Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

Clustering Analysis of Bus Fares Trans Metro Deli Medan Using Mean Shif Clustering Method

Rinanda Putri Rambe*, Ilka Zufria

Science and Technology, Computer Science, State Islamic University of North Sumatera, Medan, Indonesia
Email: [1,*]rinandaputrirambe@gmail.com, [2]ilkazufria@uinsu.ac.id
Correspondence Author Email: rinandaputrirambe@gmail.com
Submitted: 03/08/2024; Accepted: 10/08/2024; Published: 10/08/2024

Abstract−Medan City is the 3rd most populous city in Indonesia, according to data from the Central Statistics Agency, Medan has a population of 2.49 million in 2022, an increase from the previous 2.46 million in 2021. The increasing number of population inhabiting the city of Medan means that the need for transportation for the people of Medan is also increasing. Trans Metro Deli bus data can be grouped effectively using the mean shift algorithm based on several attributes, namely passenger category, payment method and fare. Each passenger group has different needs and ability to pay, which makes setting fair and efficient fares a challenge. Inappropriate pricing can lead to passenger dissatisfaction, reduce the number of public transportation users, and affect bus operators' revenue. Cluster technique is a well-known clustering technique, which aims to group data into clusters so that each cluster contains data that is as similar as possible. Mean shift belongs to the category of clustering algorithms with unsupervised learning that assigns data points to clusters iteratively by shifting the points towards the mode (mode is the highest density of data points in the region in the context of mean shift). Mean shift does not require determining the number of clusters in advance The attributes used in the clustering process, namely passenger category, payment method and fare can properly create a hyperplane between clusters, thus creating significant differences from each cluster, as evidenced by the silhouette score obtained by 0.64. By conducting this analysis, it is expected to find a more efficient and fair fare clustering pattern, and provide practical recommendations for management in setting fares that are more in line with passenger needs. In addition, this research also aims to evaluate the effectiveness of mean shift clustering in the context of transportation fare analysis.

Keywords: Classification; Mean-Shift Clustering; Cluster; Silhouette Score

1. INTRODUCTION

Medan City is the 3rd most populous city in Indonesia, according to data from the Central Statistics Agency, Medan has a population of 2.49 million in 2022, an increase from the previous 2.46 million in 2021 [1]. The increasing number of population inhabiting the city of Medan means that the need for transportation for the people of Medan is also increasing. The need for transportation is very important for the community and facilitates the process of mobilizing the transportation of goods, services and human resources to move or reside from one area to another [2].

In November 2020, the Ministry of Transportation launched the Teman Bus service in Medan City, a Bus Rapid Transit (BRT) public transportation mode that is safe, comfortable, fast, and cheap for the economy of the Medan community whose operations will cover the Pinang Baris Terminal, Amplas Terminal, Medan Belawan, Medan Tuntungan, and Tembung [3]. The Teman Bus service in Medan City is called the Trans Metro Deli Medan Bus. The Medan City Government uses the Trans Metro Deli Bus as one of the modes of transportation in Medan City [4]. The use of the Trans Metro Deli Bus has the aim of being able to serve the needs of passenger demand along the route (Dishub, 2020).

Another study was also conducted by Azdi Rihadi Harahap in 2022 entitled "Analysis of the Number of Passengers and Determination of Locations at Trans Metro Deli Corridor V Bus Stops Based on the Set Covering Problem Method" where this study determines the number and location of corridor V bus stops in Medan City and provides proper access to all passengers by adding the minimum number of stops that can meet all demand points along corridor V [5]. Based on previous research, this study was conducted with a difference, namely that this study determines the grouping of Trans Metro Deli Medan bus fares based on the Medan Tuntungan corridor IV only, while previous research was only located along corridor V and this study uses the Mean-Shift Clustering method where the advantage of this algorithm is that it can determine clustering without requiring the number of clusters first [6].

