TIERS Information Technology Journal
Vol.3, No.1, Juni 2022, pp. 35~47
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: 10.38043/tiers.v3i1.3616

Predictive Analysis of Customer Retention Using the Random Forest Algorithm

Yogasetya Suhanda[1], Lela Nurlaela[2], Ike Kurniati[3], Andy Dharmalau[4], Ita Rosita[5]
yogasetyas@swadharma.ac.id[1], lela@swadharma.ac.id[2], ikekurniati@swadharma.ac.id[3],
andy.d@swadharma.ac.id[4], itarosita17101008@gmail.com[5]
[1,4,5]Departement of Information System, Institut Teknologi Dan Bisnis Swdharma, Jakarta
[2,3]Departement of Informatic Engineering, Institut Teknologi Dan Bisnis Swdharma, Jakarta

ABSTRACT

Retaining customers is becoming a measurement focus in an industry with increasing competition. The concept of customer retention has become a research study in the sales industry, because it is difficult to retain customers and easily switch to other brands. Customer repurchase decisions in the business world of sales are very competitive. Customer satisfaction is directly proportional to the retention rate, if the customer is not satisfied then the automatic retention rate will be low. If the company is not able to meet customer expectations, it will have a serious impact on the company, namely moving customers to other services. Service factors, price, profit value, satisfaction and trust affect customer retention. One of the factors that influence consumers to become customer retention is service quality. A predictive customer retention plan is needed with data mining using the random forest algorithm. The random forest algorithm is a method that generates a number of trees from sample data, where the creation of one tree during training does not depend on the previous tree, the decision is based on the most voting. The voting results from several decision trees that are formed are the boundaries that are used as class determination in the classification process and the most votes are the winners and determine the classification class. This study aims to determine and analyze customer loyalty, customer trust and customer satisfaction. So that it can make it easier to monitor customers at the company. The results can be seen with the percentage of about 81.12% customer retention and about 18.87% customer churn. The result of feature evaluation shows that customer_activity has the highest influence on customer retention, followed by subtotal and qty.

Keywords: Churn, Customer, Customer Retention, Prediction Analysis, Random Forest, Service Quality.

Article Info

Accepted : 01-06-2022
Revised : 20-05-2022
Published Online : 25-06-2022

Correspondence Author:

Lela Nurlaela
Informatic Engineering,
Institut Teknologi Dan Bisnis Swdharma,
Jl. Malaka No 3 Roa Malaka, Kec. Tambora, West Jakarta, Jakarta 11230, Indonesia.
Email: lela@swadharma.ac.id

1. INTRODUCTION

Retaining customers is the focus of measurement in an industry that is increasingly competitive. Research on customer retention has become the study of several researchers with issues of increasingly competitive levels of competition, customers switching to other companies, dynamic customer behavior and

changes in marketing strategies from traditional to modern which are more sophisticated through today's innovations[1][2].

PT Wateru Natural Alkalindo is a company engaged in essential and non-essential fields that was founded in 2010 until now, still exists in retail and projects. Many projects have been handled, both government and private, namely pharmacies, hotels, supermarkets, restaurants, cafes, private projects, and so on with various products such as mineral water and tissue. The current situation and condition of business competition are increasingly competitive, so to maintain its business this company requires the concept of customer retention. The concept of customer retention has become a research study in the sales industry, due to the difficulty of retaining customers and easily switching to other brands. Each company tries to offer a unique product so that customers do not use the same product and make repeat purchases. Customer repurchase decisions in the sales business world are very competitive[3]. Customer satisfaction affects retention rates, as evidenced by customer loyalty not to switch to other brands. Customer satisfaction is directly proportional to the level of retention, if the customer is not satisfied then the retention rate will automatically be low, this will affect the sustainability of a business[4][5]. If the company is not able to meet customer expectations, it will have a serious impact on the company, namely moving customers to other services. For this reason, companies need to create something that can bind customers so they don't switch to competing products. Efforts in this limitation are called switching barriers. Barriers to switching are all factors that make it difficult or cost the customer to switch to another service provider. This switching barrier is the factors that influence the customer's decision to continue using the product that has been previously selected and not to switch to another product.

The definition of customer retention is an organization's efforts to make customers always buy their products[2]. Customer retention activities start from the start of the company making contact with customers until the company can establish long-term relationships. Customer retention is concerned with turning individual customer transactions into long-term customers, by keeping customers with one company rather than switching companies. Customer retention is an effort to keep customers in the company for a long period of time[1]. The fulfillment of customer satisfaction and high switching costs make customers reluctant to switch to other companies and will make a constant frequency of purchases in the long term (customer retention). The right strategy or approach in building long-term customer relationships must be owned by the company to increase customer retention.

