JOIV : Int. J. Inform. Visualization, 7(4) - December 2023 2280-2289

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Design of Prediction Model using Data Mining for Segmentation and Classification Customer Churn in E-Commerce Mall in Mall

Ilham Huda [a,*], Agus Achmad Suhendra [a], Moch Arif Bijaksana [b]

[a] *Industrial Engineering, Telkom University, Dayeuhkolot, Bandung, 40257, Indonesia*
[b] *Informatics Engineering, Telkom University, Dayeuhkolot, Bandung, 40257, Indonesia*
Corresponding author: [*]ilhamhuda@student.telkomuniversity.ac.id

Abstract—The classification of churn is driven by the potential risks e-commerce companies face, such as losing customers who discontinue their service usage or churn. Marketing specialists have shifted their efforts from acquiring new customers to retaining existing ones in order to mitigate customer churn. Predictive models are created using data mining techniques to identify customer churn patterns. This study proposes a data mining model aimed at predicting customer behavior, with the processed results utilized as suggestions for improvements and company strategies in customer retention through segmentation and classification. Segmentation and classification involve several variables: Session, Interaction with Application, Actions taken during the interaction, purchasing, claim, and discount. This study employs a clustering technique based on the Recency, Frequency, and Monetary (RFM) model, which considers factors such as the time since the last visit, the number of visits, and the total amount spent by the customer. The classification algorithm model was evaluated by comparing three classification algorithms: decision tree and Support Vector Machine (SVM). The decision tree algorithm had the highest accuracy, achieving an impressive 87% accuracy rate in customer classification. Factors influencing customer churn include purchasing behavior, session activity, claim feature utilization, adding products to cart, and discounts. Improving stock management is crucial to prevent stock shortages, likely to cause churn. Additional measures like sending emails/notifications and offering vouchers/loyalty points can be implemented for customers who added products to their carts but didn't complete the purchase, with a focus on popular products.

Keywords— Customer churn; churn prediction; data mining; e-commerce; decision tree; SVM.

Manuscript received 15 Apr. 2022; revised 23 Nov. 2022; accepted 19 Jan. 2023. Date of publication 31 Dec. 2023.

I. Introduction

E-commerce has opened numerous possibilities for consumers. The swift advancement of technology and the Internet has led to exponential growth in this industry. A company's success heavily relies on its ability to analyze customer behavior data effectively. The departure of customers, known as Customer Churn, can be viewed as a missed opportunity to generate profits. Acquiring new customers typically incurs a five to six times higher cost than retaining existing customers [1]. Consequently, marketing professionals have shifted their focus from acquiring new customers to retaining the ones they already have to maintain market share and minimize customer churn. As a result, customer churn, also referred to as customer turnover, friction, or deflection, is a significant concern in certain sectors. Consumers can easily compare products or services and switch vendors [2]

One of the challenges that *E-Commerce* companies face is the effort to reduce the number of customers who stop using the company's services and move to another company. The behavior of customers from a service that does not continue to subscribe to the company or the ratio between the number of customers who stop using the company's products/services in a certain period, with the average total number of customers in the same period and that can cut the company's profits is called the churn rate [3]. The customer segmentation process aims to classify customers based on their consumption behavior, preferences, and demands and provide targeted products and services according to the existing customer groups. It can benefit companies to develop targeted marketing programs and reduce customer retention costs.

The Data Mining method is a potential approach to addressing churn analysis and prediction in solving churn-related issues. Data mining involves using data analysis and modeling techniques to discover patterns and relationships within data, enabling companies to comprehend customer

preferences and predict future actions [4]. This study's data originates from the company's Customer Relationship Management (CRM) system. CRM refers to a system for managing and maintaining customer relationships to establish, manage, enhance, and sustain interactions between companies and customers. Analytical Customer Relationship Management (CRM) refers to implementing data mining techniques in CRM. Data mining is one of the analytical tools within analytical CRM and has demonstrated its effectiveness in helping companies predict customer attrition, anticipate churn occurrences, and evaluate the accuracy of these predictions [5]. In order to enhance marketing strategies, organizations are encouraged to incorporate analytical Customer Relationship Management (CRM) to gain competitive intelligence. By mapping customer segments, companies can effectively tailor their marketing approaches for each segment, thereby gaining deeper insights into customer behavior and facilitating a more efficient marketing process.

This research aims to employ a clustering method that relies on customer interactions with the company's services. The analysis considers several variables: session, behavior changes, interaction with the application, actions taken during the interaction, and rating. Additionally, in terms of complaints, the data is divided into three categories: data provisions, data complaints, and data repairs. To conduct the clustering process, two data mining techniques, SOM, and k-means, are utilized for comparison. It is essential to determine the most accurate classification technique for the given data to predict churn using classification techniques before forming a classification model. The data mining methods employed for classification in this study are Artificial Neural Network and Decision Tree. Following the implementation of the Clustering and Classification process in this research, various analyses have been generated, encompassing customer classification, customer churn, customer retention, and customer satisfaction analysis.

The subsequent stage involves making decisions based on the processing and analysis of customer information. This decision-making process aims to optimize data processing and management. The company's management decisions are closely tied to the outcomes of the CRM-Data Mining model, which provides the best results for devising a customer retention strategy as a recommendation for the company. This study incorporates a validation process by comparing the accuracy of various data mining techniques using a confusion matrix To ensure the high accuracy of churn prediction.

II. Materials and Method

A. Customer Churn

The behavior of customers from a service that does not continue to subscribe to the company or the comparison between the number of customers who stop using the company's products/services in a certain period, with the average number of customers in the same period and which can cut the company's profits is called the churn rate.

