Application of the K-Means Clustering Agorithm to Group Train Passengers in Labuhanbatu

Indri Cahaya Indah[1)*], **Mila Nirmala Sari**[2)], **Muhammad Halmi Dar**[3)]
[1,2,3)]Universitas Labuhanbatu,
[1)]icahaya035@gmail.com, [2)]milanirmalasari7@gmail.com, [3)]mhd.halmidar@gmail.com

Submitted : Mar 15, 2023 | **Accepted** : Mar 29, 2023 | **Published** : Apr 1, 2023

Abstract: Transportation is an activity of moving things such as humans, animals, plants and goods from one place to another. To be able to implement transportation, we need a means of transportation that suits our needs. For in Indonesia, people are more inclined to land transportation. That's because land transportation already has a lot of vehicles. Land transportation already has many vehicles that can be used, both for private and for the public. Each vehicle has its uses and risks as well. Therefore we will do a data cluster from the trains. We chose the train, because the risk from using the train is very small, meaning that there is a lot of public interest in trains. So we want to do a cluster on rail passengers. The cluster that we do is to group passenger data based on the similarity of passenger data. We will do the cluster using the K-Means method. The K-Means method is very suitable when used to perform a cluster. K-Means will process widgets that are made according to the needs of the research. So after we enter the method in the widget pattern, the widget will process it to output the results from the cluster that we created. The cluster process using the K-Means method will be applied using the orange application. After we apply it, the data will later be clustered, we will cluster data as many as 3 clusters. Then the incoming data will appear in clusters 1, 2 and 3, both from business and executive classes.

Keywords: Confusion Matrix, Data Mining, K-Means, Orange, Roc Analysis,

INTRODUCTION

Transportation is the process of physically moving people or goods from one place to another within a certain time by using or being driven by humans, animals or machines. In the process of moving, we need a tool that can be used to move goods or something else from one place to another. There are several types of transportation that can be used, transportation equipment used on land, at sea and in the air. The means of transportation used on land include buses, trains, cars, motorbikes and many others. For sea transportation there are ships, boats, and for air transportation there are planes. Each means of transportation has its own functions and uses. In terms of choosing a means of transportation, usually prospective passengers see the level of risk that will occur. Judging from the level of risk, air and sea transportation has a big risk. Meanwhile, land transportation has a small risk, especially like a train, which has a very small risk. That is because the trains run according to the rails or their paths. Therefore, many people are interested in taking the train. So we are interested in land transportation, namely trains. Therefore we want to cluster train passenger data, both business and executive class. PT. Indonesian Railroad (KAI) is one of the land transportation services that has been operating throughout Indonesia.(Adawia, Azizah, Endriastuty, & Sugandhi, 2020)

*name of corresponding author

The means of transportation that are often used by passengers is ground transportation. In this study, we will conduct a data cluster. A cluster will store data based on the similarity of the data.(Al-Ars & Al-Bakry, 2019) The data that we will cluster is train passenger data for January 2023. We will cluster passenger data from business and executive classes. In the cluster process, we will need a method that can process data for the cluster. Previously, clustering was a process of grouping data based on the similarity of a data. The cluster system will be able to minimize energy and be able to group data efficiently.(Hassen, Lafta, Noman, & Ali, 2019) A cluster can be characterized by looking at the similarities and similarities of the group attributes and dissimilarity can also be seen from the group attributes.(Hamzaoui, Amnai, Choukri, & Fakhri, 2020)

Data mining is a process of collecting and processing data with the aim of extracting important information in large data.(Bui et al., 2020)(Ghaedi, Farizani, & Ghaemi, 2021)(Uçar & Karahoca, 2021)(Dirjen et al., 2018) By doing data mining, we can get an efficient understanding of multi-view sentiment textual data.(Yassir et al., 2020) The data mining that we are doing is we will do a cluster on train passenger data. To do this data mining, of course we will need a method that is suitable and can be used to cluster train passenger data. In this study we will use the K-Means method to carry out a cluster of train passenger data.

METHOD

The K-Means method is an unsupervised learning method for defining cluster centers and grouping data based on the same data.(Riza, Rosdiyana, Wahyudin, & Pérez, 2021) So any data that has similarities and similarities will be clustered based on their respective groups.(Widiyanto & Witanti, 2021) This method is a kernel-based method as a non-probabilistic technique used to group data.(Rustam, Hartini, Pratama, Yunus, & Hidayat, 2020)

Confusion Matrix

The confusion matrix is an easy and effective tool to use to show the performance of a Classification and is very easy to use to determine the results.(Yun, 2021) Confusion matrix can be used to evaluate the work results of a model and can be used to determine the results of a data mining. The confusion matrix has several calculations, namely as follows.

Table 1. Confusion Matrix

Confusion Matrix		True Class (Actual)	
		P	N
Predicted class	Y	True Positive (TP)	False Positive (FP)
	N	False Negative (FN)	True Negative (TN)

To determine the calculation of the confusion matrix, we can do it by calculating accuracy, precision and recall.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$ (Yun, 2021)

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\%$$
(Normawati & Prayogi, 2021)

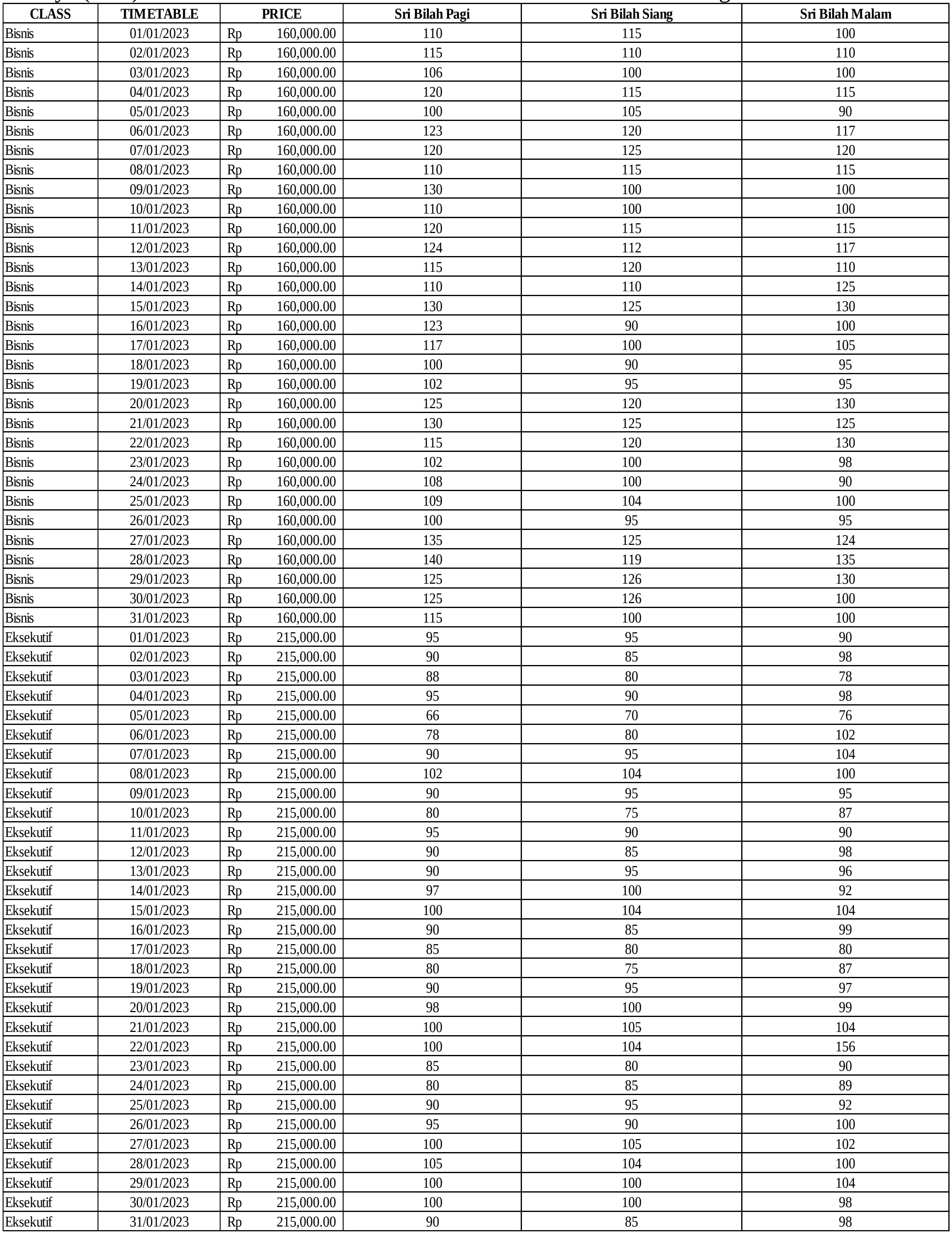
$$\text{Accuracy} = \frac{TP}{TP+FN} \times 100\%$$
(Agustina, Adrian, & Hermawati, 2021)