There are several strategies in creating long-term relationships with customers, namely building customer perceived value, customer satisfaction and customer loyalty. In addition, perceived service, perceived price, profit value, satisfaction and trust affect customer retention[6]. Customer perceived value will have an impact on customer retention. Customer retention is a form of consumer loyalty measured by consumer buying behavior which is indicated by the high frequency of product purchases by consumers[4]. Customer retention has a very big role in increasing sales, decreasing management costs, and referring new customers. Therefore, to make consumers become customer retention is part of the company's strategy. One of the factors that influence consumers to become customer retention is service quality[7][8]. For this reason, it is necessary to have a predictive customer retention plan using data mining using the random forest algorithm. Data mining is a stage of a process by applying certain methods or algorithms in finding new rules, patterns or information in a selected set of data [9].

Techniques and algorithms in data mining have many variations [10]. The accuracy in choosing the method, technique or algorithm must be adjusted to the objectives and processes that exist in data mining as a whole. The process in machine learning will refer to a method, where a computer can have automation capabilities in learning and doing a job.The machine learning process is carried out through a random forest algorithm, so that the work ordered to the computer can be done automatically. Random forest is a bagging method, which is a method that generates a number of trees from sample data where the creation of one tree during training does not depend on the previous tree then the decision is taken based on the most votes [11][10][12].This research is a follow-up study from the results of previous studies that have been carried out as material for comparison, reference and studies including:

Research conducted by Nur Hayati and Dede Suryana in 2012 with the title "The Effect of Trust and Commitment to Customer Loyalty"[4]. This research was conducted at Griya Buah Batu Bandung to determine the effect of trust and commitment on consumer loyalty at PT Nutrifood Indonesia. The research was conducted by means of surveys, interviews and questionnaires distributed to 100 samples using the incidental sampling method. The data analysis technique used is descriptive and correlation analysis. The result is that the two

hypotheses proposed have a significant influence, namely consumer trust and commitment to consumer loyalty, where the consumer trust variable has a more dominant influence [4].

Research conducted by Sadaf Nabavi, Shahram Jafari in 2013 entitled Providing a Customer Chum Prediction Model Using Random Forest and Boosted Trees Techniques(Case Study: Solico Food Industries Group)[2]. This study discusses data mining capabilities in designing and implementing a customer loss prediction model using the standard CRISP-DM methodology based on RFM (Recency, Frequency, Monetary) and Random Forest and Tree techniques, on the database of one of the Solico food industry groups. Using this model, the customer's tendency to return is identified and the planning of an effective marketing strategy. The results of the analysis of customer behavior show that the length of the relationship, relative frequency and average time of purchase are the best predictors [2].

Research conducted by Achintya Sharma; Deepak Gupta; Nikhil Nayak; Deepti Singh; Ankita Verma, with the title "Prediction of Customer Retention Rate Employing Machine Learning Techniques,"[13]. This study discusses in the telecommunications sector which has experienced a significant increase in the number of subscribers and technology content as well as competition among telecommunications companies.With the ever-increasing rate of customer churn, and it is more expensive to acquire new customers than to retain existing ones,the company makes customer retention one of the company's main focuses.This study is to compare the accuracy of traditional data mining techniques, namely Logistic Regression, Support Vector Machine (SVM), Decision Tree, XGBoost, Random Forest, Light Gradient Boosting, Gradient Descent Boosting and Cat Boost in predicting customer churn.To get an algorithm that can find the main causes of customer churn from one company to another and ways to increase customer retention.

Research conducted by Irfan Ullah, Basit Raza, Ahmad Kamran Malik,Muhammad Imran, Saif Ul Islam,Sung Won Kim with the title "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector"[14]. The amount of data generated in the telecommunications sector, every day by many client base. Business analysis and decision makers and ensure that getting new customers is more expensive than retaining an existing business, then customer relationship management (CRM) analysis and analysis is necessary to find out the causes of customer churn, as well as behavioral patterns of existing customers churn data. A churn prediction model is proposed that uses classification and clustering techniques to identify customer churn and get the factors behind customer churn in the telecommunications sector. The results show that the churn prediction proposed by the chum classification classification model is better using the RF algorithm and customer profiles using k-means clustering. From the problems above, the problem formulation of this research is how to predict customer retention using the random forest model? This study aims to determine and analyze customer loyalty [4], customer trust (customer trust) and customer satisfaction (customer satisfaction) [5]. So that it can make it easier to monitor customers at the company