1) Data Mining

Data mining is a process that uses data analysis and modeling techniques to find patterns and relationships in data so that companies can understand what customers want and predict actions that customers will take. Data mining is a decision-support process that can look for patterns in the data to obtain previously unknown information [6]

2) Business understanding: The initial phase, also known as the research understanding stage, involves defining the research project's objectives and formulating data mining problems.

3) Data Understanding: In this stage, data collection occurs, and the gathered data is analyzed and evaluated for quality.

4) Data Preparation: Once the raw data is prepared, it is transformed into the final dataset used in the subsequent phases. The desired cases and variables are selected for analysis based on the identified problems, the data is cleaned, and certain variables may be transformed as necessary.

5) Modeling: At this stage, suitable modeling techniques are applied by configuring the model to optimize the desired outcomes.

6) Evaluation: One or more models are evaluated to determine if they have achieved the objectives set in the initial stage. Decisions are made based on the results of data mining and their relevance to the research objectives.

7) Deployment: The models created in the previous stage are deployed in two ways. Simple deployments involve generating reports based on the model's output. On the other hand, complex deployments involve comparing models and conducting parallel data mining processes in other domains.

B. Classification

The customer segmentation process is when companies classify customers based on their consumption behavior, preferences, and demands and provide targeted products and services according to different customer groups. In this model, customers are divided into three levels of consumer value by analyzing consumers.

1) Decision Tree: A decision tree is a classification method in data mining. This classification technique is interesting because it involves the construction of a decision tree. A collection of decision nodes is connected by branches that extend downward from the root node until it ends at the leaf node. Starting at the root node, which by convention is placed at the top of the decision tree diagram, the attributes are tested at the decision node, with every possible outcome yielding a branch. Each branch leads to another decision node or the last leaf node.

2) SVM: The Support Vector Machine (SVM) method is a learning system that uses a hypothetical space in linear functions in a feature with high dimensions and is trained using a learning algorithm based on optimization theory. The Support Vector Machine (SVM) method is new compared to other techniques. Selection of the right and appropriate kernel function is significant and necessary because the part of the kernel will determine the feature space where the role of the classifier will be searched.

C. Previous Research

Research to predict customer churn has been carried out using various data mining techniques that are used to predict churn in multiple fields, such as e-commerce, telecommunications, banking, and others. The following are some discussions of published research that focus on e-commerce, including:

1) Berger P, Kompan M[7] *:*Develop model predictions based on user interactions. The performance of the model is predicted by Real data churn prediction. The research concludes that predictions using the proposed model outperform churn predictions based on the basic model. Berger and Kompan mention that with further research, the methodology can be developed in e-Commerce applications.

2) Fridrich M [8]*:* Propose an optimization model based on an artificial neural network to predict customer churn using Genetic Algorithms in e-Commerce. Prediction models are developed to identify customers who are at risk of defecting. The proposed model leads to increased customer churn predictability based on True Positive rate, False Positive Rate, and accuracy. Fridric suggested that techniques such as Linear Regression enhancement, Decision Tree & fuzzy logic can be used. In addition, it is recommended to add more parameters in e-Commerce for better prediction accuracy.

In this research, the customer segmentation process aims to classify customers based on customer consumption behavior, preferences, and demands and provide targeted products and services according to existing customer groups. It can benefit companies to develop targeted marketing programs and reduce customer retention costs. Several modeling techniques were used. Data Mining Techniques for Prediction/Classification using Support Vector Machine (SVM) and Decision Tree.

III. Result and Discussion

A. Operational Variable

The author constructs the model with several categories of attributes, each offering a unique perspective on user interactions and activities with applications. The authors enter six categories of attributes that can be extracted from the database of one of the E-Commerce Mall in Mall companies. Let S be the set of attributes that describe a particular session, P set of attributes that represent customer purchases, F set of attributes that describe the frequency of customer interactions with the web application, A set of attributes that describe actions performed during the session, C set of attributes that describe Claims made by customer and D is a set of attributes that describe the Discount used by the customer.

TABLE I
SESSION (S)

Variable	Description
Number of Session	Number of Sessions
Last Date Session	Last session date.
Most Use Devices	The most frequently used type of device.
Session to Purchase Ratio	Number of Sessions / Number of Successful Transactions.
Last Date Purchase	Last date of purchase
Last Date Add to Cart	Last date added to cart

TABLE II
PURCHASE (P)

Variable	Description
Number of Purchases	Number of Successful Transactions.
Total Sum Paid	Total Paid Amount of All Transactions.
Last Sum Paid	Total amount paid on the last transaction.
Most Purchases by Mall	Mall with the most orders.
Most Purchases by Brand	The brand with the most orders.
Max Amount Paid	The highest total payment in each transaction
Min Amount Paid	The smallest total payment in each transaction
Most Large Category Product by Purchases	Category Tree products purchased
Most Middle Category Product by Purchases	Category Tree Products purchased
Most Small Category Product by Purchases	Category Tree Products purchased

TABLE III
INTERACTION WITH APPLICATION

Variable	Description
Number of sessions since last purchase	The number of sessions between user's last session and his/her most recent session with the purchase event.
Time since last purchase	The time in day between user's first action of actual session and the last action of his/her most recent session with the purchase event.
Time since last Add to Cart	Time since last visit - the time in Day between last action in user's previous session and his/her first action in add to cart.