*name of corresponding author

RESULT

Analysis Data

In the picture below is the cumulative data on the number of train passengers every day, from the morning Sri Bar, afternoon Sri Bar and Evening Sri Bar trains. I got the data from the Indonesian Railways (KAI) office in Medan. We will cluster the data into 3 clusters using the K-Means method.



CLASS	TIMETABLE	PRICE	Sri Bilah Pagi	Sri Bilah Siang	Sri Bilah Malam
Bisnis	01/01/2023	Rp 160,000.00	110	115	100
Bisnis	02/01/2023	Rp 160,000.00	115	110	110
Bisnis	03/01/2023	Rp 160,000.00	106	100	100
Bisnis	04/01/2023	Rp 160,000.00	120	115	115
Bisnis	05/01/2023	Rp 160,000.00	100	105	90
Bisnis	06/01/2023	Rp 160,000.00	123	120	117
Bisnis	07/01/2023	Rp 160,000.00	120	125	120
Bisnis	08/01/2023	Rp 160,000.00	110	115	115
Bisnis	09/01/2023	Rp 160,000.00	130	100	100
Bisnis	10/01/2023	Rp 160,000.00	110	100	100
Bisnis	11/01/2023	Rp 160,000.00	120	115	115
Bisnis	12/01/2023	Rp 160,000.00	124	112	117
Bisnis	13/01/2023	Rp 160,000.00	115	120	110
Bisnis	14/01/2023	Rp 160,000.00	110	110	125
Bisnis	15/01/2023	Rp 160,000.00	130	125	130
Bisnis	16/01/2023	Rp 160,000.00	123	90	100
Bisnis	17/01/2023	Rp 160,000.00	117	100	105
Bisnis	18/01/2023	Rp 160,000.00	100	90	95
Bisnis	19/01/2023	Rp 160,000.00	102	95	95
Bisnis	20/01/2023	Rp 160,000.00	125	120	130
Bisnis	21/01/2023	Rp 160,000.00	130	125	125
Bisnis	22/01/2023	Rp 160,000.00	115	120	130
Bisnis	23/01/2023	Rp 160,000.00	102	100	98
Bisnis	24/01/2023	Rp 160,000.00	108	100	90
Bisnis	25/01/2023	Rp 160,000.00	109	104	100
Bisnis	26/01/2023	Rp 160,000.00	100	95	95
Bisnis	27/01/2023	Rp 160,000.00	135	125	124
Bisnis	28/01/2023	Rp 160,000.00	140	119	135
Bisnis	29/01/2023	Rp 160,000.00	125	126	130
Bisnis	30/01/2023	Rp 160,000.00	125	126	100
Bisnis	31/01/2023	Rp 160,000.00	115	100	100
Eksekutif	01/01/2023	Rp 215,000.00	95	95	90
Eksekutif	02/01/2023	Rp 215,000.00	90	85	98
Eksekutif	03/01/2023	Rp 215,000.00	88	80	78
Eksekutif	04/01/2023	Rp 215,000.00	95	90	98
Eksekutif	05/01/2023	Rp 215,000.00	66	70	76
Eksekutif	06/01/2023	Rp 215,000.00	78	80	102
Eksekutif	07/01/2023	Rp 215,000.00	90	95	104
Eksekutif	08/01/2023	Rp 215,000.00	102	104	100
Eksekutif	09/01/2023	Rp 215,000.00	90	95	95
Eksekutif	10/01/2023	Rp 215,000.00	80	75	87
Eksekutif	11/01/2023	Rp 215,000.00	95	90	90
Eksekutif	12/01/2023	Rp 215,000.00	90	85	98
Eksekutif	13/01/2023	Rp 215,000.00	90	95	96
Eksekutif	14/01/2023	Rp 215,000.00	97	100	92
Eksekutif	15/01/2023	Rp 215,000.00	100	104	104
Eksekutif	16/01/2023	Rp 215,000.00	90	85	99
Eksekutif	17/01/2023	Rp 215,000.00	85	80	80
Eksekutif	18/01/2023	Rp 215,000.00	80	75	87
Eksekutif	19/01/2023	Rp 215,000.00	90	95	97
Eksekutif	20/01/2023	Rp 215,000.00	98	100	99
Eksekutif	21/01/2023	Rp 215,000.00	100	105	104
Eksekutif	22/01/2023	Rp 215,000.00	100	104	156
Eksekutif	23/01/2023	Rp 215,000.00	85	80	90
Eksekutif	24/01/2023	Rp 215,000.00	80	85	89
Eksekutif	25/01/2023	Rp 215,000.00	90	95	92
Eksekutif	26/01/2023	Rp 215,000.00	95	90	100
Eksekutif	27/01/2023	Rp 215,000.00	100	105	102
Eksekutif	28/01/2023	Rp 215,000.00	105	104	100
Eksekutif	29/01/2023	Rp 215,000.00	100	100	104
Eksekutif	30/01/2023	Rp 215,000.00	100	100	98
Eksekutif	31/01/2023	Rp 215,000.00	90	85	98

Figure 1. *Passenger Data*

*name of corresponding author

In Figure 1, the data table above is cumulative data on the number of passengers each day. The data contains several attributes that are used to cluster data mining. These attributes are class, schedule, price, sri bilah pagi, sri bilah siang, dan sri bilah malam. As for the Sri attribute, the sri bilah pagi until sri bilah malam contains the number of passengers each day.

