Predicting Customer Churn and Assessing Sales Performance in Fuel Sales Using Decision Tree Algorithm and K-Means Clustering

Kania Lovia Tiarazahra[1]*, Rita Ambarwati[2],
Management Study Program, University of Muhammadiyah Sidoarjo, Indonesia[1], [2]
University of Muhammadiyah Sidoarjo
Sidoarjo, Indonesia
ritaambarwati@umsida.ac.id [2]

***Abstract*— Fuel oil, vital for community activities, sees PT Pertamina as a key supplier in Indonesia's competitive fuel market. To maintain its market edge, PT Pertamina must understand and predict customer behavior and loyalty. This study uses a combination of the RFM method, decision tree algorithms, and k-means clustering to analyze customer loyalty, salesperson performance, and potential customer churn, with a focus on the relationship between churn and salesperson expertise. The research utilizes sales transaction data from PT Pertamina Patra Niaga Regional Jatimbalinus. Findings from the decision tree algorithm show a high loyalty rate among customers, with 100% accuracy, precision, and recall in the confusion matrix. Salespersons are categorized into two performance-based clusters: 'good performance' and 'poor performance', determined by the lowest Davies-Bouldin values. A key discovery is the identification of a customer group likely to churn, a factor not influenced by salesperson performance. This insight is crucial for PT Pertamina in strategizing to reduce churn and improve market competitiveness. This study provides PT Pertamina with valuable insights into customer loyalty and sales performance. Understanding these aspects can help the company refine its strategies, potentially reducing churn and bolstering its position in the fuel market.**

Keywords*— Churn, Customer Segmentation, Salesman Clustering, Decision Trees, K-Means Clustering

I. Introduction

Fuel Oil, or often called BBM, is one of the needs that underlies all community activities, starting from personal activities, to industrial activities. Fuel plays an important role as an element of the economic wheel, both directly and indirectly. Where BBM can be an item that is traded, it can also be a mandatory item so that the operational wheels can turn. PT Pertamina is a state-owned enterprise, one of the directions of the company's movement is to provide fuel, both for the retail sector and the industrial sector. In accordance with Law No.8 of 1971 PT Pertamina is authorized by the government to produce and process oil and gas products as well as provide fuel oil and gas needs in Indonesia [1].

As time goes by, the population is increasing, the need for fuel is also getting higher. This is inversely proportional to the availability of fuel in nature which is getting thinner. So that according to Law No.22 of 2001, the government opened the door for private companies and changed the role of PT Pertamina to no longer be the only business actor in the oil and gas sector.

So PT Pertamina Patra Niaga needs to look a few steps ahead from now on to see potential profits and losses, one of which is by examining consumers who have the potential to move to competing companies or competitors, hereinafter referred to as churn. Then look for a correlation between their churn potential and the performance of the salesmen that PT Pertamina Patra Niaga has. With the hope that in the future this research can become a new reference in developing sales strategies.

II. Literature Review

A. Churn

Churn is a term that refers to the meaning of switching consumers to competing companies. As written in the research conducted by Radius Poniman, Opim Salim, and Sri Mulyani in the article entitled "Churn Analysis of Cellular Telecommunication Broad Band Services in Sumbagut", the definition of churn is the number of service users who no longer subscribe to service usage [2]. According to Agung Rezkina's view stated in his research entitled "Telkomsel Surakarta's Marketing Public Realtions Strategy in Maintaining Customer Loyalty", it is stated that churn is a word used to define customer switching to another operator [3]. So, if churn is placed in this research, the meaning of churn will be adjusted to become the movement of consumers to other competing companies.

This research is focused on the corporate sector where customers are grouped into several groups, namely Industrial Fuel Agents, Bunker Agents, Sea Transportation, BU-PIUNU, Industry, Government Agencies, Public Services, Fisheries, PLN, POLRI, TNI-AD (Army), TNI-AL (Navy), TNI-AU (Airforce), and Land Transportation. Also accompanied by research on salesmen in the corporate sector sales department totaling 8 people. The data used in the research is fuel sales transaction data for a period of 2 years, namely 2021-2022. The method used in finding answers to these objectives uses several methods, namely starting with scoring customer groups through the RFM model (Recency, Frequency, Monetary) to find out

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.2030, Copyright ©2024
Submitted : December 6, 2023, Revised : January 23, 2023, Accepted : January 26, 2023, 2024, Published : February 15, 2024

which costomer groups have the potential to move to competitors or not. The same method is also used in scoring salesmen in interpreting their performance track records which will then be interpreted in Decision Tree C4.5 and K-Means Clustering modeling.

B. RFM Model

The first method carried out in the research is an analysis method using the RFM model. The RFM model is used to identify characters in each customer group with the aim of forming a class of each customer group [4]. Analysis with this model is a classic approach that is carried out to determine the interaction patterns and customer patterns in general. After knowing the classes of customers with each different category, then calculated using the C4.5 decision tree algorithm to visualize the results of data processing in the form of a decision tree based on decision-forming criteria [5].

C. Decision Tree C4.5

Decision tree is one of the methods used to classify data. This algorithm is in the form of a decision tree consisting of nodes and edges. The nodes in this tree are further divided into 3, namely the root node marked with the attribute name, then the branching nodes labeled with attribute values, and the leaf nodes marked with different classes [22] In this research, the decision tree method is paired with the C4.5 algorithm. The C4.5 algorithm is a refined form of the ID3 algorithm. The advantages of C4.5 over ID3 lie in the following 4 parts: C4.5 is more resistant to data noise, C4.5 is better able to handle variables that have missing values, C4.5 is able to handle variables with discrete and continuous types, and C4.5 is able to prune branches of the decision tree.