2. RESEARCH METHOD

The random forest is a method that generates a number of trees from sample data, where the creation of one tree during training does not depend on the previous tree, then the decision is based on the most votes. The two basic concepts of the random forest algorithm are building an ensemble of trees via bagging with replacement and random selection of features from each tree. Ensemble-based classification has maximum performance if there is a low correlation between basic learners. An ensemble must build a weak basic learner, because a strong learner is likely to have a high correlation and usually also causes overfit. The random forest algorithm minimizes correlations and maintains the power of classification by randomizing the training process, namely by selecting a number of features at random from all the existing features during the training tree, then using the selected features to get the optimal branching tree. Random forest has two main parameters, namely m the number of trees to be used and k, which is the maximum number of features that are considered when branching. The more m values, the better the classification results, while the recommended k value is the square root or logarithm of the total number of features. Using a dataset T with a number of m trees as a basic leaner and k features chosen randomly from the total features for branching in each tree. The training process on each tree uses the T dataset which is the result of bootstrap from the dataset which is used as a parameter for the random forest. Bootstrap is the process of selecting a sample from the dataset to be used in the training tree process. The ensemble, bootstrap method is a sampling process with replacement, so that the sample taken for one training tree process can still be used for another training tree process. For details, see Figure 1 below.

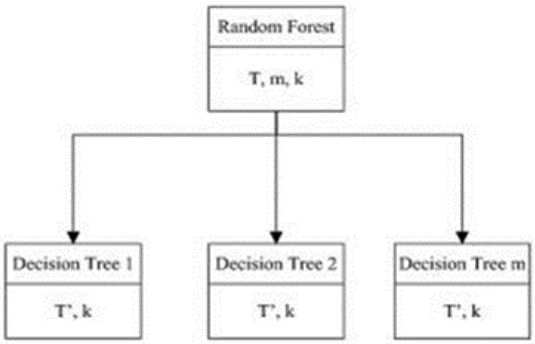

Figure 1. Random forest illustration

The more trees used, the better the accuracy value will be. To make predictions on a new sample with a random forest, it is done by inserting the sample into an existing decision tree to determine the class of the sample. This step is carried out repeatedly on the entire decision tree contained in the random frrest. The voting results from several decision trees that are formed are the boundaries that are used as class determination in the classification process and the most votes are the winners of determining the classification class.

This study is to predict customer retention with the random forest algorithm at PT Wateru Natural Alkalindo which was carried out from January 2020-March 2021 data. The primary data used is activity data and sales documents which will be processed to produce predictive data, to predict a variable in the future based on consideration of past data. The variable that will be predicted is customer retention, by referring to predictive data that can be used as support in making business decisions. The report output data is used as historical data on sales of Wateru Bamboo Tissue and Mineral Water for the period January 2020 - March 2021, for the full details, see table 1.

Table 1. Sample of historical sales data

tgl	no_id	jenis_pelanggan	nama_pelanggan	item	merk	jenis	qty	satuan	harga	subtotal	diskon	harga_nett	keaktifan_pelanggan	retensi_pelanggan
Aug-20	W0223	caffee	Abdi_Nagri_Coffee_Rostery	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	0	0
Sep-20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Oct-20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nov-20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dec-20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jan-21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Feb-21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mar-21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aug-20	W0234	caffee	Bakmi_Congkee	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	1	1
Sep-20	W0234	caffee	Bakmi_Congkee	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	1	1
Oct-20	W0234	caffee	Bakmi_Congkee	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	1	1
Nov-20	W0234	caffee	Bakmi_Congkee	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	1	1
Dec-20	W0234	caffee	Bakmi_Congkee	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	1	1
Jan-21	W0234	caffee	Bakmi_Congkee	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	1	1
Feb-21	W0234	caffee	Bakmi_Congkee	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	1	1
Mar-21	W0234	caffee	Bakmi_Congkee	Air	Eternalplus	500ml	10	Dus	140,000	1,400,000	0.00	1,400,000	1	1
Aug-20	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	67	Dus	140,000	9,380,000	0.00	9,380,000	1	1
Sep-20	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	6	Dus	140,000	840,000	0.00	840,000	1	1
Oct-20	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	21	Dus	140,000	2,940,000	0.00	2,940,000	1	1
Nov-20	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	68	Dus	140,000	9,520,000	0.00	9,520,000	1	1
Dec-20	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	80	Dus	140,000	11,200,000	0.00	11,200,000	1	1
Jan-21	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	60	Dus	140,000	8,400,000	0.00	8,400,000	1	1
Feb-21	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	80	Dus	140,000	11,200,000	0.00	11,200,000	1	1
Mar-21	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	100	Dus	140,000	14,000,000	0.00	14,000,000	1	1
Aug-20	W0276	caffee	Coffe_Tree	Air	Eternalplus	500ml	100	Dus	104,500	10,450,000	0.00	10,450,000	1	1
Aug-20	W0276	caffee	Coffe_Tree	Air	Eternalplus	250ml	10	Dus	135,000	1,350,000	0.00	1,350,000	1	1
Sep-20	W0276	caffee	Coffe_Tree	Air	Eternalplus	500ml	150	Dus	104,500	15,675,000	0.00	15,675,000	1	1
Sep-20	W0276	caffee	Coffe_Tree	Air	Eternalplus	250ml	15	Dus	135,000	2,025,000	0.00	2,025,000	1	1

Table 1 above is in the form of historical sales data that will be used as research, consisting of 1,950 rows and 15 columns containing: ***field* tanggal, no_id, jenis pelanggan, nama pelanggan, item, *merk*, jenis, *qty*, satuan, harga, subtotal, diskon, harga_nett, keaktifan_pelanggan, retensi_pelanggan**.