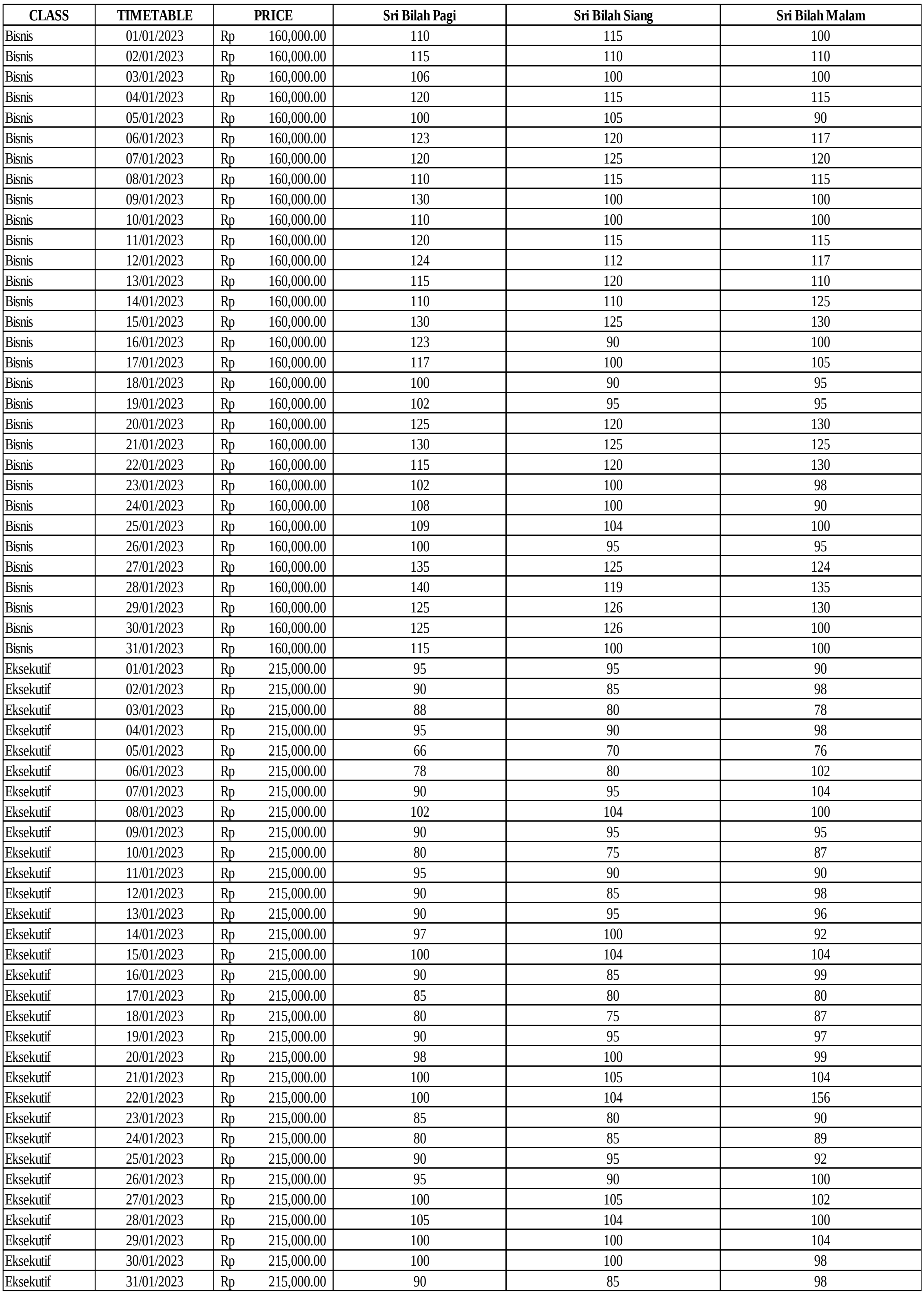
TABLE 2
PASSENGER DATA ATTRIBUTES

No	Atribut	Type	Deskripsi
1	Class	Category	The type of train used
2	Timetable	DateTime	Train departure time
3	Price	Numeric	Cost required
4	Sri Bilah Pagi	Numeric	The number of passengers on the morning train
5	Sri Bilah Siang	Numeric	The number of passengers on the afternoon train
6	Sri Bilah Malam	Numeric	Night train passengers

In the attribute table. The attribute of this research is the data we obtained from the Indonesian Railways (KAI) office in Medan. The attribute data is equipped with the type and description of each attribute.

Data Training

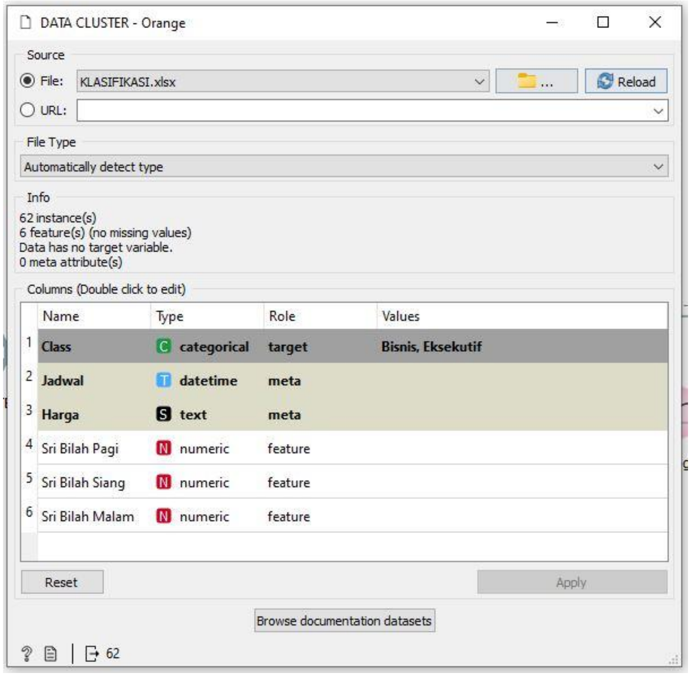
Training data is data that will be used as a sample in this study. The data we have obtained has the file.xlsx format. Then we arrange the data according to the needs of the k-means method so that we can cluster the data.