Forming a C4.5 decision tree involves several key steps. First, the algorithm determines the root attribute from which the tree will start growing. Then, it forms branches based on each possible value of the selected attribute. After that, the algorithm sorts the cases for each branch, essentially assigning the data points to different paths based on their attribute values. This process is repeated continuously, forming each class, until the stopping criteria are met. Through these steps, the C4.5 algorithm builds a decision tree that can be used for classification and prediction.

In the process of determining the attribute to be the root of the decision tree, it is taken from the highest gain value of each existing attribute. And to calculate the gain can be obtained from the formula in the following equation:.

$$Gain(S, A) = Entropy(\mathrm{S}) - \sum_{\mathrm{n}=1}^{\mathrm{n}} \frac{|S_i|}{|S|} * Entropy(Si)$$

Description:
S: Set of Cases
A: Attributes n: Number of partitions of attribute A
$|Si|$: Number of cases in the i-th partition
|S|: Number of cases in S

But before getting the gain value, we must first find the entropy value to generate an attribute. The basic formula that can be used to find entropy can be found in the following equation:

$$Entropy(S) = \sum_{i=1} -pi * log_2\, pi$$

Description:
S: The set of cases
n: Number of partitions of S
pi: Proportion of Si to S

D. K-means Clustering

K-means clustering is one of the data mining modeling in grouping data according to their closest characteristics [6]. This method also uses a descriptive model to explain the algorithm for grouping objects with similar characteristics. The calculations carried out in this method begin by taking x

as an input parameter, each of which will later become the

center of cluster x. In this k-means clustering method, clustering depends on the initial centroid point which is then calculated by the Euclidean formula against the centroid. Data elements that have the smallest distance to the centroid are grouped into one cluster. the process continues to repeat until there is no more cluster movement.In this study, using the davies bouldin method, 2 groups were found with the closest characteristic distance to the centroid which is translated as "good performance" and "poor performance" which will then be further investigated regarding its correlation with customer groups that have the potential to churn.

Decision tree modeling and k-means clustering have provided many results in previous research. For example, in 2018 research was carried out by Ni Wayan Wardani, Gede Rasben Dantes, and Gede Indrawan regarding predictions of consumer loyalty in a retail company which was also carried out using decision tree modeling, and the results of this test showed that there were 3 predictions that were lacking. from the 259 cases studied, which means that data modeling with decision trees is quite accurate [5]. Also the same for k-means clustering. Research has been carried out in 2021 by Elly Muningsih, Ina Maryani, and Vembria Rose Handayani regarding provincial clustering based on village potential, and the division of clusters in this research is accurate which is proven by the Davies Bouldin method which shows the optimization results for the number of clusters is 3 with the Davies Bouldin value index (DBI) which shows the number 0.175, which is smaller than the number of DBIs in other clusters [7].

III. Research Methods

Decision tree modeling and k-means clustering have provided many results in previous studies. For example, in 2018 research was conducted by Ni Wayan Wardani, Gede Rasben Dantes, and Gede Indrawan regarding the prediction of consumer loyalty in a retail company which was also carried out with decision tree modeling, and the results of the test showed that there were 3 incorrect predictions out of 259 cases studied, which means that data modeling with this decision tree

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.2030, Copyright ©2024
Submitted : December 6, 2023, Revised : January 23, 2023, Accepted : January 26, 2023, 2024, Published : February 15, 2024

is quite accurate [5]. Also the same for k-means clustering. Research has been carried out in 2021 by Elly muningsih, Ina Maryani, and Vembria Rose Handayani regarding provincial clustering based on village potential, and the division of clusters in this study is accurate which is proven by the davies bouldin method which shows the results of optimizing the number of clusters is 3 with a davies bouldin index (DBI) value that shows a number of 0.175, which is smaller than the number of DBIs in other clusters [7].

This research focuses on the attachment of 2 different data processing results, namely between data processing results related to customer groups, with data processing results related to salesmen. Researchers have a simple assumption that the customer group that has the potential to churn has a big influence from salesmen who are less skilled in attracting customers. So from this research it will be proven, is it true if the group of customers who have the potential to churn is caused by salesmen who are less skillful in attracting consumers. Because if examined through past research, the shrewdness of a salesman at work, especially in communication skills, is one of the important things that can support marketing success, as conducted by Ahmat Arif Syaifudin and Tutik Al-fiyah in their research which discusses the implementation of integrated marketing communication in supporting marketing success in one of the companies in 2022 [8].

The data used as research material in this journal is secondary data originating from related companies. The data used is data that describes purchases made by consumers which are grouped into several groups based on the type of consumer itself, which will then be segmented again based on the consumption level of each consumer. The time span taken is for the past 2 years, to be precise in 2021-2022.