Data mining is a technique that allows to obtain patterns or models from collected data [7]. This technique is applied in all types of environments such as biology, educational and financial applications, industry, police, and political processes. In data mining there are several techniques, including rule induction and decision trees which according to various studies conducted are among the most widely used [8]. Clustering is one of the methods of data mining and clustering has become a valid instrument for solving complex problems of computer science and statistics. Clustering is the process of grouping data points into two or more groups so that data points that fall into the same group are more similar to each other than in different groups, based only on the information available with the data points [9].

The process of data mining includes collecting raw data from basic data such as relations, data warehouses, information reservations and more advanced reservations, object-oriented, and object-relational, transactional and spatial, heterogeneous and legacy, multimedia and streaming, words, word mining and web

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

mining [10]. The process involves data mining to produce insights that make the information better known and understood. The process is like knowledge discovery, information retrieval, pattern analysis and knowledge extraction that allows understanding of data, leading to constructive measurements of the areas involved [11].

Clustering is the process of creating groupings so that all members of each partition have similarities based on a certain matrix. Cluster analysis or group analysis is a data analysis technique that aims to group individuals or objects located in one group will have relatively homogeneous properties. The purpose of cluster analysis is to group these objects [12].

Initially, the clustering or average algorithm was discovered by several people such as Lloyd (1957-1982), Forgey (1965), Friedman and Rubin (1967), and McQueen (1967). The idea of clustering was first discovered by Lloyd in 1957. However, it was only published in 1982. In 1965, Forgey also published the same technique so that it is sometimes known as Lloyd-Forgyn in some sources. In addition, in the midst of the rapid development of artificial intelligence technology, it consists of several branches, one of which is machine learning [13]. Machine learning technology is one branch of AI that is very interesting, because machine learning is a machine that can learn like humans. Machine learning is a technology that is able to study existing data and perform certain tasks according to what is learned [14].

The basic concept of mean shift clustering is based on the concept of moving the center point (centroid) of a cluster to its average position (mean) [15]. This transfer is done repeatedly until convergence, where there is no more significant change in the position of the center point. Mean shift is a non-parametric method, which means that there is no need to make assumptions about the distribution of the data. This makes it more flexible in handling complex and unstructured data [14]. Kernel density estimation This method uses density estimation kernel density estimation allows the algorithm to determine the probability of the existence of a new center point based on the density of the surrounding data.

Python is a popular and easy-to-learn programming language. Python is widely used in software development, artificial intelligence, web development, machine learning, and data analysis [19]. Python provides various libraries, such as NumPy for numeric computing and pandas for data analysis, making it easier for someone to perform certain tasks quickly and efficiently. At this stage is the last stage in testing through the Python language, this stage is carried out in the form of the results of the total cluster and the final centroid [16].

This study aims to apply mean shift clustering in grouping Trans Metro Deli Medan bus fares. By conducting this analysis, it is expected to find a more efficient and fair fare clustering pattern, and provide practical recommendations for management in setting fares that are more in line with passenger needs. In addition, this research also aims to evaluate the effectiveness of mean shift clustering in the context of transportation fare analysis.

2. RESEARCH METHODOLOGY

2.1 Research Stages

In conducting research, a research methodology is needed that contains a research model [20]. The method used in this study is a quantitative method where the data is taken from the Medan Tuntungan corridor IV. The research framework contains a description of the steps taken when conducting research, so that the research carried out runs systematically and the expected goals can be achieved.

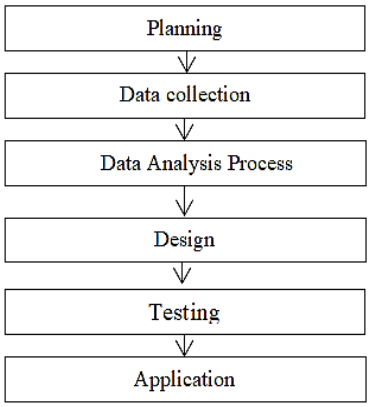

Figure 1. Research Framework

The iterative process of the mean shift algorithm iterates where each data point is moved to a new position which is the center of a denser region in data density. This process continues until there is no more significant change in the position of the center point [6].