CLASS	TIMETABLE	PRICE		Sri Bilah Pagi	Sri Bilah Siang	Sri Bilah Malam
Bisnis	01/01/2023	Rp	160,000.00	110	115	100
Bisnis	02/01/2023	Rp	160,000.00	115	110	110
Bisnis	03/01/2023	Rp	160,000.00	106	100	100
Bisnis	04/01/2023	Rp	160,000.00	120	115	115
Bisnis	05/01/2023	Rp	160,000.00	100	105	90
Bisnis	06/01/2023	Rp	160,000.00	123	120	117
Bisnis	07/01/2023	Rp	160,000.00	120	125	120
Bisnis	08/01/2023	Rp	160,000.00	110	115	115
Bisnis	09/01/2023	Rp	160,000.00	130	100	100
Bisnis	10/01/2023	Rp	160,000.00	110	100	100
Bisnis	11/01/2023	Rp	160,000.00	120	115	115
Bisnis	12/01/2023	Rp	160,000.00	124	112	117
Bisnis	13/01/2023	Rp	160,000.00	115	120	110
Bisnis	14/01/2023	Rp	160,000.00	110	110	125
Bisnis	15/01/2023	Rp	160,000.00	130	125	130
Bisnis	16/01/2023	Rp	160,000.00	123	90	100
Bisnis	17/01/2023	Rp	160,000.00	117	100	105
Bisnis	18/01/2023	Rp	160,000.00	100	90	95
Bisnis	19/01/2023	Rp	160,000.00	102	95	95
Bisnis	20/01/2023	Rp	160,000.00	125	120	130
Bisnis	21/01/2023	Rp	160,000.00	130	125	125
Bisnis	22/01/2023	Rp	160,000.00	115	120	130
Bisnis	23/01/2023	Rp	160,000.00	102	100	98
Bisnis	24/01/2023	Rp	160,000.00	108	100	90
Bisnis	25/01/2023	Rp	160,000.00	109	104	100
Bisnis	26/01/2023	Rp	160,000.00	100	95	95
Bisnis	27/01/2023	Rp	160,000.00	135	125	124
Bisnis	28/01/2023	Rp	160,000.00	140	119	135
Bisnis	29/01/2023	Rp	160,000.00	125	126	130
Bisnis	30/01/2023	Rp	160,000.00	125	126	100
Bisnis	31/01/2023	Rp	160,000.00	115	100	100
Eksekutif	01/01/2023	Rp	215,000.00	95	95	90
Eksekutif	02/01/2023	Rp	215,000.00	90	85	98
Eksekutif	03/01/2023	Rp	215,000.00	88	80	78
Eksekutif	04/01/2023	Rp	215,000.00	95	90	98
Eksekutif	05/01/2023	Rp	215,000.00	66	70	76
Eksekutif	06/01/2023	Rp	215,000.00	78	80	102
Eksekutif	07/01/2023	Rp	215,000.00	90	95	104
Eksekutif	08/01/2023	Rp	215,000.00	102	104	100
Eksekutif	09/01/2023	Rp	215,000.00	90	95	95
Eksekutif	10/01/2023	Rp	215,000.00	80	75	87
Eksekutif	11/01/2023	Rp	215,000.00	95	90	90
Eksekutif	12/01/2023	Rp	215,000.00	90	85	98
Eksekutif	13/01/2023	Rp	215,000.00	90	95	96
Eksekutif	14/01/2023	Rp	215,000.00	97	100	92
Eksekutif	15/01/2023	Rp	215,000.00	100	104	104
Eksekutif	16/01/2023	Rp	215,000.00	90	85	99
Eksekutif	17/01/2023	Rp	215,000.00	85	80	80
Eksekutif	18/01/2023	Rp	215,000.00	80	75	87
Eksekutif	19/01/2023	Rp	215,000.00	90	95	97
Eksekutif	20/01/2023	Rp	215,000.00	98	100	99
Eksekutif	21/01/2023	Rp	215,000.00	100	105	104
Eksekutif	22/01/2023	Rp	215,000.00	100	104	156
Eksekutif	23/01/2023	Rp	215,000.00	85	80	90
Eksekutif	24/01/2023	Rp	215,000.00	80	85	89
Eksekutif	25/01/2023	Rp	215,000.00	90	95	92
Eksekutif	26/01/2023	Rp	215,000.00	95	90	100
Eksekutif	27/01/2023	Rp	215,000.00	100	105	102
Eksekutif	28/01/2023	Rp	215,000.00	105	104	100
Eksekutif	29/01/2023	Rp	215,000.00	100	100	104
Eksekutif	30/01/2023	Rp	215,000.00	100	100	98
Eksekutif	31/01/2023	Rp	215,000.00	90	85	98

Figure 2. *Data Training*

Figure 2 contains the cumulative data of train passengers every day. We will use this data to cluster data using the k-means method.



DATA CLUSTER - Orange

Source
● File: KLASIFIKASI.xlsx
○ URL:

File Type
Automatically detect type

Info
62 instance(s)
6 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

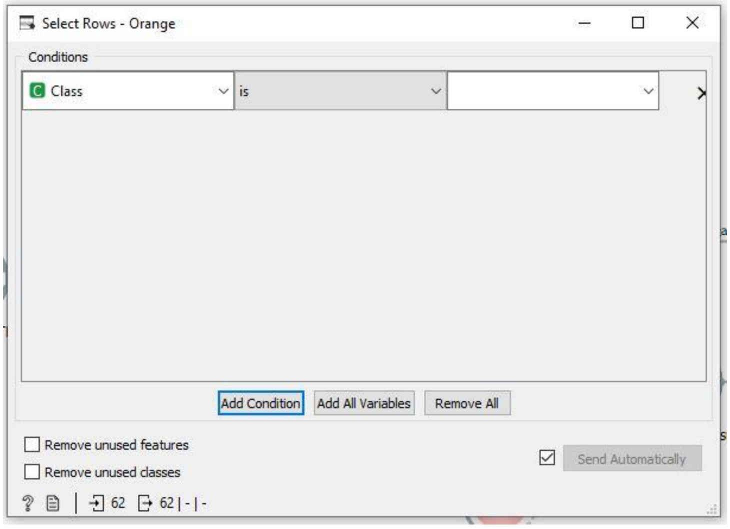
	Name	Type	Role	Values
1	Class	C categorical	target	Bisnis, Eksekutif
2	Jadwal	T datetime	meta	
3	Harga	S text	meta	
4	Sri Bilah Pagi	N numeric	feature	
5	Sri Bilah Siang	N numeric	feature	
6	Sri Bilah Malam	N numeric	feature	

Figure 3. *Selection of training data targets*

Figure 3 contains the attributes that we have compiled from the data we obtained from the Indonesian Railways (KAI) office in Medan and those attributes that we will use to carry out a data cluster using the k-means method and change the type of class attribute which was originally a target feature so that we can cluster correctly.

Data Row Selection Process

The row selection process is a process for selecting and determining which row we will make a condition for later to be clustered.



Select Rows - Orange

Conditions
C Class | is |

☐ Remove unused features
☐ Remove unused classes
☑ Send Automatically

Figure 4. *Data Selection Process/ Preprocessing*

In Figure 4 it contains the part of the row we selected which is contained in the training data. The selection is made to determine the part of the row that we will cluster.

*name of corresponding author

Data Mining Process

The data mining process is carried out using the data cluster model using the k-means method. To do this cluster we use the orange application. This is done to cluster cumulative data on Indonesian Railways (KAI) passengers.

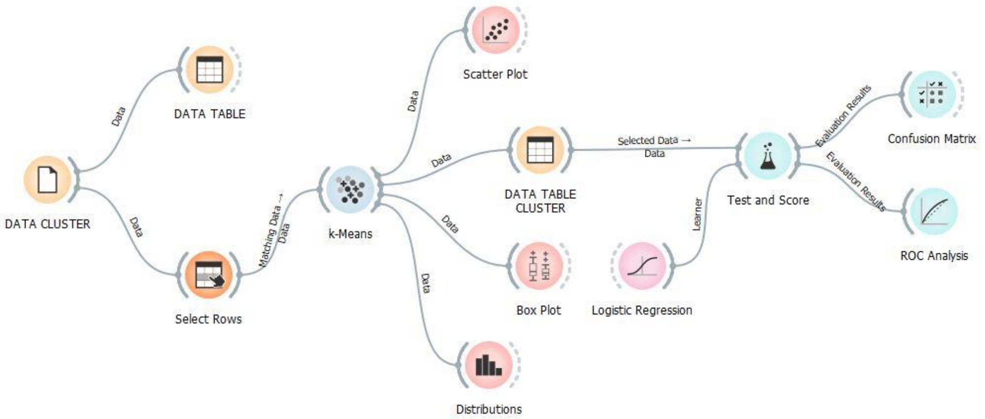


Figure 5. *Data Mining Process*

In Figure 5 is a widget that is used to carry out a data cluster using the k-means method. We will cluster cumulative data on train passengers based on similarities between data.

Proses Pengujian Model Klaster

In the data testing process, the neural network method will be used to classify community data. To carry out this classification we will need training data and test data which will be sample data, this data is data from the Kotapinang Subdistrict community which will be carried out using the k-means method for cluster data. To do a cluster we don't need to use training and testing data, but we only need 1 data that already contains the target attribute that we will cluster later.