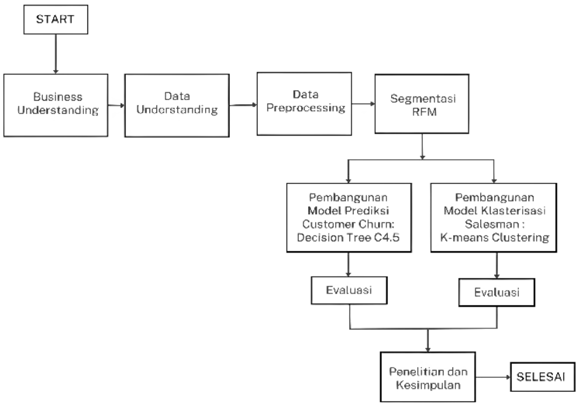

START
Business Understanding
Data Understanding
Data Preprocessing
Segmentasi RFM
Pembangunan Model Prediksi Customer Churn: Decision Tree C4.5
Pembangunan Model Klasterisasi Salesman : K-means Clustering
Evaluasi
Evaluasi
Penelitian dan Kesimpulan
SELESAI

Fig. 1. Research overview diagram

In analyzing the data, several sequential analysis techniques are used with the aim of interpreting the results of data processing that is easy to understand. The first thing to do is Business understanding. Business understanding is a stage of understanding related to the business to be studied [2]. Understanding covers business objectives, the actual situation in the field, and a general design related to the purpose of conducting research using existing data. The next stage is Data Understanding. Data Understanding is a stage where an understanding of the data that has been obtained is carried out. At this stage the researcher gets help from the data owner source to understand the data provided. The data obtained is daily consumer transaction data with various types of materials, located in the East Java, Bali and Nusa regions for 2 years. 113,270 data records for 2021 and 110,364 data records for 2022. For the third stage, Data Preprocessing is carried out. Which is a stage where data is sorted and prepared for the data mining process. All unused attributes will be cleaned up to leave attributes that are important for the data processing stage. In RFM modeling, both in the customer segment and the salesman segment, the attributes used are the customer group, salesman, calendar per day, amount of income, and description of materials purchased. Then next enter the RFM segmentation stage. RFM is a method of grouping consumer classes based on transaction loyalty. In this segmentation process, a score of 1 to 3 is given to each recency, frequency, and monetary domain in each customer group. Score 3 for the highest score and score 1 for the lowest score where the final score is determined by a combination of scores in each domain [3].

The recency score is taken from the distance between the last transaction and the current date. The customer group that has the closest transaction date to the present has the highest score, which is a score of 3. Also related to the frequency value is taken from the frequent number of transactions from each customer group. The customer group that has the most frequent number of transactions is given the highest score, which is score 3. As for the monetary value, it is taken from the total number of transactions. Related to the score range of each domain is obtained from processing raw data in excel with the help of the pivot data feature. And obtained the score range as shown in table 1.

TABLE I. RFM CLASS SCORE AND RANGE FOR CUSTOMER GROUP SEGMENT

Atribut	Nilai	Skor	Arti
Recency	o - ≤44658	1	Longest
	≤44659 - ≤44922	2	Somewhat Longer
	≤44923 - 44927	3	New
Frequency	o - ≤3550	1	Rare
	≤3551 - ≤14109	2	Medium
	≤14110 - 84794	3	Often
monetary	0 - 186520927985	1	low
	≤186520927986 - ≤1463821002601	2	Medium
	≤1463821002602 - 15370672220547	3	high

As for the scores for the salesman segment, there are only differences in the frequency and monetary domains, not for recency, because in the raw data, all salesmen have the last transaction data record on the same day. So the table for the salesman segment is as shown in table 2 below.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.2030, Copyright ©2024
Submitted : December 6, 2023, Revised : January 23, 2023, Accepted : January 26, 2023, 2024, Published : February 15, 2024

TABLE II. RFM SCORE AND CLASS RANGE FOR SALESMAN SEGMENT

Atribut	Nilai	Skor	Arti
Frequency	0 - 18124	1	Rare
	≤18125 - ≤37254	2	Medium
	≤37255 - 47947	3	Often
monetary	0 - 2515383469894	1	low
	≤2515383469895 - ≤3958875694211	2	Medium
	≤3958875694212 - 15472713667590	3	high

A. *Building a Customer Churn Prediction Model with Decision Tree C4.5*

Data mining on research on customer group segments using C4.5 decision tree modeling. The data processed in this modeling uses data that is categorical and not numerical [4]. After determining the score in each domain, the next step is determining the customer label [3]. Customer labels in this study are determined in several categories. The first is "Superstar", which is a category aimed at the customer class that has the highest score. Then there is the "Golden" category, which is a category intended for the class that has the highest monetary score after superstar, a high frueqency score, and has an average recency score. Next there is the "Typical" category which is a category intended for customer classes that have average scores in each domain. Next is "Occasional" which is a category assigned to the customer class that has the second lowest monetary score after dormant, but has a high score for frequency, and a low score for recency. Next is the "Everyday" category which is for the customer class that has an average recency score. The frequency score is low, and the monetary score is from medium to low. And the last is "Dormant" which is intended for the customer class that has the lowest mix of scores.

There are 27 customer classes seen from the combination of RFM scores. The highest score has a combination of 333 and the lowest score has a combination of 111. The score table with this category label will determine the average customer class owned by the company is loyal or not, and can be a determinant in the preparation of further strategies. The score table with category labels is shown in table 3 as follows.