3. RESULTS AND DISCUSSION

Data collection is done by identifying data needs to achieve customer retention predicted goals/goals. This study uses historical sales data owned by PT. Wateru Natural Alkalindo for the period January 2020 – March 2021. The dataset contains 1,950 customers and 15 data points (fields) for each customer. Next, the total number of rows and columns is displayed in Figure 2.

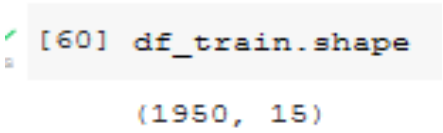

Figure 2. Display of the number of customers and fields

Figure 2 above shows the script df_train.shape which means that the results of the data frame form 1,950 rows and 15 columns.

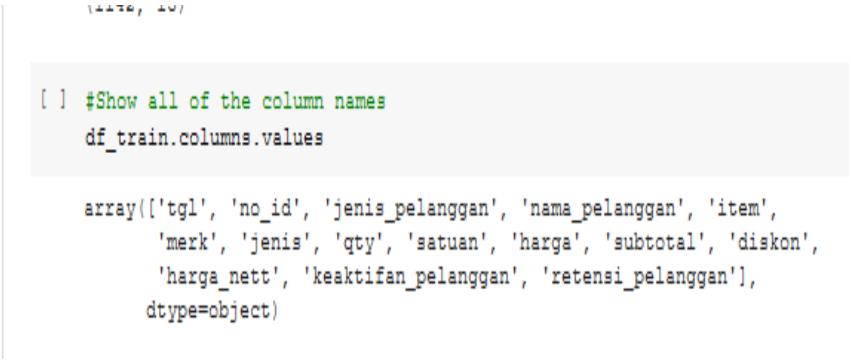

Figure 3. Display of all columns in the dataset

Figure 3 above shows the script df_train.colums.values which means from the training dataframe by using the column name or column type. So that a set of column names appears from the training data with the object data type. This division is based on the model that will be made on customer retention, entered into the case classification to supervise learning which is needed by algorithms in machine learning.

The data in this algorithm is generally divided into 2 parts, namely training data and testing data. The training data will be used to train the algorithm in finding the appropriate model, while the testing data will be used to test and determine the performance of the model obtained at the testing stage. The following figure 4 is a flowchart of the data collection mechanism:

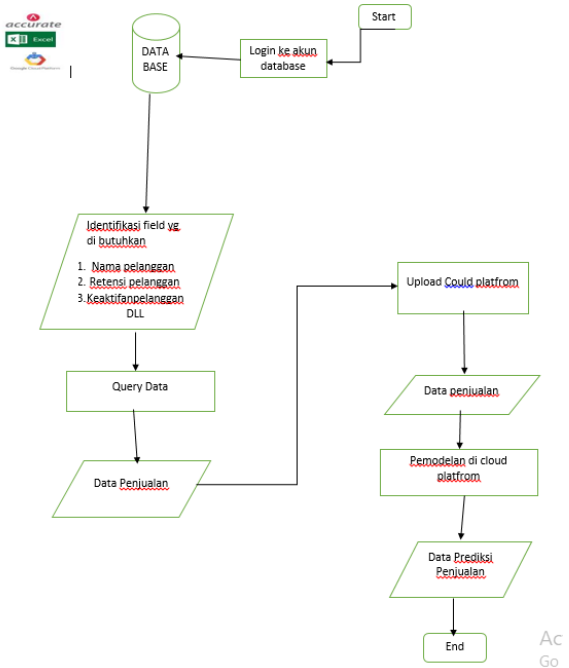

Figure 4. Flowchart display of data collection mechanism

The explanation of the process carried out first is database login to identify what fields will be needed, after selecting the data, we enter the query stage, changing data such as create, select, alter, drop, delete, insert, update. After querying the database, we get the sales data, which is uploaded to the google platform like we uploaded to google drive, after it is automatically uploaded to the save on google drive the sales data, we apply it using google collab for the modeling process on the cloud platform, after After the modeling is complete, we can get predictive data from the modeling.

