *PDGC 2018 - 2018 5th Int. Conf. Parallel, Distrib. Grid Comput.*, no. October 2019, pp. 776–782, 2018, doi: 10.1109/PDGC.2018.8745763.

[17] P. P. Singh, F. I. Anik, R. Senapati, A. Sinha, N. Sakib, and E. Hossain, "Investigating customer churn in banking: A machine learning approach and visualization app for data science and management," *Data Sci. Manag.*, vol. 7, no. 1, pp. 7–16, 2024, doi: 10.1016/j.dsm.2023.09.002.