TABLE III. SCORE TABLE WITH CATEGORY LABELS

Class	SCORE			Label Cust Group
	R	F	M	
K1	New	Rare	low	everyday a
K2	New	Rare	Medium	typical b
K3	New	Rare	high	golden c
K4	New	Rare	low	everyday a
K5	New	Some what often	Medium	superstar c
K6	New	Some what often	high	superstar b
K7	New	often	low	typical d
K8	New	often	Medium	golden a
K9	New	Rare	high	golden c
K10	New	Rare	low	everyday a
K11	New	Rare	Medium	typical b
K12	New	Rare	high	golden c
K13	New	Some what often	low	typical c
K14	New	Some what often	Medium	superstar c
K15	New	often	high	a superstar
K16	New	often	low	typical d
K17	New	Rare	Medium	typical b
K18	New	Rare	high	golden c
K19	New	Rare	low	everyday a
K20	Some what Longer	Rare	Medium	everyday c
K21	Some what Longer	Some what often	high	golden b
K22	Some what Longer	Some what often	low	everyday d
K23	Some what Longer	often	Medium	everyday e
K24	Longest	often	high	golden d
K25	Longest	Rare	low	a dormant
K26	Longest	Rare	Medium	dormant b
K27	Longest	Rare	high	golden e

Furthermore, this table will be processed into the RapidMiner application for C4.5 decision tree modeling, whose detailed process will be shown in the results and conclusions chapter.

B. *Building a Customer Churn Prediction Model with Decision Tree C4.5*

Modeling using the k-means clustering algorithm for the salesman segment is not much different from the customer group segment. Before finally entering k-means clustering modeling, the data must be processed first with RFM modeling to be able to highlight the characteristics required by each salesman. So from RFM modeling, it is known that the value possessed by each salesman from each domain is as follows as shown in table 4.

TABLE IV. RFM VALUE FOR EACH SALESMAN

Wilayah SBM	R	F	M
SBM Industry I	0.75890411	41503	8460422329347
SBM Industry II	0.75890411	20346	3958875694211
SBM Industry III	0.75890411	47947	2515383469894
SBM Industry IV	0.75890411	18124	1305280048298
SBM Industry V	0.75890411	37254	2620658189947
SBM Industry VI	0.75890411	33424	15472713667590

SBM Industry VII	0.75890411	12190	2411370220183
SBM Industry VIII	0.75890411	12846	2642144188420

From this table, it can be seen that the recency domain has the same value, because from the raw data itself, each salesman has the same track record for dates, and there is no accompanying hour information. After obtaining the value of each domain for each salesman, the next is data processing in RapidMiner with the k-means clustering algorithm to determine the cluster of each data.

To determine the number of clusters for determining the centroid point, researchers use the Davies Bouldin method which looks at the proximity distance between the data and the center point of each cluster from the predetermined center. This method has an internal scheme test, where the suitability of the number of clusters is seen through comparing the Davies Bouldin value between the number of clusters. This method was first disseminated by David L. Davies and Donald W. Bouldin in 1979 [7]. As for the salesman data in this study, the Davies Bouldin method is explained in the following table.

TABLE V. DAVIES BOULDIN VALUE FOR EACH CLUSTER

Davies Bouldin	
k	**db**
2	1.549
3	2.900
4	3.592
5	3.829

It can be seen from the table above, the smallest closeness value is owned by cluster 2 which is worth 1.549, which means that the right cluster division for salesman data is 2 clusters. Henceforth, k-means clustering processing on RapidMiner is divided into 2 clusters, and will be explained in more detail in the results and conclusions chapter.

Clustering of salesman data is done to distinguish between groups of salesmen who have good performance and groups of salesmen who have poor performance. If it has been clustered, then it will be easier for the author to decide the answer to the research question "is the group of customers who have the potential to churn caused by the lack of salesman skills?".

IV. RESULT AND DISCUSSION

A. Test Results on Customer Group Data

Testing on customer group data is done in 2 ways, namely with RFM modeling and classification using the C4.5 decision tree algorithm. RFM modeling, in addition to being used for data preprocessing before modeling using a decision tree, RFM modeling that includes scoring can also be used to designate 1 customer group that has the highest percentage of churn. Meanwhile, modeling through the C4.5 decision tree algorithm can be used to identify the character of the BBM industry customer group at PT Pertamina Patra Niaga Regional Jatimbalinus as a whole, in order to get the conclusion that the majority of customers owned by the company have a high level of loyalty or not, as well as one of the determinants of whether the current strategy is appropriate.

As stated in chapter II, RFM modeling is used one of them to prepare data which is then processed in decision tree modeling. RFM modeling can also be used to determine which customer group has the highest percentage to churn.

TABLE VI. RFM LABEL TABLE FOR CUSTOMER GROUP DATA

Customer group	Max of Calendar Day	Frequency	Monetary	Label
PLN	44926	84794	15,370,672,220,547	Superstar
Industri	44926	23419	11,069,681,813,245	Superstar
Agen BBM Industri	44926	37068	4,167,413,285,932	Superstar
Angkutan Laut	44926	22285	3,973,001,549,093	Superstar
TNI - AL (Navy)	44926	12003	1,463,821,002,601	Golden
Agent Bunker	44923	2833	1,086,402,885,241	Typical
POLRI	44926	13242	678,250,613,276	Golden
TNI - AD (Army)	44922	8532	579,156,467,053	Golden
Transportasi Darat	44926	14109	556,924,549,043	Golden
TNI - AU (Airforce)	44923	3550	186,520,927,985	Typical
Instansi Pemerintah	44926	1169	141,688,074,946	Everyday
BU-PIUNU	44658	451	87,070,804,854	Everyday
Perikanan	44848	97	13,986,458,975	Everyday
Layanan Umum	44463	82	12,257,155,099	Dormant

The table above shows the RFM value for each customer group, as well as the label determined based on the value range of each customer group. So it can be seen, the customer group that has great potential in churning is the customer group in the "general service" segment, where the RFM value of this customer group is the lowest compared to other customer groups.