Here are the general steps of how the mean shift clustering algorithm works:

1. Initialization: each data point is considered as the initial center of a cluster.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

2. Cluster center estimation: for each cluster center, a kernel function (usually Gaussian) is applied to measure the density of data points around it.
3. Center shift: the cluster center is moved to the center of mass (mean) of the points that have the highest density within the kernel range.
4. Iteration: steps 2 and 3 are repeated until there is no more significant change in the position of the cluster center, or until the maximum number of iterations is reached [11].

The formula used in the mean shift algorithm is as follows:

1. Kernel function (K(x)): a function that measures the density of data points around a point x [16].

   $$K(x) = \frac{1}{h^d} \bullet f\left(\frac{||x - x_i||^2}{h^2}\right)$$

   xi is the cluster center, h is the bandwidth, which is a parameter that controls the effective distance of the kernel function, || . || indicates the Euclidean norm, f (.) is the kernel function (usually Gaussian) and d is the dimension of the data space.
2. Cluster center update:

   $$m(x) = \frac{\sum_{i=1}^{N} K(x - x_i) \cdot x_i}{\sum_{i=1}^{N} K(x - x_i)} m$$

   (x) is the new cluster center, xi are the data points and K (x – xi) is the kernel function that measures the density of data points around point x [17]
3. Cluster center shift

   $$x_{i+1} = m(x_i)$$

   xi is the current cluster center and xi +1 is the updated cluster center.

The mean shift algorithm will continue to repeat the above steps until there is no significant change in the position of the cluster center. After convergence, data points that are within a specified distance from the cluster center will be considered as members of that cluster [18].

3. RESULT AND DISCUSSION

In conducting research, a research methodology is needed which contains a research model [21]. The method used in this research is a quantitative method where the data is taken from corridor IV Medan Tuntungan. In the research framework there is a description of the steps taken when conducting research, so that the research carried out runs systematically and the expected goals can be achieved.

1. Planning
   The initial process of this research is planning, namely determining what topics will be discussed. The research topic that I discussed was the analysis of the grouping of Trans Metro Deli Medan bus fares using the mean shift clustering method.
2. Data Collection
   The data collected was 1100 data, then verified to ensure accuracy and consistency, and to avoid duplication or input errors. The large and representative amount of data allows clustering analysis using the Mean-Shift method to identify relevant fare patterns and provide a clear picture of the bus fare structure. With adequate data, this study can provide comprehensive insights for more effective bus fare planning and management. The following is a sample of the data that will be used.

Table 1. Raw Data

No	Passenger name	Date / time	Passenger Category	Payment method	Cost
1	Nadia	2024-03-15	Student	QRIS	2000
2	Yusuf	2024-07-07	Student	QRIS	2000
3	Anita	2024-04-01	Student	QRIS	2000
4	Rizki	2024-01-18	Student	QRIS	2000
5	Siska	2024-04-01	Student	QRIS	2000
6	Arif	2024-05-21	General	Electronic card	4300
7	Nurul	2024-05-31	Elderly	Electronic card	2000
8	Fahmi	2024-01-02	Elderly	Electronic card	2000
9	Risma	2024-01-02	Elderly	Electronic card	2000
..	..	..	..	..	..
1096	Dewi	2024-05-28	Student	QRIS	2000
1097	Yuda	2024-05-14	Student	QRIS	2000
1098	Dini	2024-06-20	Student	Electronic card	2000
1099	Ari	2024-07-20	Elderly	QRIS	2000

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

No	Passenger name	Date / time	Passenger Category	Payment method	Cost
1100	Andri	2024-04-16	Elderly	Electronic card	2000

In Table 1, passenger names were used to identify individuals in this study, while trip times recorded the specific times when fares were paid, providing temporal context to the data. Passenger categories, such as general, student, or elderly, provide insights into fare segmentation by demographic group. Payment methods, which include options such as QRIS or electronic cards, help in understanding payment preferences and habits.