TABLE IV
ACTION MADE IN INTERACTION (A)

Variable	Description
Number of Add to Cart	Total Add to Cart.
Add to Cart to Success Purchase	Number of Add to Cart / Number of Successful Transactions.
Add to Cart to Success Payment	Total Add to Cart / Total Payment Successful.
Add to Cart since Last Purchase	The number of Add to Cart since the last successful transaction.
Most Large Category Product by Add to Cart	Category Tree Most added products to cart
Most Middle Category Product by Add to Cart	Category Tree Most added products to cart
Most Small Category Product by Add to Cart	Category Tree Most added products to cart
Most Product Add to Cart by Mall	Mall category of most products added to cart

TABLE V
DISCOUNT (D)

Variable	Description
Most Use Voucher Type	The most used types of vouchers.
Number of Voucher Uses	Number of vouchers used.
Average Discount Amount	Average amount discount.
Min Discount Amount	The smallest discount amount.
Number of Promo	The total number of promos that have been used.

TABLE VI
CLAIM (C)

Variable	Description
Number of Claim	Total number of claims.
Last Claim Reason	Last reason to claim.
Most Claim Reason	Most claim reasons.

B. Exploring Data Analysis

In Exploring Data Analysis, in addition to examining the data variables used in this study. The author also conducted interviews with the Data Analyst at the E-Commerce Mall in Mall to identify the causes or indicators of customer churn. After conducting interviews, it was determined what variables were used in this study which was also in line with previous studies.

C. Data Pre-process

After collecting the required data, data is processed through data pre-processing. Data pre-processing involves cleaning, and eliminating missing data, misclassification, information inconsistencies, or outliers. The data cleaning process in this study was manually conducted using PHP MYSQL and SQL Developer. This study comprises four stages of data pre-processing: data collection, grouping, cleaning, and division into two types of datasets: RFM for segmentation and classification for prediction, as illustrated in Figure 1.

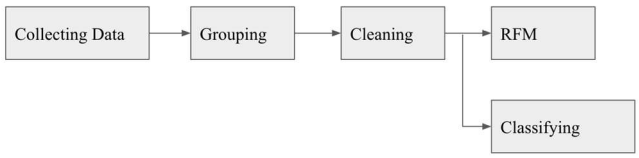

Fig. 1 Pre-processing

The first step, collecting data, aims to collect data based on the variables that have been determined in the Exploring Data Analysis. This company has two separate databases, for the category table is in database 1, while the session, voucher, discount, purchase, and add to cart tables are in database 2. The category table in database 1 shows data on categories' names. The category's name is related to the identity of the small category product, medium category product, large category product, mall, and brand variables contained in database 2.

Each table is exported in CSV format in the data collection process, which will then be processed using PHP MYSQL. After the Collect Data process is carried out, the id on the variable small category product, medium category product, large category product, mall, and brand, which is in the add to cart and purchase table, is related to the category table to get the name of each category. After that, the variable tables contained in the session, voucher, discount, purchase, and add-to-cart tables are combined into one CSV file using a query in PHP MYSQL.

After the Grouping process, the next step is to complete the data cleaning process. Data cleaning is also known as data cleansing or scrubbing. The data cleaning process can be used to determine inaccurate, incomplete, or incorrect data and improve data quality through error detection. Dealing with missing values by omitting the data from records can be dangerous because the missing values may be systematic and lead to biased data subsets. From the results of data cleaning, which is done manually with PHP MYSQL and SQL Developer from 14,806 customers to 12,806 data due to separating with variables that have data outliers.

D. Variable Analysis

Variable analysis is to ensure the relationship between variables. This section aims to explain the measurement of the value of the variables obtained starting with determining what variables are used. The variables used are variables that have been tested in previous studies. In addition, the variables used are variables obtained based on the results of interviews with experts, in this case the Head of Data Analyst, Marketing Analyst and Marketing Channel at E-Commerce Mall in Mall. It is obtained several variables that have been included in the model which later these variables will be useful in the process of classifying churn and non-churn customers.

TABLE VII
RFM MODEL

Variable	Description
Recency	Lag in days from the last data retrieval date to the last customer session date.
Frequency	Number of customers purchases.
Monetary	Total amount issued by the customer for the purchase

E. Clustering

In this section, there is a distinction made between churn customers and non-churn customers. However, this distinction is made manually rather than using data mining algorithms. During interviews with the company, non-churn customers are identified as customers who have had interactions within the past three months. Interactions are defined as activities such as signing in, adding products to the cart (add to cart), or making purchases. If none of these interactions occur, the customer is classified as churn. For this study, customer data from September 28, 2021, to January 26, 2022, will be utilized. The data processing is based on customer records from the E-Commerce Mall in Mall, focusing on customers who have been registered for at least the last six months and have made at least one transaction. The initial dataset consists of 14,806 customer records, which is reduced to 12,806 after pre-processing.

To apply the RFM (Recency, Frequency, Monetary) model, it is necessary to assign numerical rankings to customers in each of the three categories. Typically, a scale of 1 to 4 is used, where higher numbers indicate better results.

TABLE VIII
RECENCY SCALE

Value	Scale
1	More than 90 days
2	More than 60 days and less than 90 days
3	More than 30 days and less than equal to 60 days
4	Less than 30 days

TABLE IX
FREQUENCY SCALE

Value	Scale
1	Less than 2
2	More than with 2 and less than equal to 5
3	More than 6 less than equal to 10
4	More than 10

TABLE X
MONETARY SCALE

Value	Scale
1	Less than IDR. 50.000 or 3.5 US Dollar
2	More than IDR 50.000 or 3.5 US Dollar and less than equal to IDR 100.000 or 6.9 US Dollar
3	More than IDR 100.000 or 6.9 US Dollar and less than IDR. 1.000,000 or 69.9 US Dollar
4	More than IDR. 1,000,000 or 69.9 US Dollar

After forming four segments, these segments are grouped into RFM Group. Each rfm group has a different category.