The selection for modeling with the decision tree algorithm, taken from the test results using stratified random sampling technique. This technique is a process that starts from a dataset that is divided into traing data and testing data, which then makes data distribution for each part. The comparison is divided into three comparisons, namely 80:20, 70:30, and 60:40. After the data is successfully split, it will be tested with each different model, then evaluated with the performance operator to see the accuracy of each test result.

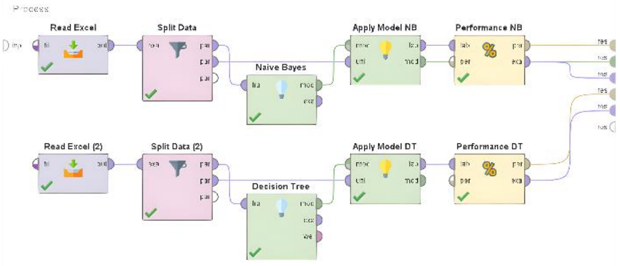


Fig. 2. Comparison of 2 algorithms

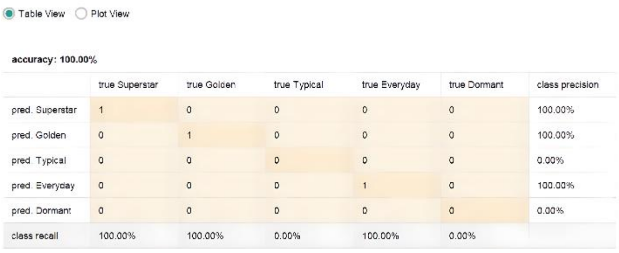
So, the results of the tests on each model with different comparisons are shown in the following table:

TABLE VII. COMPARISONS OF TWO ALGORITHM

Comparison	accuracy	
	Decision Tree	Naïve Bayes
80:20	100.00%	66.67%
70:30	75.00%	75.00%
60:40	83.33%	83.33%

It can be seen that the highest value is owned by the decision tree model with a ratio of 80:20. The accuracy rate of decision tree is very high up to 100% which is inversely proportional to naïve bayes which is only 66.67%. Therefore, decision tree modeling with a ratio of training data and tasting data of 80:20 will be used in this test.

Data processing begins with RFM modeling which then produces 5 class categories, namely dormant, everyday, typical, golden, and superstar. Which is then processed using the C4.5 decision tree algorithm in RapidMiner, and the following results are obtained:



● Table View ○ Plot View

accuracy: 100.00%

	true Superstar	true Golden	true Typical	true Everyday	true Dormant	class precision
pred. Superstar	1	0	0	0	0	100.00%
pred. Golden	0	1	0	0	0	100.00%
pred. Typical	0	0	0	0	0	0.00%
pred. Everyday	0	0	0	1	0	100.00%
pred. Dormant	0	0	0	0	0	0.00%
class recall	100.00%	100.00%	0.00%	100.00%	0.00%	

Fig. 3. PerformanceVector Classification Table for Customer Group Data

The test results are shown in the table above. The table is a table generated from the performance classification operator. This operator is useful as a validity tester of the test data. The parameters tested consist of accuracy, precision, and recall. And the results obtained from this table where the total accuracy level is 100.00%. Which means that this test has a high level of validity. The calculations in this table use confusion matrix comparisons for datasets processed using the decision tree method.

From data processing using decision tree, the following decision tree results can be seen.

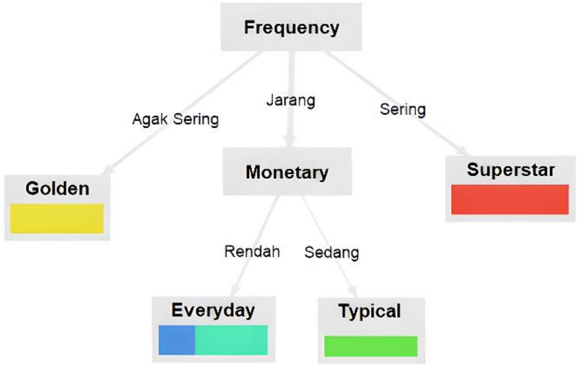


Fig. 4. Class Category Decision Tree from Customer Group Data

The decision tree shows a factor that has the most influence on the top branch, and it can be seen that the category grouping is a factor of frequency or in other words the number of frequency of purchases, which is then followed by the monetary factor. And from this decision tree, it can be concluded that the majority of PT Pertamina Patra Niaga regional Jatimbalinus customer groups in the Fuel Industry segment fall into the superstar and golden categories, which means that there is little chance for these customer groups to churn. In other words, it can be said that when viewed through this research method, the strategy implemented in the last 2 years has been included good, judging from the majority of customers who have high loyalty.

B. Test Results on Salesman Data

Salesman data testing is tested through 2 methods, through RFM modeling and through k-means clustering algorithm modeling. Through the RFM method, transaction data that has been carried out by salesmen is processed in such a way that it is then modeled in RapidMiner to extract the essence of the data processing results. In this study, researchers used k-means clustering to find out the lowest salesman performance among the salesmen that the company has, which will then be investigated in relation to the purchasing power of customer groups in the same segment.