The training data will use input data consisting of data withdrawals that have been carried out, it is known that the total sales data of PT Wateru Natural Alkalindo recorded during August 2020 - March 2021 is 1,950 data. There are 15 fields that are recorded as shown in Figure 5 below:

```
load_data_train=pd.read_excel(path_data_train)
load_data_train
```

	tgl	no_id	jenis_pelanggan	nama_pelanggan	item	merk	jenis	qty	satuan	harga	subtotal	diskon	harga_nett	keaktifa
0	2020-08-01	W0223	caffee	Abdi_Nagri_Coffee_Rostery	Air	Eternalplus	500ml	10	Dus	140000	1400000	0.0	1400000.0	
1	2020-09-27	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
2	2020-10-06	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
3	2020-11-11	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
4	2020-12-12	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1945	2020-11-28	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
1946	2020-12-28	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
1947	2021-01-28	W0246	caffee	cocoon	Air	Eternalplus	500ml	10	Dus	140000	1400000	0.0	1400000.0	
1948	2021-02-28	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
1949	2021-03-28	0	0	0	0	0	0	0	0	0	0	0.0	0.0	

1950 rows × 15 columns

Figure 5. Display of sales training data

Data Testing will use the testing phase and will use data from the data collection process carried out by

researchers. The testing data used is sales data in the January 2020-July 2020 period, as many as 1,809 rows and 14 columns, the following in Figure 6 is a display of sales testing data:

```
[6] load_data_test=pd.read_excel(path_data_test)
    load_data_test
```

	tgl	no_id	jenis_pelanggan	nama_pelanggan	item	merk	jenis	qty	satuan	harga	subtotal	diskon	harga_nett	keaktifan_j
0	2020-01-12	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	67	Dus	140000	9380000	0.0	9380000.0	
1	2020-02-22	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	6	Dus	140000	840000	0.0	840000.0	
2	2020-03-10	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	21	Dus	140000	2940000	0.0	2940000.0	
3	2020-04-11	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	68	Dus	140000	9520000	0.0	9520000.0	
4	2020-05-12	W0335	Resto	CV.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	80	Dus	140000	11200000	0.0	11200000.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1804	2020-03-28	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
1805	2020-04-28	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
1806	2020-05-28	0	0	0	0	0	0	0	0	0	0	0.0	0.0	
1807	2021-06-28	W0246	caffee	cocoon	Air	Eternalplus	500ml	10	Dus	140000	1400000	0.0	1400000.0	
1808	2021-07-28	0	0	0	0	0	0	0	0	0	0	0.0	0.0	

1809 rows × 14 columns

Figure 6. Display of sales testing data

From Figure 6 above, it can be seen that the testing data used has fields that are almost similar to the training data, including field date, no_id, type_customer, customer_name, item, brand, type, qty, unit, price, subtotal, discount, price_nett, active_customer but the difference is there is no customer activity column.

Data from the database is in the form of raw data to be used as modeling input, and a set of techniques applied to the database to remove noise, missing values, and inconsistent data so that the resulting model is of high quality. In order to improve the quality of the data to be analyzed, it is necessary to carry out data preprocessing steps and process the model with the scheme shown in Figure 7 below:

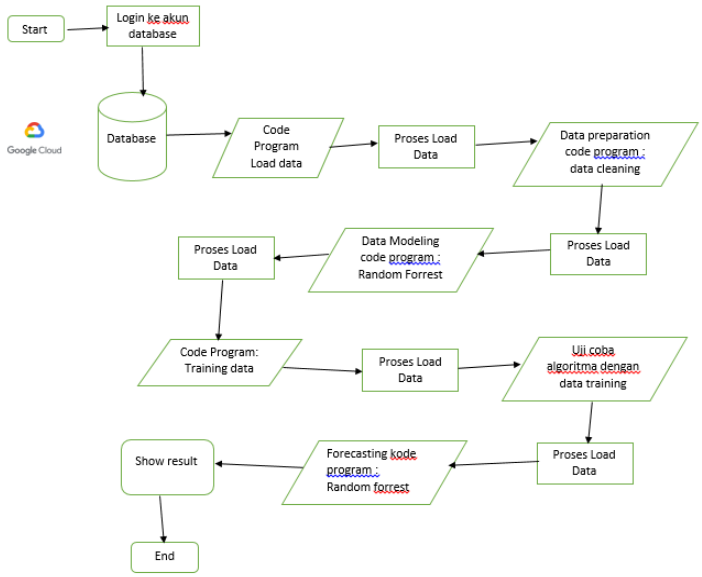

Figure 7. The flowchart of the predictive research process in the database

Data modeling is used to determine and analyze the data requirements needed, then enter the data modeling stage. The statistical model that we will use is the random forest algorithm. To use this algorithm,