TABLE XI
RFM CATEGORY

Category	RFM Score	Description
Best Customer	444	Customers with the highest score of visit time, number of purchases, and nominal purchases.
Loyal Customer	X4X	Customer with the highest purchase amount score
Potential Big Spender	4X4	Customers with the highest visit time score and purchase amount.
Need Attention Big	XX4	Customers with the highest nominal purchase score
Recent Customer	4XX	Customer with the highest visit time score
Reguler Customer	Others	Customers with the highest score of visit time, number of purchases and nominal purchases.
Churn Customer	1XX	Customer with the lowest visit score

The clustering results will be used as a data reference in data mining classification in determining churn and non-churn customers.

F. Classification

In the preceding section, customer churn clustering and predictions for customer non-churn were performed. Subsequently, the customer classification process was carried out utilizing a data mining classifier method. This model embodies the knowledge that will be employed to forecast classes for new data. The data mining model utilized in this study incorporates the optimal clustering technique, with the results being applied using the finest classification technique. The classification algorithms assessed in this research encompass Support Vector Machine and a decision tree.

To establish a classification model from the data, evaluating which classification technique yields the highest accuracy for this dataset is necessary. The data mining methods employed for classification encompass Artificial Neural Networks and Decision Tree. The classification process in this study was conducted using the Orange3 software.

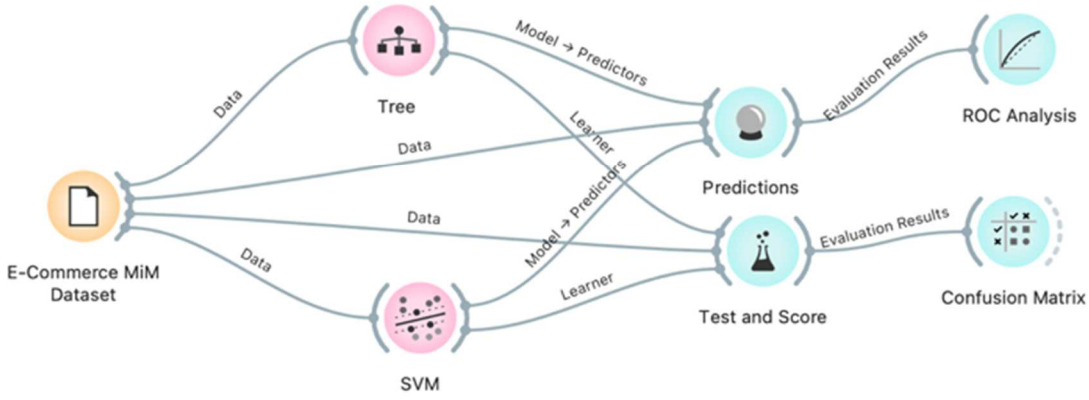

Fig. 2 The process of testing the classification method using Orange3

In the previous testing phase, it was established that 80% of the data was allocated for the training set and 20% for the test set. This allowed for evaluating the accuracy results of the three classification models in the 'Test & Score' section, as depicted in Figure 1. The outcomes of the classification algorithm test are presented in Figure 3.

Model	AUC	CA	F1	Precision	Recall
Tree	0.745	0.875	0.850	0.840	0.875
SVM	0.621	0.756	0.783	0.818	0.756

Fig. 3 The results of testing the classification algorithm in Orange 3

In this study, two different classification methods have been introduced for churn prediction. Using Orange3 software, both methods were tested to suit the dataset used in the study. Both use a standard model, a decision tree with a split subset of at least 1, and a Support Vector Machine. This part of the study aims to compare the performance and accuracy of the two methods. Figure 4. shows the test results of the three algorithms. The decision tree algorithm has a higher classification accuracy (CA) value than other algorithms. The decision tree value is always superior to the Support Vector Machine method in this dataset.

		Predicted		
		churn	non-churn	Σ
Actual	churn	969	5051	6020
	non-churn	1368	43892	45260
	Σ	2337	48943	51280

Fig. 4 Confusion matrix decision tree

		Predicted		
		churn	non-churn	Σ
Actual	churn	1992	4028	6020
	non-churn	8464	36796	45260
	Σ	10456	40824	51280

Fig. 5 Confusion matrix Support Vector Machine

The suitability rate for churn customers from the decision tree is much higher than other methods, with an accuracy rate of 87%. In comparison, the Support Vector Machine has a customer churn rate of 75%. This classification accuracy value is shown in Figure 4. And Figure 5. So, considering that it has the highest accuracy compared to other methods, it can be concluded that the decision tree is chosen as the best technique among other techniques for classifying churn customers. The model testing phase includes predicting test data instances that depend on the right classifier decision tree. Classifier testing is also based on the cluster results with the highest accuracy level.