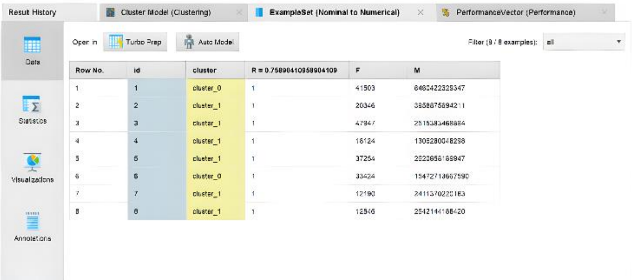
Tests carried out on salesman data using the RFM method, produce the following output.

TABLE VIII. TEST RESULTS USING RFM MODELING ON SALESMAN DATA

Wilayah SBM	R	F	M
SBM Industry I	0.75890411	41503	8460422329347
SBM Industry II	0.75890411	20346	3958875694211
SBM Industry III	0.75890411	47947	2515383469894
SBM Industry IV	0.75890411	18124	1305280048298
SBM Industry V	0.75890411	37254	2620658189947
SBM Industry VI	0.75890411	33424	15472713667590
SBM Industry VII	0.75890411	12190	2411370220183
SBM Industry VIII	0.75890411	12846	2642144188420

The test results of hundreds of thousands of salesman transaction data summarized according to each salesman are shown in the table. It is found from the test results, there are 8 salesmen who each have RFM scores according to the transaction traces they have made in the last 2 years. Furthermore, the results of this test will be processed in RapidMiner to be able to retrieve test results according to the objectives of this study.

Salesman data test results with RFM modeling are processed in RapidMiner using the k-menas clustering algorithm. In this algorithm, each salesman will be grouped based on the similarity of the characteristics of each data with a predetermined centroid point. To determine the number of centroid points, researchers used the davies bouldin method as described in chapter II. So from this method, it is found that the division of the number of centroid points for the number of clusters for salesman data is 2 points, which then these 2 points become the number of clusters that will be called clusters with 'good performance', and clusters with 'poor performance'. After processing with RapidMiner, the following modeling results were obtained.



Row No.	id	cluster	R = 0.75890410958904109	F	M
1	1	cluster_0	1	41503	6480422328347
2	2	cluster_1	1	20346	3858875894211
3	3	cluster_1	1	47947	2515383465884
4	4	cluster_1	1	16124	1306280046298
5	5	cluster_1	1	37254	2820656189947
6	6	cluster_0	1	33424	15472713667590
7	7	cluster_1	1	12190	2411370220183
8	8	cluster_1	1	12546	2542144188420

Fig. 5. Modeling Results of K-means Clustering on Salesman Data

A new yellow column labeled cluster_0 and cluster_1 appears between the existing columns. This column shows the grouping of data according to the closeness of their respective characteristics. So that it can be seen which salesmen are included in cluster 0 and which salesmen are included in cluster 1. Then from the table, new knowledge is obtained regarding the cluster, with details of cluster 0 containing salesmen with good performance categories, which are characterized by their RFM values that tend to be higher. Meanwhile, cluster 1 contains salesmen with poor performance categories, which are characterized by their RFM values which tend to be lower. And from the table, it can be seen that salesmen IV and VII have the least good performance among others, which is characterized by the lowest RFM value among other salesmen.

C. *Identification of the Relationship between the Results of the Two Types of Data*

Thus far, the research has revealed specific findings regarding the customer group data studied to identify potential churn customers and the salesmen data studied to determine the underperforming salesmen. It was found that the customer group with a high potential for churn is the 'general service' category, and the salesmen with the least satisfactory performance are 'salesman IV and VII'. Therefore, to address the research question 'whether the potential churn customers are caused by the lack of salesmen's skill', further investigation will be conducted based on the relevant data from these two results.

Starting with the exploration of the 'general service' customer group data, an anomaly was found. The entire 'general service' customer group was served by salesman I, who belongs to cluster 0 (cluster with good performance). Another anomaly was found when focusing on the transaction dates. Upon examining the raw data, it was noticed that the 'general service' customer group made their last purchase on September 24, 2021, and there were no further transaction records thereafter. This result raised a question as to why the transactions ceased after September 24, 2021, despite consistent monthly transactions before that date.

Continuing with the exploration of the data for salesman IV and salesman VII, who have the least satisfactory performance among other salesmen, the analysis of the raw data revealed that the majority of customers served by salesman IV and salesman VII belong to the small potential churn customer group.

V. Conclusion

The results show that analyzing customer data using a decision tree algorithm with a ratio of 80:20 shows that the majority of the company's customers are loyal customers. This finding is validated by the 100% accuracy rate on the confusion matrix. In addition, a specific group of customers in the "General Services" segment was identified as having high churn potential, as indicated by the lowest RFM value and this group was categorized as inactive. Furthermore, this study shows that optimal clustering divides salespeople into two groups: 'good' and 'bad' performance groups. This division is based on the Davies Bouldin method, which shows the smallest value at 1.549. An important insight from linking customer transaction data and salesperson performance data is that high churn rates in certain customer groups are not necessarily related to salesperson performance. This is exemplified in the "General Service" segment, where customers with high churn potential are served by salesperson I, who is categorized in the 'good' performance cluster. Positive comments about consistent service quality, product satisfaction, and brand trust can support the high level of loyalty indicated by the decision tree analysis. This study has limitations due to the researcher's indirect involvement with the company's operational environment. For a more comprehensive understanding, further information gathering is recommended, such as conducting in-depth interviews with relevant stakeholders or collecting broader data from various accredited sources.