3. Data Analysis Process
   The grouping process that will be carried out will not use all data columns as a reference for grouping, the columns that will be the reference for the grouping process are the Passenger Category, Payment Method and Tariff columns. Then for categorical data (which is not in the form of numbers), a transformation is carried out into numeric form so that the algorithm can process the data with the following provisions.
   a. Student = 1;
   b. General = 2;
   c. Elderly = 3;
   d. Electronic card = 1;
   e. QRIS = 2;

   After deleting columns and converting, the following is a representation of the data to be processed.

Table 2. Data To Be Analyzed

No	Passenger Category	Payment Method	Cost
1	1	2	2000
2	1	2	2000
3	1	2	2000
4	1	2	2000
5	1	2	2000
6	2	1	4300
7	3	1	2000
8	3	1	2000
9	3	1	2000
..	..	..	..
1096	1	2	2000
1097	1	2	2000
1098	1	1	2000
1099	3	2	2000
1100	3	1	2000

4. Design
   In this research, the design was carried out to create a system which is explained in a flowchart as follows:

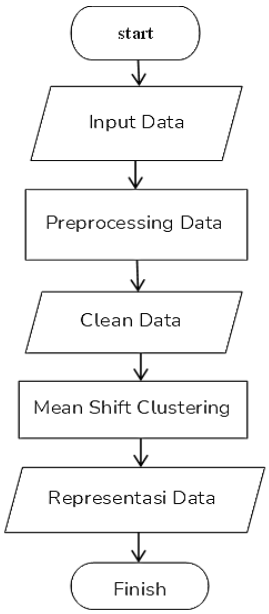

Figure 2. Flowchart System

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

3.1 Implementation System

Before conducting the analysis, the collected data needs to go through a pre-processing stage to ensure that the data is ready to be used in further analysis. This stage involves several important steps such as data cleaning, handling missing values, and data normalization. Data cleaning is done to eliminate duplication and inconsistencies, while handling missing values is done by filling in missing values using the imputation method or deleting incomplete entries. Data normalization is done to ensure that all variables have the same scale, which is important to avoid bias in clustering analysis [22].

For cleaning and eliminating duplication is not necessary because Trans Metro Deli has filtered the data so that there is no empty or duplicate data. The step that will be taken in this preprocessing stage is data normalization. The normalization technique used in this study is the z-score technique [23].
Here is the formula for the z-score

$$x_{norm} = \frac{x - \mu}{\delta}$$

Explanation:
x_{norm} = normalized data
x = initial data
μ = column mean
δ = standard deviation

Using this formula, the normalized value for the entire dataset can be obtained. Here is an example calculation for the first data point.
Step 1: Calculate the Mean of the First Column

$$\mu = \frac{1+1+1+1+1+2+3+3+3\ ..+1+1+1+1+3+3}{1100} = 1.9763$$

After obtaining the mean value of the first column, calculate the standard deviation of the first column.

$$\delta = \sqrt{\frac{(1-1.9763)^2+(1-1.9763)^2+(1-1.9763)^2\ +..+\ (1-1.9763)\text{^}2+\ (3-1.9763)\text{^}2+(3-1.9763)\text{^}2}{1100}} = 0.83$$

Step 2: Calculate the Normalized Value of the First Data Point

$$x_{norm} = \frac{1-\ 1.9763}{0.83} = -1.175$$

The normalized value for the last data point in the "Passenger Category" column:

$$x_{norm} = \frac{3\ -\ 1.9763}{0.83} = 1.232$$

Calculate the Mean of the "Payment Method" Column

$$\mu = \frac{2+2+2+2+2+1+1+1+1\ ..+1+2+2+1+2}{1100} = 1.513$$

After obtaining the mean value of the first column, calculate the standard deviation of the first column.