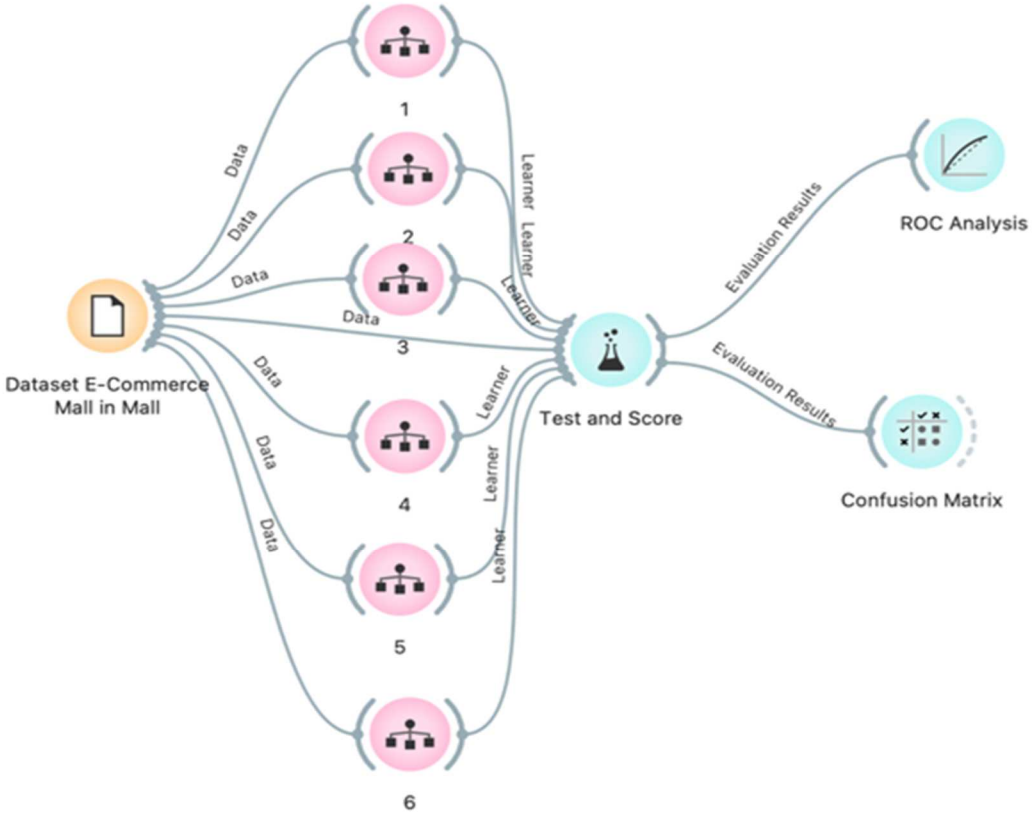

Fig. 6 Classification Decision Tree model process using Orange 3

The process in Figure 6 shows that there are six decision tree models. This learning model will be tested with test data in the 'Test & Score' section. As mentioned, the test data values are proportional to 20% of the overall dataset file. This division helps evaluate the prediction accuracy of the classification model made. Furthermore, the test results are visualized in the form of ROC Analysis and Confusion Matrix.

The decision tree models are different because of the results of the pruning technique. Pruning is a machine learning technique that reduces the decision tree's size by removing parts that have little power to classify instances. Pruning can reduce complexity. Besides that, it can improve prediction accuracy by reducing overfitting. The accuracy of the classification generated from the pruning technique in the decision tree model is shown in Figure 7.

Model	AUC	CA	F1	Precision	Recall
6	0.780	0.881	0.847	0.842	0.881
5	0.781	0.880	0.847	0.842	0.880
4	0.780	0.881	0.847	0.842	0.881
3	0.781	0.880	0.846	0.842	0.880
2	0.779	0.881	0.847	0.843	0.881
1	0.742	0.877	0.847	0.838	0.877

Fig. 7 Decision tree model evaluation results

The best results from comparing the 6 different models fell to model 2, although the other models had very slight differences.

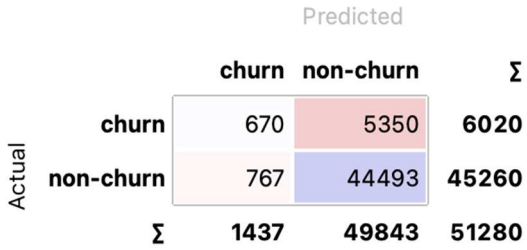

Fig. 8 Confusion Matrix model 2 Decision Tree

The level of accuracy generated by the classification model built is 88.3%. This shows that the performance of the customer classification model is quite good.

G. Data Analysis

The analysis phase involves interpreting the processed data and utilizing the analysis results for decision-making purposes, marking the culmination of this research. Decision-making entails providing recommendations for improvement and strategies to be implemented by the company's management in customer retention. This serves as a response to the problem initially formulated in the study. The initial analysis to be conducted involves examining customer profile data.

The Customers in this research have signed up for PT XYZ's services by September 28, 2021, and have conducted at least one transaction. The customer data used in this study is sourced from the company's database and has undergone pre-processing using SQL Developer, PHP Hypertext Pre-processing Apache, and Orange. As a result, the dataset now consists of 12,806 customer records that are prepared for further analysis.

The analysis conducted in this journal focuses on customer data obtained from historical records of various activities performed while utilizing the service. These activities include interactions with features such as add to cart, purchasing, voucher, discount, and claim. The dataset also contains data records related to the products, specifically the small category product, medium category product, large category product, brand, and mall. The analysis of each category within the respondent profile is described as follows:

1) Small Category Product: Small Category Product is the lowest category of product categories in PT.XYZ. There are 426 categories recorded from the data recorded during the last three months of transactions.

TABLE XII
SMALL CATEGORY PRODUCT

Category Name	Number of Customer	Number of Customer Churn
Skin care	1.347	100
Serums & Treatments	972	54
Sling bag	826	64
Sachet	266	7
Face cleanser	527	25
Face mask	467	31
Face Moisturizer & Cream	395	12
Coconut oil	150	1
13 years old	59	0
Superior	316	36
Run	289	14
Shampoo	225	14
Face	220	5
Pack	75	1

Table XII shows 14 categories with the highest transactions, including 1-3 years, tops, face masks, coconut oil, moisturizers & facial creams, facial cleansers, skincare, sachets, and serums & treatments and sling bags. The skincare category has the highest purchase, namely 1,347 customers, with a percentage of 10.5%. The following order is Serum & Treatment, with 972 customers with 7.5%.