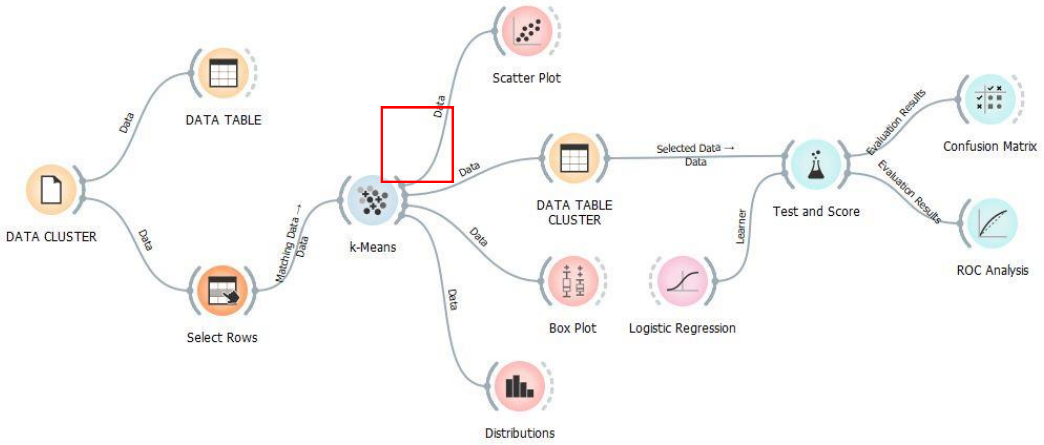


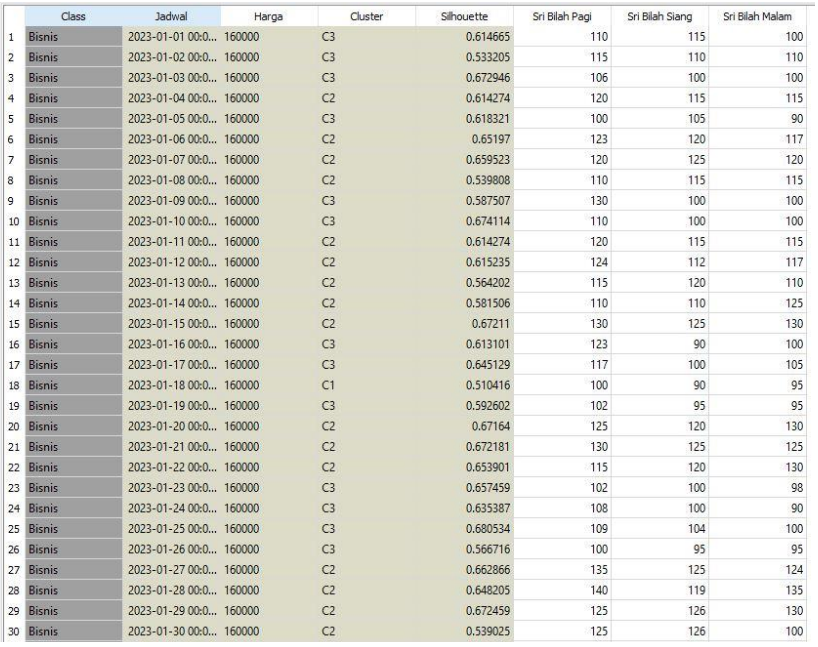
Figure 6. *Train Passenger Cluster Widget Design Model*

In Figure 6 is the widget pattern needed when carrying out a data cluster. The widget in the red box is the k-means method that we use to cluster data. We will cluster data based on the similarity of the data.

*name of corresponding author

Cluster Model Prediction Process

This process is a process carried out to carry out a prediction by clustering data using the k-means method. The data will be clustered based on the similarity of the data. For the results of predictions Classification of this data can be seen in Figure 7.



	Class	Jadwal	Harga	Cluster	Silhouette	Sri Bilah Pagi	Sri Bilah Siang	Sri Bilah Malam
1	Bisnis	2023-01-01 00:0...	160000	C3	0.614665	110	115	100
2	Bisnis	2023-01-02 00:0...	160000	C3	0.533205	115	110	110
3	Bisnis	2023-01-03 00:0...	160000	C3	0.672946	106	100	100
4	Bisnis	2023-01-04 00:0...	160000	C2	0.614274	120	115	115
5	Bisnis	2023-01-05 00:0...	160000	C3	0.618321	100	105	90
6	Bisnis	2023-01-06 00:0...	160000	C2	0.65197	123	120	117
7	Bisnis	2023-01-07 00:0...	160000	C2	0.659523	120	125	120
8	Bisnis	2023-01-08 00:0...	160000	C2	0.539808	110	115	115
9	Bisnis	2023-01-09 00:0...	160000	C3	0.587507	130	100	100
10	Bisnis	2023-01-10 00:0...	160000	C3	0.674114	110	100	100
11	Bisnis	2023-01-11 00:0...	160000	C2	0.614274	120	115	115
12	Bisnis	2023-01-12 00:0...	160000	C2	0.615235	124	112	117
13	Bisnis	2023-01-13 00:0...	160000	C2	0.564202	115	120	110
14	Bisnis	2023-01-14 00:0...	160000	C2	0.581506	110	110	125
15	Bisnis	2023-01-15 00:0...	160000	C2	0.67211	130	125	130
16	Bisnis	2023-01-16 00:0...	160000	C3	0.613101	123	90	100
17	Bisnis	2023-01-17 00:0...	160000	C3	0.645129	117	100	105
18	Bisnis	2023-01-18 00:0...	160000	C1	0.510416	100	90	95
19	Bisnis	2023-01-19 00:0...	160000	C3	0.592602	102	95	95
20	Bisnis	2023-01-20 00:0...	160000	C2	0.67164	125	120	130
21	Bisnis	2023-01-21 00:0...	160000	C2	0.672181	130	125	125
22	Bisnis	2023-01-22 00:0...	160000	C2	0.653901	115	120	130
23	Bisnis	2023-01-23 00:0...	160000	C3	0.657459	102	100	98
24	Bisnis	2023-01-24 00:0...	160000	C3	0.635387	108	100	90
25	Bisnis	2023-01-25 00:0...	160000	C3	0.680534	109	104	100
26	Bisnis	2023-01-26 00:0...	160000	C3	0.566716	100	95	95
27	Bisnis	2023-01-27 00:0...	160000	C2	0.662866	135	125	124
28	Bisnis	2023-01-28 00:0...	160000	C2	0.648205	140	119	135
29	Bisnis	2023-01-29 00:0...	160000	C2	0.672459	125	126	130
30	Bisnis	2023-01-30 00:0...	160000	C2	0.539025	125	126	100

Figure 7. *Results from predictions using the K-Means method*

Figure 7 shows the predicted results from the data cluster process using the k-means method.