$$\delta = \sqrt{\frac{(2-1.513\ )^2+(2-1.513\ )^2+(2-1.513\ )^2\ +..+\ (2-1.513\ )\text{^}2+\ (1-1.513\ )\text{^}2+(2-1.513\ )\text{^}2}{1100}} = 0.499$$

Calculate the Normalized Value for the First Data Point in the "Payment Method" Column $x_{norm} = \frac{2-\ 1.513}{0.499} = 0.97$
The normalized value for the last data point in the "Payment Method" column:

$$x_{norm} = \frac{2\ -\ 1.513}{0.499} = 0.97$$

Calculate the Mean of the "Fare" Column

$$\mu = \frac{2000+2000+2000+\ ..+2000+2000+2000}{1100} = 2710.9$$

After obtaining the mean value of the first column, calculate the standard deviation of the first column.

$$\delta = \sqrt{\frac{(2000-2710)^2+(2000-2710\ )^2\ +..+\ (2000-2710\ )\text{^}2+\ (2000-2710)\text{^}2+(2000-2710\ )\text{^}2}{1100}} = 1062$$

Calculate the Normalized Value for the First Data Point in the "Cost" Column

$$x_{norm} = \frac{2000-\ 2710.9}{2710} = -0.66$$

The normalized value for the last data point in the "Cost" column:

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

$$x_{norm} = \frac{2000 - 2710}{2710} = -0.66$$

Table 3. Data Normalization Results

No	Passenger Category	Payment Method	Cost
1	-1.17511	0.973089	-0.66886
2	-1.17511	0.973089	-0.66886
3	-1.17511	0.973089	-0.66886
4	-1.17511	0.973089	-0.66886
5	-1.17511	0.973089	-0.66886
6	0.028448	-1.02765	1.49509
7	1.232001	-1.02765	-0.66886
8	1.232001	-1.02765	-0.66886
9	1.232001	-1.02765	-0.66886
..	..	..	..
1096	1095	1.232001	-1.02765
1097	1096	-1.17511	0.973089
1098	1097	-1.17511	0.973089
1099	1098	-1.17511	-1.02765
1100	1099	1.232001	0.973089

After knowing the flow of the process to be carried out. Then the flow is applied to a system built using the python programming language.

a. Import and initialize libraries
To facilitate the grouping process to be carried out, several libraries from python are utilized. The following is the python code to call the library to be used.

```
%pip install pandas scikit-learn openpyxl
import pandas as pd
from sklearn.cluster import MeanShift
from sklearn.metrics import silhouette_score
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import numpy as np
```

Figure 3. *Import* and Inisialisasi *Library*

b. Reading dataset
After importing and initializing the library, the next step is the process of reading the dataset. The dataset used is data stored in .xlsx format. Here is the python code to read the existing dataset.

```
# Ubah path ke lokasi file Excel Anda
excel_file = 'data_penumpang.xlsx'

# Membaca data dari file Excel
df = pd.read_excel(excel_file, engine='openpyxl')

# Tampilkan data untuk memastikan pembacaan berhasil
print("Data sebelum clustering:")
print(df.head(20))
```

Figure 4. Reading the dataset

Here is a display of the dataset that has been called using python.

```
Data sebelum clustering:
     Nama Penumpang Hari/Tanggal Kategori Penumpang Metode Pembayaran  Tarif
0             Nadia   2024-03-15            Pelajar              QRIS   2000
1             Yusuf   2024-07-07            Pelajar              QRIS   2000
2             Anita   2024-04-01            Pelajar              QRIS   2000
3             Rizki   2024-01-18            Pelajar              QRIS   2000
4             Siska   2024-04-01            Pelajar              QRIS   2000
...             ...          ...                ...                ...    ...
1095          Andri   2024-04-16            Lansia  Kartu Elektronik   2000
1096           Dewi   2024-05-28            Pelajar              QRIS   2000
1097           Yuda   2024-05-14            Pelajar              QRIS   2000
1098           Dini   2024-06-20            Pelajar  Kartu Elektronik   2000
1099            Ari   2024-07-20            Lansia              QRIS   2000

[1100 rows x 5 columns]
```