2) Medium Category Product: Medium Category Product is a category that includes a small category of products that are in PT.XYZ. There are 162 categories recorded from the data recorded during the last three months' transactions.

TABLE XIII
MEDIUM CATEGORY PRODUCT

Category Name	Number of Customer	Number of Customer Churn
Skin care	2590	143
Beauty	1579	108
Woman's bag	997	68
Shoe	856	60
Selling Baby Formula Susu	529	52
accessories	200	2
Cooking oil	458	57
Clothes	432	24
Perfume	318	26
Hair Care	324	7
make up	78	1
Instant coffee	112	2
Instant noodles	245	7
Health Support Tools & Accessories	2590	143

Categories with the highest transactions in the medium category include Accessories, Baby Formula Milk, Beauty Cooking Oil, Clothing, Perfume, Skin Care, Hair Care, Shoes and Women's Bags. The skincare category has the highest purchase, namely 2,590 customers with a percentage of 20%. The following order is Beauty, with 1,579 customers with 12%.

3) Large Category Product: Large Category Product is a category that includes small and medium product categories in PT.XYZ. There are 43 categories recorded from the data recorded during the last three months of transactions.

TABLE XIV
LARGE CATEGORY PRODUCT

Category Name	Number of Customers	Number of Customer Churn
Beauty & Health	4087	206
Beauty, Shoes & Bags	1696	114
Fashion	1472	93
Man	1381	128
Food and Ingredients	508	11
Baby Needs	472	26
English books	286	7

Category Name	Number of Customers	Number of Customer Churn
Woman	451	34
Home Supplies	348	28
Beauty	179	16
Coffee and Tea	63	0
Electronics & Home	121	3
Health	201	20
Sports & Hobbies	161	11

There are 14 categories with the highest transactions, including Beauty, English Books, Fashion, Home Needs, Beauty & Health, Beauty, Shoes & Bags, Baby Supplies, Food and Cooking Ingredients, Men and Women. The beauty & health category became the category with the highest purchase, namely 4,087 customers with 31%. The following order is beauty, shoes & bags with 1,773 customers with 13%.

4) Most Devices: When customers engage with services provided by e-Commerce Malls, particularly at PT. XYZ, they have the option to access these services through four different platforms. These platforms include the Android Application, Website (WEB EC), Mobile Website (Web MC), and iPhone Application. The perceived user experience can differ for each platform. Mobile users can download the Mobile Apps from the respective application marketplaces, namely the Playstore and the App Store.

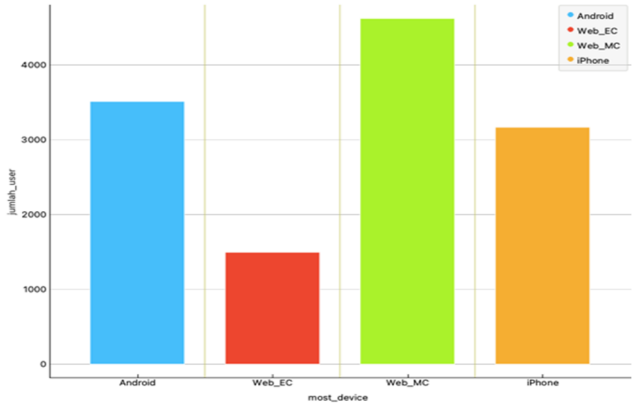

Fig. 9 Most Device

Most users, totaling 4,625 customers or 36%, primarily utilize the Web MC platform, which involves accessing services by opening a website using a mobile phone browser. Android is the second most popular platform, with 3,514 customers accounting for 27% of the user base. The iPhone platform ranks third, with 3,170 customers (24.7%) who have installed the application on their devices. The Web EC platform has the lowest user count, with 1,497 users representing 11.6% of the total.

5) Claim: When making a transaction, the customer is in the process of claiming or returning goods. Customers used three categories of claims for the last three months, including cancel, return, and exchange, with 14 types of claim reasons.

TABLE XV
CLAIM FAULT

Category Name	Number of Customers	Number of Customer Churn
Cancel	1214	112
Return	4	0
Exchange	195	22

Customers most widely use the claim category in category 2. It is Cancel, which means the purchase is canceled. 1, 214 customers used the Cancel Claim category, with 112 in churn.

TABLE XVI
CLAIM REASON

Category Name	Number of Customers	Number of Customer Churn
Change of mind	244	18
Price dissatisfaction	31	0
Option/Size/Qty wrong choice	203	14
Similar product purchase	43	6
Shortage	1038	94
Orders Confirmation Unprocessed	48	4
Other product shortage	20	2
Forgot to input the voucher	201	18
Discount/Promo not working	72	7
Force Return	90	12
Defect product	24	1
Specification difference with website	61	6

In Table XVI, the most reason customers cancel is a shortage or insufficient stock of products ordered by customers. The Shortage category was the most claim, 1,038 customers, of which 94 customers is churn.

H. Customer Loyalty

This study goes beyond examining churn customers and also investigates non-churn customers using RFM variables for segmentation. The author identifies six categories of customer loyalty within the non-churn group: best customer, loyal customer, potential big spender, need attention big, recent customer, and regular customer.

Figure 10 is a graph of customer segmentation at the E-Commerce mall in the mall company. The category with the highest number of customers is in the potential big spender category, with a total of 3,539 customers, which means that the most customer categories are customers who have the highest recency score and monetary score.