Cluster Model Evaluation Results

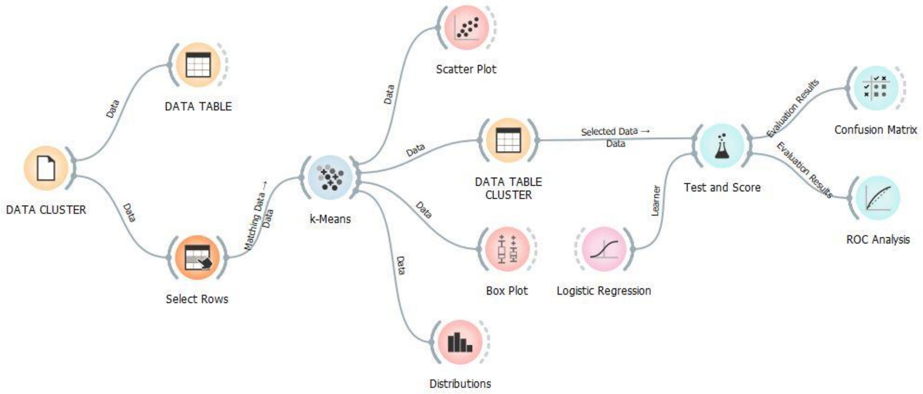
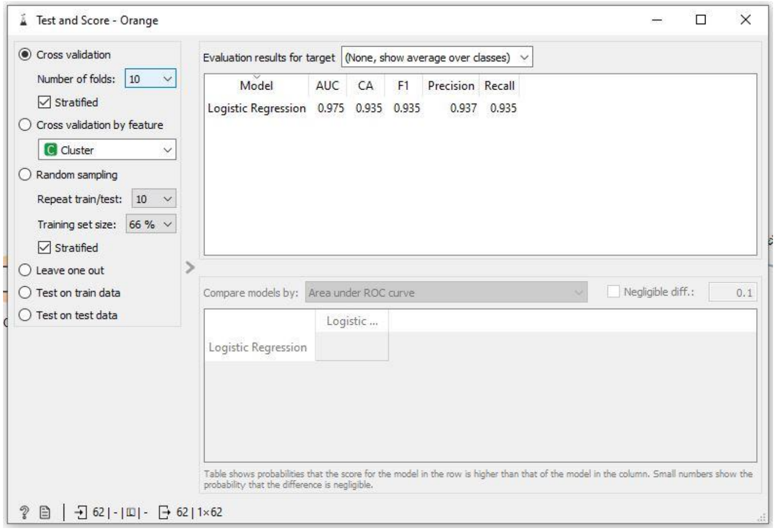


Figure 8. *Cluster Evaluation Widget*

Figure 8 contains the results of the data cluster evaluation which consists of several widgets needed to determine test scores and scores. After we get the test scores and scores, later we will look for values from the confusion matrix and ROC analysis. To get these three results, we need to use a widget called Logistic Regression so that the results from tests and scores, Confusion Matrix and ROC Analysis can come out.

*name of corresponding author



Test and Score - Orange

● Cross validation
Number of folds: 10
☑ Stratified
○ Cross validation by feature
Cluster
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.975	0.935	0.935	0.937	0.935

Compare models by: Area under ROC curve   □ Negligible diff.: 0.1

	Logistic ...
Logistic Regression	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

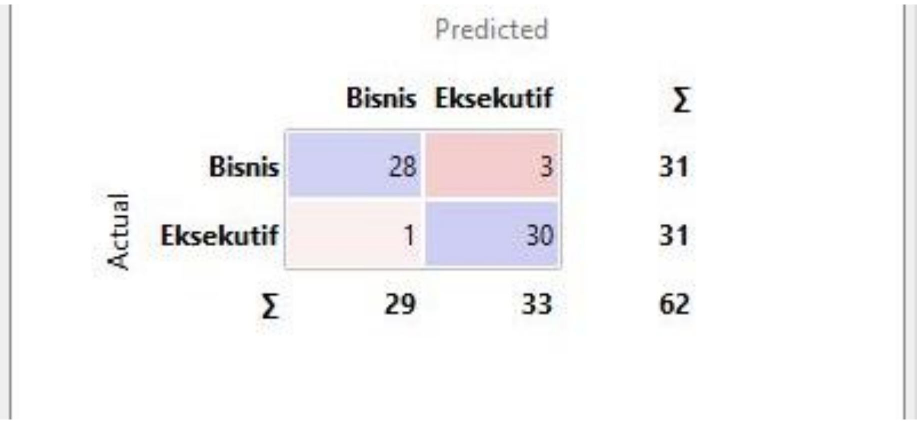
62 | - | - | - 62 | 1×62

Figure 9. *Results of Test and Score*

In Figure 9 are the results of the test and the score we get from 62 cumulative data on train passengers, so we will get the result of an AUC of 0.975.

Evaluation Results with Confusion Matrix

The Confusion Matrix is a measuring tool for making predictions by adjusting data based on data similarity using the k-means method.



		Predicted		
		Bisnis	Eksekutif	Σ
Actual	Bisnis	28	3	31
	Eksekutif	1	30	31
	Σ	29	33	62

Figure 10. *Confusion Matrix value of K-Means method*

Figure 10. The True Positive (TP) result is 28. True Negative (TN) is 30, False Positive (FP) is 3 and False Negative (FN) is 1. Then the values for accuracy, precision and recall are as follows:

$$\boldsymbol{Accuracy} = \frac{\mathbf{28+30}}{\mathbf{28+30+3+1}} + \mathbf{100\%}$$ Then the Accuracy value = 93%

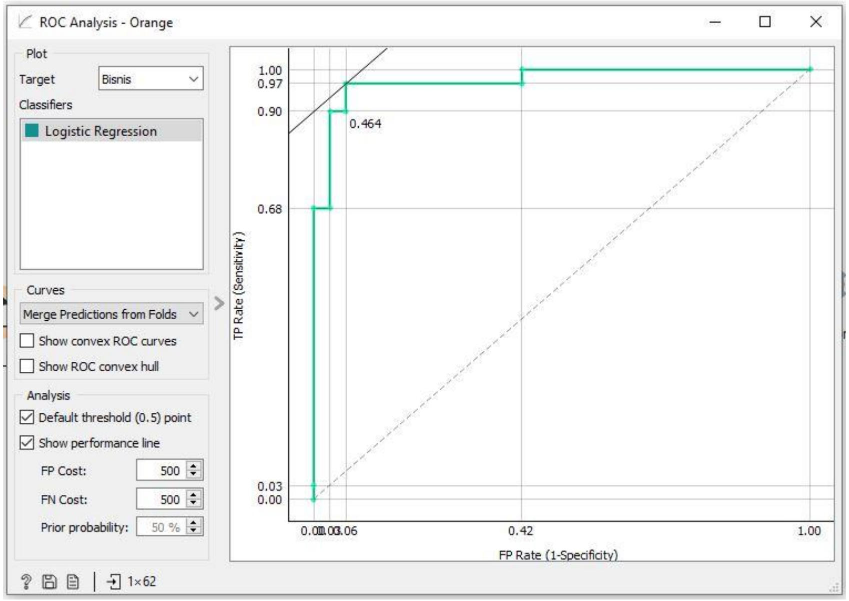
$$\boldsymbol{Presisi} = \frac{\mathbf{28}}{\mathbf{28+3}} + \mathbf{100\%}$$ Then the Precision value = 90%

$$\boldsymbol{Recall} = \frac{\mathbf{28}}{\mathbf{28+1}} + \mathbf{100\%}$$ Then the Recall value = 96%

*name of corresponding author

Evaluation Results with ROC Curve

The Roc Curve is obtained from the true signal (sensitivity) and (1 specificity) over the entire cut off point range to obtain the ROC curve visualized from the Confusion Matrix. The results of the ROC graph can be seen in Figures 11 and 12.