we call the [fit] method and pass in the value of the feature set (x) and the corresponding set of labels (y). Then you can use the prediction method to make predictions on the test set (x_test), in Figure 8 below is the result of modeling using a random forest.

```
[31] from sklearn.ensemble import RandomForestClassifier
     #Create the model
     classifier = RandomForestClassifier(n_estimators=200, random_state=0)
```

```
[32] #Train the model
     model=classifier.fit(X_train,y_train)
     predictions = classifier.predict(X_test)
```

```
[33] print(predictions)
```

```
[1 1 1 1 0 1 1 1 0 0 1 1 0 0 1 1 0 0 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1
 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 0
 1 1 0 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 0 0 1 1 1 1 1 1 0
 1 1 0 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 0 1 1 1 1 0 1
 1 0 1 1 1 0 0 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 0 0 1 1 1 0 0 1 0 1 0 1 1 1 1
 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1
 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 0 1 0
 1 1 1 1 0 1 1 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1
 0 0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0]
```

Figure 8. Display of dataset prediction results

From Figure 8 the prediction results above show more number 1, meaning that there are more customer retention than customers who do not retain or churn.

Next, we enter the model evaluation stage, to evaluate the random forest model, we will score the accuracy and see test statistics such as precision, recall, and f-1 score, in Figure 9 the following is the display of the model evaluation results:

```
[34] from sklearn.metrics import accuracy_score
     print(classification_report(y_test,predictions ))
     print(accuracy_score(y_test, predictions ))
```

```
              precision    recall  f1-score   support

           0       0.92      0.88      0.90        78
           1       0.97      0.98      0.98       312

    accuracy                           0.96       390
   macro avg       0.95      0.93      0.94       390
weighted avg       0.96      0.96      0.96       390

0.9615384615384616
```

Figure 9. Display of random forest model evaluation results

From Figure 9, it can be seen that the recall value of the model is around 98%, which means that the model correctly identifies about 96% of retention customers. The precision of the model is about 97% and the f1-score is about 98%. The company may want to increase its monthly discount for at least new customers for the first 6 months, this may be a good strategy to help retain their customers.

The next step in feature evaluation is to make judgments about a program, improve its effectiveness, and to inform programming decisions. This examination involves gathering and analyzing information about program activities, characteristics, and outcomes. Let's see which features play the most important role in identifying customer retention. The random forest classifier contains an attribute named feature_importance which contains information about the most important features for a particular classification as shown in Figure 10 below:

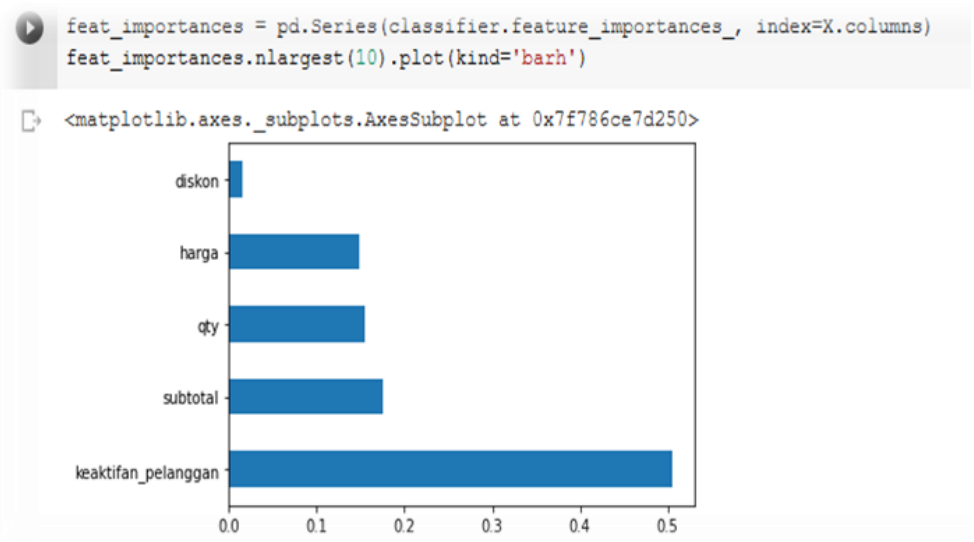

Figure 10. Display of feature evaluation results

Based on Figure 10 above, it can be seen that customer_activity has the highest influence on customer retention, followed by subtotal and qty. Machine learning is able to detect and select problems that exist in the program, including the collection of original data in the form of excel and so on, from the modeling results to form an accuracy of 390 lines in Figure 11 below:

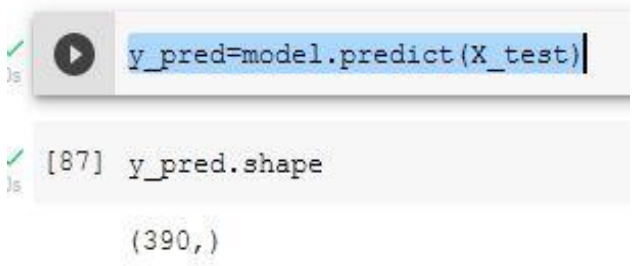

Figure 11. Display to see modeling accuracy

For the display of actual and predicted data frames, it can be seen in Figure 12 below:

```
df_test = pd.DataFrame({'Actual' : y_test, 'Prediction' : y_pred})
df_test
```

	Actual	Prediction
1614	1	1
1405	1	1
974	1	1
1055	1	1
307	0	0
...	...	...
618	1	1
426	1	1
966	1	1
849	1	1
1600	0	0

390 rows × 2 columns

Figure 12. Display of the actual results of the prediction

From Figure 12 the actual and predicted results using test data from the evaluation results of the random forest model there are 390 accuracy lines and to see the actual results or the original data with the predicted results seen in the figure if the actual and predictions show the same results. Furthermore, after the training process is carried out on a machine learning algorithm, the next step is to evaluate the performance of the algorithm or commonly called testing can be seen in Figure 13 below:

```
x_testing = data_testing[['qty','harga','diskon','subtotal','keaktifan_pelanggan']]
x_testing.head()
```

	qty	harga	diskon	subtotal	keaktifan_pelanggan
0	67	140000	0.0	9380000	1
1	6	140000	0.0	840000	1
2	21	140000	0.0	2940000	1
3	68	140000	0.0	9520000	1
4	80	140000	0.0	11200000	1

Figure 13. Display of algorithm performance results in data testing

From Figure 13 above, the fields used as testing are qty, price, discount, subtotal, customer activity. The value of customer activeness is 1, which means that 1 of the fields is a customer who is always active in the qty and price. Next, we will know the set of testing data with predictions seen in Figure 14 below.

```
[92] y_testing_pred = model.predict(x_testing)
     y_testing_pred

     array([1, 1, 1, ..., 0, 1, 0])

[93] df_testing_pred = pd.DataFrame({'Prediction': y_testing_pred}).round(2)
     df_testing_pred
```

	Prediction
0	1
1	1
2	1
3	1
4	1
...	...
1804	0
1805	0
1806	0
1807	1
1808	0

1809 rows × 1 columns

Figure 14. Display of the test data prediction set

From Figure 14 the results of the set of predictions with data testing, produce a prediction field that shows the results of number 1 (yes) number 0 (no), meaning that from these data it shows that which line of customers is still experiencing retention and is not churn. Next, to combine the prediction testing frame data above with testing data, we will use concat from the pandas library as shown in Figure 15 below:

```
[94] # buat var baru
     df_result_predict = pd.concat([data_testing, df_testing_pred], sort=False, axis=1)
     df_result_predict
```

nama_pelanggan	item	merk	jenis	qty	satuan	harga	subtotal	diskon	harga_nett	keaktifan_pelanggan	retensi_pelanggan	Prediction
V.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	67	Dus	140000	9380000	0.0	9380000.0	1	1	1
V.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	6	Dus	140000	840000	0.0	840000.0	1	1	1
V.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	21	Dus	140000	2940000	0.0	2940000.0	1	1	1
V.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	68	Dus	140000	9520000	0.0	9520000.0	1	1	1
V.Sari_Rasa_Nusantara	Air	Eternalplus	500ml	80	Dus	140000	11200000	0.0	11200000.0	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	0.0	0.0	0	0	0
0	0	0	0	0	0	0	0	0.0	0.0	0	0	0
0	0	0	0	0	0	0	0	0.0	0.0	0	0	0
cocoon	Air	Eternalplus	500ml	10	Dus	140000	1400000	0.0	1400000.0	1	1	1
0	0	0	0	0	0	0	0	0.0	0.0	0	0	0

Figure 15. Display of the results of combining the frame testing data and the original testing data

Figure 15 above is the result of a combination of frame testing data and testing data, we can also export it to an excel file to make it a saved file and it will automatically go to Google Drive. The following can be concluded the stages of customer retention prediction using the random forest algorithm in table 2 below:

Table 2. Conclusion of the prediction stage and results

No	Prediction stage	Result
1	Data collection	* Data retrieved from database in time period January 2020-March 2021 there are 1,950 rows and 15 columns. * Data is divided into 2 training data and testing data August 2020-March 2021 training data as many as 1,950 rows and 15 columns January 2020-July 2020 training data as many as 1,890 and 14 columns.
2	Data *preparation*	* Setting up the library used there are 10 libraries. * From the data there is no missing value. * From the results of statistical data, the biggest discount is 12%, taking the most qty will get the biggest discount according to the company tier. * From the visualization results, there are 81.12% customer retention, and 18.87% customers who do not retain or churn.
3	Data *cleaning*	* From the training data the deleted or irrelevant column, namely no_id.
4	Data *modeling*	* Prediction results using random forest number 1 (yess retention) is more than number 0 (no retention). * Evaluation of features, customer_activity fields that affect customer retention.
5	Machine learning	* The actual results and predictions of random forest models, showing 96% accuracy.