Fig. 10 Graph of the number of customers based on customer loyalty

I. *Managerial Implication*

The managerial implication derived from this research pertains to enhancing productivity within the e-Commerce Mall in Mall company at PT XYZ, focusing on decision-making. Decision-making serves as the culmination of this research and represents a crucial step in maximizing available resources, particularly information resources, for data processing and management.

This research investigates the impact of various factors on customer churn. The findings of this study reveal that several factors significantly influence customer churn, as identified through data mining classification. These factors include characteristics related to purchasing behavior, session activity, utilization of the Claim feature, add-to-cart actions, and discount usage. Based on these outcomes, there are opportunities to enhance effectiveness through some strategies.

The first is considering the attributes of claims and the outcomes derived from the decision tree analysis; various categories of Claims have been identified as having a substantial impact on churn. Specifically, customers who file claims by canceling transactions due to product shortage are associated with a 50% probability of churn, indicating a significant influence on customer attrition. This category comprises a total of 1,038 customers who have utilized this specific claim. Consequently, it is recommended that the company enhance its stock management processes for each mall to prevent future product shortages or insufficient stock levels.

The second is claims with a higher probability of churn account for 60% of cases, but they are associated with a relatively low amount. Specifically, issues in the Discount/Promo category, where the discount or promotional offer does not function as intended, and differences in product size between what was purchased and received have been identified as factors influencing churn. Consequently, the company should prioritize improving the functional systems, including both the website and mobile platforms, to ensure the proper implementation of the discount feature. Additionally, individual malls should focus on ensuring that products shipped match the size specified by customers in their orders.

The last is customers who engage in a single session and make purchases in the oil and clothing categories exhibit a 67% churn probability. However, if a customer adds the product back to the cart following the transaction, it indicates an intention to repurchase the same item. Based on these characteristics, the company can implement follow-up measures such as sending Email/Notifications and offering vouchers/points to customers who have added products to the cart but have not completed the purchase process.

IV. CONCLUSION

The conclusions in this study answer the questions that exist in the formulation of the problem to meet the research objectives based on data processing and analysis that has been carried out. This research produces theoretical implications regarding the design of the CRM-Data Mining model in the E-Commerce industry. The findings from the segmentation process using RFM revealed that there were 1,503 customers classified as churn, accounting for 11% of the total, while there were 11,303 customers classified as non-churn. Among the variables analyzed, the product category purchased, Claim Reason, Session, and Add to Cart were identified as significant factors influencing customer churn. These variables were found to be present in the characteristics of the company's services. The processed data using the accurate decision tree classification model demonstrated that the variables above had a notable impact on customer churn based on the research hypothesis.

A precise predictive model for customer behavior prediction is designed using a Data Mining architecture model. The accuracy of this model has been thoroughly examined through the process of segmentation and classification, along with the influential variable model. The variable model employed for segmenting churn and non-churn customers is the Recency, Frequency, and Monetary (RFM) model, which considers the time elapsed since the last visit, the frequency of purchases, and the monetary value of purchases.

The classification algorithm model was evaluated by comparing three classification algorithms: the decision tree and Support Vector Machine (SVM). Among these algorithms, the decision tree algorithm exhibited the highest level of accuracy. The decision tree model achieved an impressive accuracy rate of 87% in accurately classifying customers.

Factors influencing customer churn include purchasing behavior, session activity, utilization of the Claim feature, adding products to the cart, and discounts. The company needs to enhance the stock management process for each mall, ensuring that stock shortages, which have the highest likelihood of causing churn, are avoided in the future. Furthermore, the company can implement follow-up measures such as sending emails or notifications and offering vouchers or loyalty points to customers who have added products to their carts but have not completed the purchase, particularly focusing on the most popular products.

The model validation results, obtained through ROC analysis and Confusion Matrix, indicate that the six decision tree models achieved an accuracy rate of 88%. This value signifies a strong performance as it surpasses the threshold of 75%, indicating the model's effectiveness.

REFERENCES

[1] M. Chinnu, J. P. G. Scholar, M. Paul, and P. Mathai, "Customer Churn Prediction:A Survey," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, [Online]. Available: www.ijarcs.info

[2] M. Pondel et al., "Deep Learning for Customer Churn Prediction in E-Commerce Decision Support," Business Information Systems, pp. 3–12, Jul. 2021, doi: 10.52825/bis.v1i.42.

[3] M. Malleswari, R. J. Manira, P. Kumar, and . M., "Comparative Analysis of Machine Learning Techniques to Identify Churn for Telecom Data," International Journal of Engineering & Technology, vol. 7, no. 3.34, p. 291, Sep. 2018, doi: 10.14419/ijet.v7i3.34.19210.

[4] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2011.

[5] F. Buttle and S. Maklan, Customer Relationship Management. Routledge, 2019. doi: 10.4324/9781351016551.

[6] K. Parsaye and M. Chignell, *Intelligent Database Tools & Applications: Hyperinformation Access, Data Quality, Visualization, Automatic Discovery*. 1993.

[7] P. Berger and M. Kompan, “User Modeling for Churn Prediction in E-Commerce,” IEEE Intelligent Systems, vol. 34, no. 2, pp. 44–52, Mar. 2019, doi: 10.1109/mis.2019.2895788.

[8] M. Fridrich, “Experimental Parameter Tuning of Artificial Neural Network in Customer Churn Prediction,” Trends Economics and Management, vol. 11, no. 28, p. 9, Jun. 2017, doi:10.13164/trends.2017.28.9.

[9] S. Pratidina, "A Design of Data Mining Model in Customer Relationship Management for Patient Churn Segmentation and Classification in Obstetrics and Gynecology Clinic of Hospitals," in Proceedings of the 2020 4th International Conference on Computational Intelligence and Applications, 2020, pp. 15-19, doi:10.1109/ICCIA49802.2020.9189784.