ROC Analysis - Orange

Plot

Target: Bisnis

Classifiers

🟩 Logistic Regression

Curves

Merge Predictions from Folds

☐ Show convex ROC curves

☐ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

FP Cost: 500

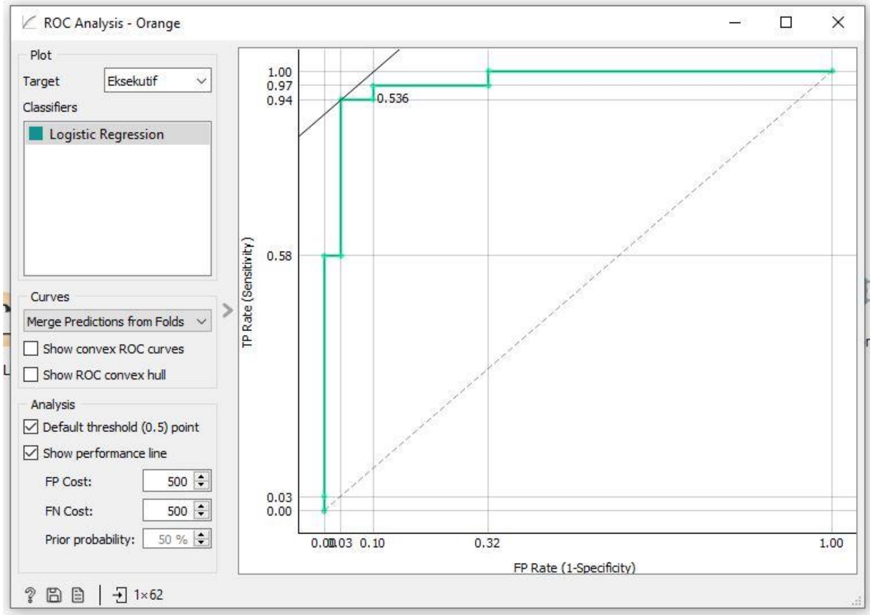
FN Cost: 500

Prior probability: 50 %

1×62

***Figure 11.** ROC Analysis Targeting Social Assistance Eligible People*

Figure 11 states that the results of the ROC Analysis of cumulative passengers in business class using the k-means method, the result is 0.464.



ROC Analysis - Orange

Plot

Target: Eksekutif

Classifiers

🟩 Logistic Regression

Curves

Merge Predictions from Folds

☐ Show convex ROC curves

☐ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

FP Cost: 500

FN Cost: 500

Prior probability: 50 %

1×62

***Figure 12.** ROC analysis targeting people who are not eligible for social assistance*

Figure 12 states that the results of the ROC analysis of cumulative passengers in the executive class using the k-means method, the result is 0.536.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12260

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Results of Scatter Plots

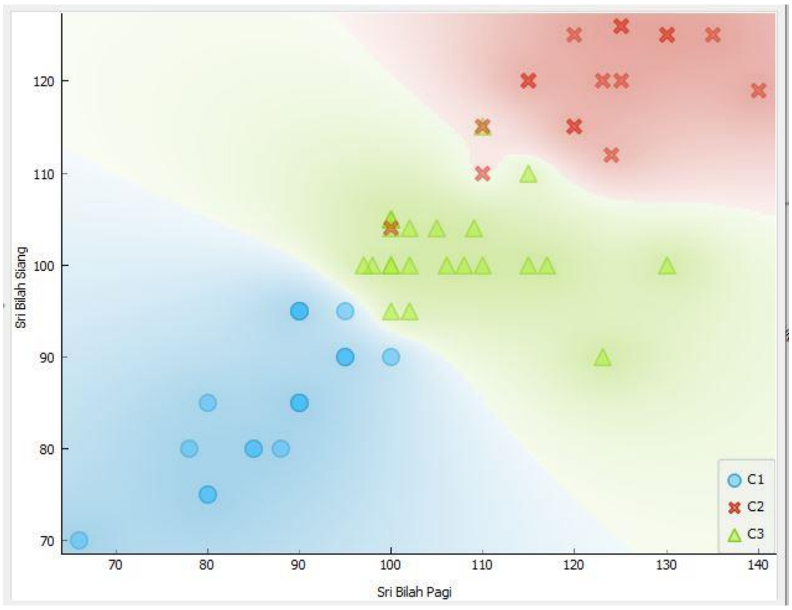
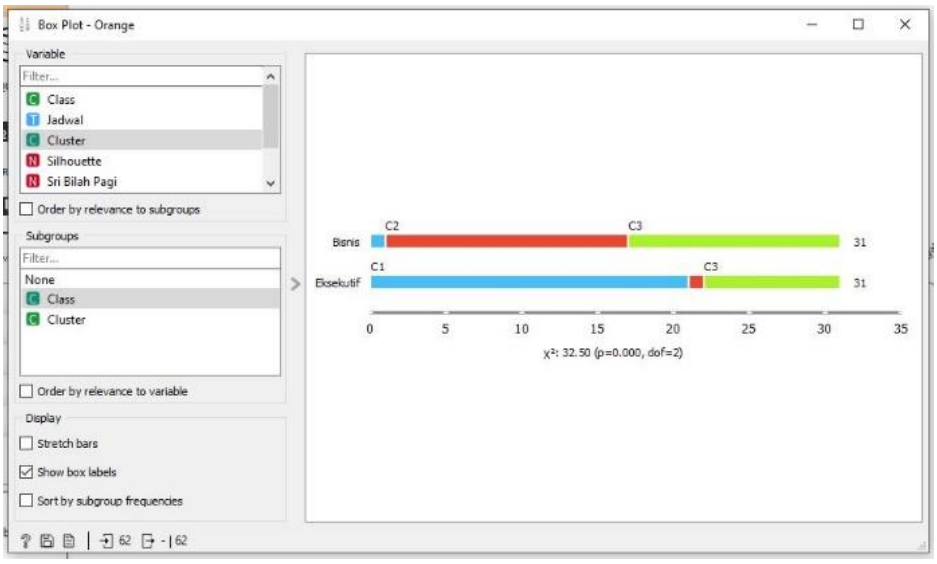


Figure 13. *Hasil Scatter Plot*

Figure 13 contains the results of the cumulative data cluster for train passengers. Each cluster has different shapes and the results of the three clusters that we have done using the k-means method can already be seen in the scatter plot.

Results of Box Plots


Box Plot - Orange

Variable

Filter...
- Class
- Jadwal
- Cluster
- Silhouette
- Sri Bilah Pagi

☐ Order by relevance to subgroups

Subgroups

Filter...
- None
- Class
- Cluster

☐ Order by relevance to variable

Display

☐ Stretch bars
☑ Show box labels
☐ Sort by subgroup frequencies

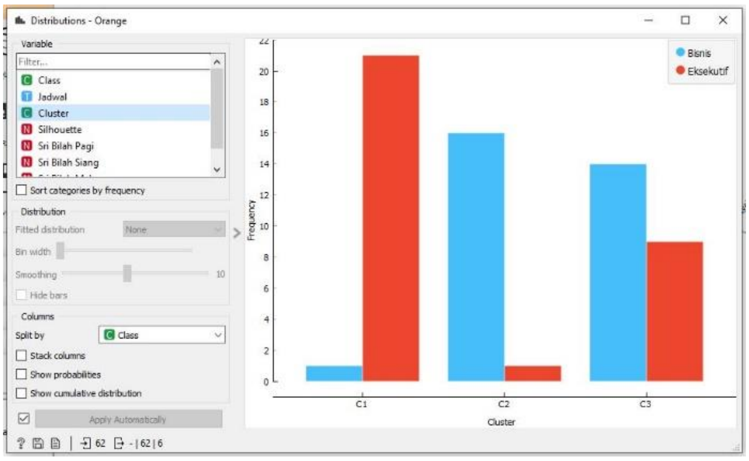
χ²: 32.50 (p=0.000, dof=2)

Figure 14. *Cluster Evaluation Widget*

Figure 14 contains the amount of data entered in clusters 1, 2 and 3, both business and executive classes, but the shape is horizontal. It is clear that most data for the executive class is included in cluster 1, but some executive data is included in cluster 2 and cluster 3. Then the business class data is almost divided into two, namely cluster 2 and cluster 3, but there is also a small amount of data business class is included in cluster 1.