Acknowledgment

The researcher would like to express gratitude to all parties who have assisted in this research, enabling it to be published as it should be. Special thanks to PT Pertamina Patra Niaga in the Jatimbalinus region for providing all the necessary information to ensure the smooth progress of this research. Also, the researcher extends appreciation to the university for facilitating the research, allowing it to be carried out to the best of its ability.

References

[1] N. D. Japari, A. Zafrullah TN, and F. R. Djoemadi, "The Role of Pt. Pertamina as a fuel oil supply provider in Indonesia," Calyptra, vol. 7, no. 2, pp. 1-12, 2019

[2] R. Muliono and Z. Sembiring, "Data Mining Clustering Using K-Means

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.2030, Copyright ©2024
Submitted : December 6, 2023, Revised : January 23, 2023, Accepted : January 26, 2023, 2024, Published : February 15, 2024

Algorithm for Clustering Lecturer Teaching Tridarma Level," CESS (Journal Comput. Eng. Syst. Sci., vol. 4, no. 2, pp. 2502-714, 2019.

[3] E. Muningsih, I. Maryani, and V. R. Handayani, "Application of K-Means Method and Optimization of Number of Clusters with Davies Bouldin Index for Clustering Provinces Based on Village Potential," J. Sains dan Manaj., vol. 9, no. 1, pp. 95-100, 2021, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/view/10428/4839.

[4] A. A. Syaifuddin and T. Al-Fiyah, "Implementation of Integrated Marketing Communication in Supporting Marketing and Brand Equity of BMT Mandiri Artha Sejahtera East Java," Juornal Econ. Policy ..., vol. 03, no. 02, pp. 37-51, 2022, [Online]. Available: https://ejournal.uinsatu.ac.id/index.php/jesk/article/view/7126%0Ahttps://ejournal.uinsatu.ac.id/index.php/jesk/article/download/7126/2115.

[5] F. Soufitri, E. Purwawijaya, E. H. Hasibuan, and R. N. Singarimbun, "Testing C4.5 Algorithm Using Rapid Miner Applications in Determining Customer Satisfaction Levels," J. Infokum, vol. 9, no. 2, pp. 510-517, 2021, [Online]. Available: http://infor.seaninstitute.org/index.php/infokum/index.

[6] Y. Primawati, I. Verdian, and G. W. Nurcahyo, "K- Means Clustering on Based Classification Method of Sales Agent," J. Comput. Since Inf. Technol., vol. 7, no. 2, pp. 1-6, 2021, doi: 10.35134/jcsitech.v7i2.1.

[7] P. Algorithm, C. Using, and A. Rapid, "In determining the level of customer satisfaction," vol. 0, no. 2, pp. 510-517, 2021.

[8] M. A. Kadafi, "Evaluate the potential bankruptcy of Indonesian oil and gas mining companies in the 2013-2015 period," Forum Ekon., vol. 21, no. 2, pp. 154-164, 2019, [Online]. Available: http://journal.feb.unmul.ac.id/index.php/FORUMEKONOMI

[9] R. Novendri, R. Andreswari, and ..., "Implementation of Data Mining to Predict Customer Churn Using Naive Bayes Algorithm," eProceedings ..., vol. 8, no. 2, pp. 2762-2773, 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/download/146 78/14455 sdfdsfdsf.

[10] N. L. Widyastuti and H. Nugroho, "The Impact of Covid-19 on the Oil and Gas Industry:
Policy Recommendations for Indonesia," J. Perenc. Pembang. Indonesia. J. Dev. Plan., vol. 4, no. 2, pp. 166-176, 2020, doi: 10.36574/jpp.v4i2.116dfdsfdsfdsf.

[11] E. K. Tarigan, "Juridical analysis of retail fuel oil sales according to the oil and gas law (law number 22 of 2001)," J. Lex Justitia, vol. 2, no. 2, pp. 121-134, 2021, [Online]. Available: http://ejournal.potensi-utama.ac.id/ojs/index.php/LexJustitia/article/view/1347.

[12] A. Munawar et al., "Cluster Application with K-Means Algorithm on the Population of Trade and Accommodation Facilities in Indonesia," J. Phys. Conf. Ser., vol. 1933, no. 1, 2021, doi: 10.1088/1742-6596/1933/1/012027S.

[13] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto,
"Application of Data Mining Technique using K-Medoids in the Case of Export of Crude Petroleum Materials to the Destination Country," IOP Conf. Ser. Mater. Sci. Eng., vol. 835, no. 1, 2020, doi: 10.1088/1757-899X/835/1/012058.

[14] J. R. Riwukore, L. Marnisah, and F. Habaora, “Employee Performance Analysis Based on the Effect of Discipline, Motivation, and Organizational Commitment at the Regional Secretariat of the Kupang City Government,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 12, no. 1, p. 76, 2022, doi: 10.30588/jmp.v12i1.1009.

[15] P. A. Sampurna and T. Miranti, “The Effect of Service Quality, Banking Digitalization, and Customer Relationship Management (CRM) on Customer Loyalty,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 12, no. 1, p. 303, 2022, doi: 10.30588/jmp.v12i1.1138

[16] J. F. Sofyan and M. Rianty, “Karakteristik Manajemen dan Kepemimpinan Transformasional sebagai Penentu Kreativitas Karyawan yang Dimediasi oleh Kepuasaan Kerja,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 12, no. 2, p. 448, 2023, doi: 10.30588/jmp.v12i2.1186.