[10] M. Azeem and M. Usman, “A fuzzy based churn prediction and retention model for prepaid customers in telecom industry,” International Journal of Computational Intelligence Systems, vol. 11, no. 1, p. 66, 2018, doi: 10.2991/ijcis.11.1.6.

[11] J. Vijaya and E. Sivasankar, “An efficient system for customer churn prediction through particle swarm optimization-based feature selection model with simulated annealing,” Cluster Computing, vol. 22, no. S5, pp. 10757–10768, Sep. 2017, doi: 10.1007/s10586-017-1172-1.

[12] E.-B. Lee, J. Kim, and S.-G. Lee, “Predicting customer churn in mobile industry using data mining technology,” Industrial Management & Data Systems, vol. 117, no. 1, pp. 90–109, Feb. 2017, doi: 10.1108/imds-12-2015-0509.

[13] A. Babkin and I. Goldberg, “Incorporating Time-Dependent Covariates into BG-NBD Model for Churn Prediction in Non-Contractual Settings,” SSRN Electronic Journal, 2017, doi:10.2139/ssrn.2905307.

[14] V. R. Hananto, A. D. Churniawan, and A. P. Wardhanie, “Perancangan Analytical CRM untuk Mendukung Segmentasi Pelanggan di Institusi Pendidikan,” Jurnal Ilmiah Teknologi Informasi Asia, vol. 11, no. 1, p. 79, Feb. 2017, doi: 10.32815/jitika.v11i1.55.

[15] A. Yulianto and F. Firmansyah, “Prediksi Customer Churn Pada Bisnis Retail Menggunakan Algoritma Naïve Bayes,” remik, vol. 6, no. 1, pp. 41–47, Nov. 2021, doi: 10.33395/remik.v6i1.11196.

[16] E. A. el Kassem, S. Ali, A. Mostafa, and F. Kamal, “Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 5, 2020, doi: 10.14569/ijacsa.2020.0110567.

[17] B. Zhang, “Customer Churn in Subscription Business Model—Predictive Analytics on Customer Churn,” BCP Business & Management, vol. 44, pp. 870–876, Apr. 2023, doi:10.54691/bcpbm.v44i.4971.

[18] E. Zdravevski, P. Lameski, C. Apanowicz, and D. Ślęzak, “From Big Data to business analytics: The case study of churn prediction,” Applied Soft Computing, vol. 90, p. 106164, May 2020, doi:10.1016/j.asoc.2020.106164.

[19] H. Dang Tran, N. Le, and V.-H. Nguyen, “Customer Churn Prediction in the Banking Sector Using Machine Learning-Based Classification Models,” Interdisciplinary Journal of Information, Knowledge, and Management, vol. 18, pp. 087–105, 2023, doi: 10.28945/5086.

[20] M. Arowolo, B. Jimada-Ojuolape, S. Yakub, A. S. Olaniyi, A. M. Olaolu, and S. Y. Kayode, "Customer Churn Prediction in Banking Industry Using K-Means and Support Vector Machine Algorithms," International Journal of Multidisciplinary Sciences and Advanced Technology, vol. 1, no. 1, 2020, doi: 10.5281/zenodo.4543690.

[21] T. J. Shen, A. Samad, and B. Shibghatullah, "Developing Machine Learning and Deep Learning Models for Customer Churn Prediction in Telecommunication Industry," 2022. [Online]. Available: www.kaggle.com.

[22] A. Mishra and U. S. Reddy, “A comparative study of customer churn prediction in telecom industry using ensemble based classifiers,” 2017 International Conference on Inventive Computing and Informatics (ICICI), Nov. 2017, doi: 10.1109/icici.2017.8365230.

[23] N. Jajam and N. Panini Challa, “Dynamic Behavior-Based Churn Forecasts in the Insurance Sector,” Computers, Materials & Continua, vol. 75, no. 1, pp. 977–997, 2023, doi:10.32604/cmc.2023.036098.

[24] A. Dingli, V. Marmara, and N. S. Fournier, “Comparison of Deep Learning Algorithms to Predict Customer Churn within a Local Retail Industry,” International Journal of Machine Learning and Computing, vol. 7, no. 5, pp. 128–132, Oct. 2017, doi:10.18178/ijmlc.2017.7.5.634.

[25] Y. Liu, J. Fan, J. Zhang, X. Yin, and Z. Song, “Research on telecom customer churn prediction based on ensemble learning,” Journal of Intelligent Information Systems, vol. 60, no. 3, pp. 759–775, Sep. 2022, doi: 10.1007/s10844-022-00739-z.

[26] X. Zhao, “Research on E-Commerce Customer Churning Modeling and Prediction,” The Open Cybernetics & Systemics Journal, vol. 8, no. 1, pp. 800–804, Dec. 2014, doi: 10.2174/1874110x01408010800.

[27] W. Pedrycz, “Introducing WIREs Data Mining and Knowledge Discovery,” WIREs Data Mining and Knowledge Discovery, vol. 1, no. 1, pp. 1–1, Jan. 2011, doi: 10.1002/widm.17.

[28] D. T. Larose and C. D. Larose, "Discovering Knowledge in Data an Introduction to Data Mining Second Edition Wiley Series on Methods and Applications in Data Mining."

[29] Yulianti, "Metode Data Mining Untuk Prediksi Churn Pelanggan," Jurnal ICT Akademi Telkom Jakarta, no. 17.

[30] A. Kolomiiets, O. Mezentseva, and K. Kolesnikova C. A., "Customer Churn Prediction in the Software by Subscription models IT business using machine learning methods."