4. CONCLUSION

The results showed that the Random Forest Algorithm can be used to measure customer retention. The data consists of 1,950 rows and 15 columns with the fields, namely, field date, no_id, type_customer, customer_name, item, brand, type, qty, unit, price, subtotal, discount, price_nett, customer_activity, customer_retention. The results of data preparation processing can be seen that there are 1582 customers who retain retention and 368 customers who do not, with a percentage of 81.12% customer retention and 18.87% customer churn. The results of the feature evaluation in the random forest classifier method, show that customer_activity has the highest influence on customer retention, followed by subtotal and qty.

Suggestion: Based on the results of research that has been done, this system still has several shortcomings that need to be followed up. Suggestions for further development require more data so that the results displayed are more precise.

REFERENCES

[1] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, pp. 273–281, 2018.

[2] S. Nabavi and S. Jafari, "Providing a Customer Churn Prediction Model Using Random Forest and Boosted TreesTechniques (Case Study: Solico Food Industries Group)," *J. Basic. Appl. Sci. Res*, vol. 3, no. 6098, pp. 1018–1026, 2013.

[3] P. W. Anuang and P. K. Dyah, "Pengaruh Adopsi Teknologi dan Social Media Marketing Terhadap Minat Beli Serta Dampaknya Terhadap Keputusan Pembelian (Studi Pada Perusahaan Niluh Djelantik)," *Tiers Inf. Technol. J.*, vol. 1, no. 1, pp. 25–32, 2020.

[4] N. Hayati and D. Suryana, "Pengaruh Kepercayaan dan Komitmen Terhadap Loyalitas Pelanggan," *J.*

Sains Manaj. Akunt., vol. IV, no. 2, pp. 52–67, 2012.

[5] A. Ambarini, D. Novirani, and A. Bakar, "Upaya Peningkatan Kepuasan Pelanggan Indosat Berdasarkan Telecommunication Service Quality dengan Menggunakan Structural Equation Modelling (SEM)," *J. Online Inst. Teknol. Nas.*, vol. 02, no. 01, pp. 204–216, 2014.

[6] A. Dharmalau, Y. Suhanda, and lela Nurlaela, "Perancangan Sistem Informasi Pelayanan Purna Jual berbasis Customer Relationship management," *J. Rekayasa Inf. Swadharma (JRIS)*, vol. 01, no. 01, pp. 1–8, 2021.

[7] M. Asqia and T. Nabarian, "Pemanfaatan Google Sheets dan Google Form untuk Layanan Administrasi Mahasiswa Menggunakan Konsep Electronic Service Quality," *J. Teknol. Terpadu*, vol. 7, no. 1, pp. 15–22, 2021.

[8] N. Wayan, A. Karmila, W. Sunia, F. Ekonomi, and D. Bisnis, "Pengaruh E-Service Quality, Word Of Mouth, Price, dan Promotion Terhadap Minat Konsumen Menggunakan Layanan Jasa Go-Jek (Studi Kasus Pada Masyarakat Pengguna Go-Jek Di Kota Denpasar)," *TIERS Inf. Technol. J.*, vol. X, No.X, no. 39, pp. 41–54, 2020.

[9] S. Mulyana, E. Winarko, P. Studi, I. Komputer, and U. G. Mada, "TEKNIK VISUALISASI DALAM DATA MINING," vol. 2009, no. semnasIF, pp. 100–106, 2009.

[10] B. Prasojo and E. Haryatmi, "Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman dengan Metode Random Forest," *J. Nas. Teknol. dan Sist. Inf.*, vol. 7, no. 2, pp. 79–89, 2021.

[11] C. I. Agustyaningrum, W. Gata, R. Nurfalah, U. Radiyah, and M. Maulidah, "Komparasi Algoritma Naive Bayes, Random Forest Dan Svm Untuk Memprediksi Niat Pembelanja Online," *J. Inform.*, vol. 20, no. 2, pp. 164–173, 2020.

[12] I. Afdhal, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 1, pp. 49–54, 2022.

[13] A. Sharma, G. Deepak, N. Nikhil, S. Deepti, and V. Ankita, "Prediction of Customer Retention Rate Employing Machine Learning Techniques," *Int. Conf. Informatics*, vol. 1, no. 1, pp. 103–107, 2022.

[14] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019.