*name of corresponding author

Results from Distributions



Distributions - Orange

Variable

Filter...

- Class
- Jadwal
- Cluster
- Silhouette
- Sri Bilah Pagi
- Sri Bilah Siang

☐ Sort categories by frequency

Distribution

Fitted distribution: None

Bin width

Smoothing: 10

☐ Hide bars

Columns

Split by: Class

☐ Stack columns

☐ Show probabilities

☐ Show cumulative distribution

☑ Apply Automatically

62 | - | 62 | 6

Figure 15. *Cluster Evaluation Widget*

In Figure 15, the contents are almost the same as the explanation and form of the Boxplot results, it's just that the Distribution results have a vertical shape like a bar chart in general. So the explanation is the same as the Box Plot. The two widgets display the amount of data from each class included in clusters 1, 2 and 3.

DISCUSSIONS

Land transportation is very popular in Indonesia, because many people think that riding a land vehicle can make passengers comfortable, especially if they use rail vehicles. The train is one of the favorites for people who are afraid to drive. Therefore, we will try to cluster the passenger data into 3 clusters. To carry out a train passenger data cluster, we will use the K-Means cluster method. This method is often used by people to do a data cluster. so this data becomes a very suitable method for clustering data. After we use the k-means method, we will find the results of the passenger data clusters.

CONCLUSION

The means of transportation that are often used by passengers is ground transportation. In this study, we will conduct a data cluster. A cluster will store data based on the similarity of the data. The data that we will cluster is train passenger data for January 2023. We will cluster passenger data from business and executive classes. In the cluster process, we will need a method that can process data for the cluster. Previously, clustering was a process of grouping data based on the similarity of a data. Cluster systems will be able to minimize energy and be able to group data efficiently. A cluster can be characterized by looking at the similarities and similarities of the group attributes and dissimilarity can also be seen from the group attributes. This research was conducted to cluster train passenger data. By using the K-Means method we will cluster the data into 3 clusters. Each cluster will have its own group based on the similarity of the data. The K-Means method is a method used to cluster data based on data similarity.

REFERENCES

Adawia, P. R., Azizah, A., Endriastuty, Y., & Sugandhi, S. (2020). Pengaruh Kualitas Pelayanan Dan Fasilitas Terhadap Kepuasan Konsumen Kereta Api Commuter Line (Studi Kasus Commuter Line Arah Cikarang Ke Jakarta Kota). *Sebatik*, *24*(1), 87–95. https://doi.org/10.46984/sebatik.v24i1.869

Agustina, N., Adrian, A., & Hermawati, M. (2021). Implementasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Berita Palsu pada Sosial Media. *Faktor Exacta*, *14*(4), 1979–276. https://doi.org/10.30998/faktorexacta.v14i4.11259

Al-Ars, Z. T., & Al-Bakry, A. (2019). A web/mobile decision support system to improve medical

*name of corresponding author

diagnosis using a combination of K-mean and fuzzy logic. *Telkomnika (Telecommunication Computing Electronics and Control)*, *17*(6), 3145–3154. https://doi.org/10.12928/TELKOMNIKA.v17i6.12715

Bui, X. N., Nguyen, H., Choi, Y., Nguyen-Thoi, T., Zhou, J., & Dou, J. (2020). Prediction of slope failure in open-pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm. *Scientific Reports*, *10*(1), 1–17. https://doi.org/10.1038/s41598-020-66904-y

Dirjen, S. K., Riset, P., Pengembangan, D., Dikti, R., Yaumi, A. S., Zulfiqkar, Z., & Nugroho, A. (2018). *Terakreditasi SINTA Peringkat 4 Klasterisasi Karakter Konsumen Terhadap Kecenderungan Pemilihan Produk Menggunakan K-Means*. *3*(1), 195–202.

Ghaedi, H., Farizani, S. R. K. T., & Ghaemi, R. (2021). Improving power theft detection using efficient clustering and ensemble classification. *International Journal of Electrical and Computer Engineering*, *11*(5), 3704–3717. https://doi.org/10.11591/ijece.v11i5.pp3704-3717

Hamzaoui, Y., Amnai, M., Choukri, A., & Fakhri, Y. (2020). Enhancenig OLSR routing protocol using K-means clustering in MANETs. *International Journal of Electrical and Computer Engineering*, *10*(4), 3715–3724. https://doi.org/10.11591/ijece.v10i4.pp3715-3724

Hassen, B. S., Lafta, S. A. S., Noman, H. M., & Ali, A. H. (2019). Analyzing the performances of WSNs routing protocols in grid- based clustering. *International Journal on Advanced Science, Engineering and Information Technology*, *9*(4), 1211–1216. https://doi.org/10.18517/ijaseit.9.4.8900

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *5*(2), 697–711. Retrieved from http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369

Riza, L. S., Rosdiyana, R. A., Wahyudin, A., & Pérez, A. R. (2021). The k-means algorithm for generating sets of items in educational assessment. *Indonesian Journal of Science and Technology*, *6*(1), 93–100. https://doi.org/10.17509/ijost.v6i1.31523

Rustam, Z., Hartini, S., Pratama, R. Y., Yunus, R. E., & Hidayat, R. (2020). Analysis of architecture combining Convolutional Neural Network (CNN) and kernel K-means clustering for lung cancer diagnosis. *International Journal on Advanced Science, Engineering and Information Technology*, *10*(3), 1200–1206. https://doi.org/10.18517/ijaseit.10.3.12113

Uçar, T., & Karahoca, A. (2021). Benchmarking data mining approaches for traveler segmentation. *International Journal of Electrical and Computer Engineering*, *11*(1), 409–415. https://doi.org/10.11591/ijece.v11i1.pp409-415

Widiyanto, A. T., & Witanti, A. (2021). Segmentasi Pelanggan Berdasarkan Analisis RFM Menggunakan Algoritma K-Means Sebagai Dasar Strategi Pemasaran (Studi Kasus PT Coversuper Indonesia Global). *KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi*, *1*(1), 204–215. https://doi.org/10.24002/konstelasi.v1i1.4293

Yassir, A. H., Mohammed, A. A., Alkhazraji, A. A. J., Hameed, M. E., Talib, M. S., & Ali, M. F. (2020). Sentimental classification analysis of polarity multi-view textual data using data mining techniques. *International Journal of Electrical and Computer Engineering*, *10*(5), 5526–5534. https://doi.org/10.11591/IJECE.V10I5.PP5526-5534

Yun, H. (2021). Prediction model of algal blooms using logistic regression and confusion matrix. *International Journal of Electrical and Computer Engineering*, *11*(3), 2407–2413. https://doi.org/10.11591/ijece.v11i3.pp2407-2413