[17] F. P. Ferdy Pangestu, N. Y. Nur Yasin, R. C. Ronald Chistover Hasugian, and Y. Yunita, “Penerapan Algoritma K-Means Untuk Mengklasifikasi Data Obat,” J. Sisfokom (Sistem Inf. dan Komputer), vol. 12, no. 1, pp. 53–62, 2023, doi: 10.32736/sisfokom.v12i1.1461.

[18] A. A. Syahidah and M. F. Aransyah, “Pengaruh E-Service Quality dan E-Trust Terhadap E-Customer Loyalty Pada Pengguna Dompet Digital DANA Melalui E-Satisfaction Sebagai Variabel Intervening,” J. Sisfokom (Sistem Inf. dan Komputer), vol. 12, no. 1, pp. 36–44, 2023, doi: 10.32736/sisfokom.v12i1.1593.

[19] A. Rosyida and T. Bayu Sasongko, “Deteksi Dini Penyakit Alzheimer dengan Algoritma C4.5 Berbasis BPSO (Binary Particle Swarm Optimization),” J. SISFOKOM (Sistem Inf. dan Komputer), vol. 12, pp. 341–349, 2023.

[20] D. E. Sondakh, R. C. Maringka, F. P. Ayorbaba, J. S. C. B. T. Mangi, and S. R. Pungus, “Emotion Mining Review Pengguna Aplikasi Mobile Banking BRImo Menggunakan Algoritma Decision Tree,” vol. 12, pp. 350–355, 2023.

[21] F. Warda, F. N. Fajri, and A. Tholib, “Classification of Final Project Titles Using Bidirectional Long Short Term Memory at the Faculty of Engineering Nurul Jadid University,” J. Sisfokom (Sistem Inf. dan Komputer), vol. 12, no. 3, pp. 356–362, 2023, doi: 10.32736/sisfokom.v12i3.1723.

[22] R. Vannya and A. Hermawan, “Analisis Performa klasifikasi Kesegaran Daging Ayam menggunakan Naïve Bayes , Decision Tree , dan,” vol. 12, pp. 394–400, 2023

[23] M. W. Sari, N. N. Deswira, and A. Risdwiyanto, “Determinasi Kepuasan Kerja dan Implikasinya terhadap Turnover Intention: Studi pada PT Hayati Pratama Mandiri Padang,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 12, no. 2, p. 346, 2023, doi: 10.30588/jmp.v12i2.430.

[24] E. Fatmasari and B. S. Dwiyanto, “Analisis Kinerja Keuangan dengan Metode Economic Value-Added pada Studi Kasus Perusahaan Subsektor Pertambangan Minyak dan Gas Bumi yang Terdaftar di Indeks Saham Syariah Indonesia (ISSI),” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 9, no. 1, p. 17, 2019, doi: 10.30588/jmp.v9i1.435.

[25] K. S. Utami, “Green consumers behavior: Consumer behavior in purchasing environmentally friendly products,” Manag. Coop. Entrep., vol. 9, no. 2, pp. 208–223, 2020, [Online]. Available: http://ejournal.up45.ac.id/index.php/maksipreneur/article/download/499/526

[26] R. B. Prakarsa, W. Yadiati, and N. R. H. Suciati, “Pengaruh Risk Profile, Good Corporate Governance, Earning, Capital terhadap Value of Firm di Bursa Efek Indonesia,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 9, no. 2, p. 137, 2020, doi: 10.30588/jmp.v9i2.530.

[27] I. L. Natapermana, W. Yadiati, and E. Nurhayati, “Pengaruh Implementasi Good Corporate Governance dan Strategi Bisnis terhadap Kinerja Perusahaan: Studi Kasus BUMN di Indonesia Tahun 2013-2018,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 9, no. 2, p. 153, 2020, doi: 10.30588/jmp.v9i2.579.

[28] A. Junaidi, “Sisfokom-Maret2019-604-1539-1-PB,” vol. 08, pp. 61–67, 2019.

[29] W. Marantika and S. Sarsono, “Pengaruh Kualitas Produk, Word of Mouth, dan Store Image terhadap Keputusan Pembelian: Studi pada Pengunjung Toko Amigo Pedan,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 10, no. 1, p. 114, 2020, doi: 10.30588/jmp.v10i1.633.

[30] A. Khomsiyah and S. Sanaji, “Pengaruh Loyalitas dan Fanatisme Supporter pada Klub terhadap Keputusan Pembelian Merchandise Orisinal: Studi pada Supporter Persela Lamongan,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 10, no. 2, p. 242, 2021, doi: 10.30588/jmp.v10i2.756

[31] D. M. D. Prama Yanti and I. G. Sanica, “Menelisik Pengelolaan Human Capital di Dunia Bisnis dalam Era New Normal: Studi Kasus pada Generasi Milenial di Bali,” J. Maksipreneur Manajemen, Koperasi, dan Entrep., vol. 11, no. 1, p. 122, 2021, doi: 10.30588/jmp.v11i1.840.

[32] M. A. S. Arifin, H. Oktafia, and L. Wijaya, “Deteksi Botnet IoT Menggunakan Autoencoder dan Decision Tree,” vol. 12, pp. 329–334, 2023.