Figure 5. Dataset Display

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

c. Dataset transformation

The initial dataset is still in the 'raw' category. Therefore, dataset transformation is carried out. Where unused columns will be discarded and data in the form of "strings" / "categorical" will be changed to a number format according to the previous provisions. The following is the code to transform the dataset.

```
# Mengonversi nilai kategorikal menjadi angka
df['Kategori Penumpang'] = df['Kategori Penumpang'].map({'Pelajar': 1,
                                                         'Umum': 2,
                                                         'Lansia': 3})
df['Metode Pembayaran'] = df['Metode Pembayaran'].map({'Kartu Elektronik': 1,
                                                       'QRIS': 2})
df[['Kategori Penumpang', 'Metode Pembayaran', 'Tarif']]
```

Figure 6. Dataset Transformation

By implementing this code, the appearance of the dataset will change to look like the following image.

	Kategori Penumpang	Metode Pembayaran	Tarif
0	1	2	2000
1	1	2	2000
2	1	2	2000
3	1	2	2000
4	1	2	2000
...	...	...	...
1095	3	1	2000
1096	1	2	2000
1097	1	2	2000
1098	1	1	2000
1099	3	2	2000

1100 rows × 3 columns

Figure 7. Dataset Transformation Results

d. Data normalization

After the transformation, the dataset is now a model that can be analyzed, but to improve the analysis results, data normalization is carried out first by implementing the following code.

```
# Pilih fitur yang akan digunakan untuk clustering
X = df[['Kategori Penumpang', 'Metode Pembayaran', 'Tarif']]
# Normalisasi data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_scaled
```

Figure 8. Normalization of Dataset

With the code above, all data in the dataset will be normalized. Here is the data display after the normalization process.

```
array([[-1.17510534,  0.97308923, -0.66885605],
       [-1.17510534,  0.97308923, -0.66885605],
       [-1.17510534,  0.97308923, -0.66885605],
       ...,
       [-1.17510534,  0.97308923, -0.66885605],
       [-1.17510534, -1.02765498, -0.66885605],
       [ 1.23200058,  0.97308923, -0.66885605]])
```

Figure 9. Dataset Normalization Results

e. Clustering process

After normalization, the data is now ready to be analyzed [24]. Next, implement the mean shift algorithm on the data to group the data into several groups using the following code.

```
# Inisialisasi dan fitting model Mean Shift
ms = MeanShift()
ms.fit(X_scaled)

# Menambahkan hasil clustering ke dalam DataFrame
df['Cluster'] = ms.labels_

# Tampilkan hasil clustering
print("\nHasil clustering:")
print(df[['Nama Penumpang', 'Kategori Penumpang',
          'Metode Pembayaran', 'Tarif', 'Cluster']])
```

Figure 10. Clustering Process

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

By using the code above, the data grouping results will be displayed.

```
Hasil clustering:
     Nama Penumpang  Kategori Penumpang  Metode Pembayaran  Tarif  Cluster
0             Nadia                   1                  2   2000        0
1             Yusuf                   1                  2   2000        0
2             Anita                   1                  2   2000        0
3             Rizki                   1                  2   2000        0
4             Siska                   1                  2   2000        0
...             ...                 ...                ...    ...      ...
1095          Andri                   3                  1   2000        1
1096           Dewi                   1                  2   2000        0
1097           Yuda                   1                  2   2000        0
1098           Dini                   1                  1   2000        0
1099            Ari                   3                  2   2000        1

[1100 rows x 5 columns]
```

Figure 11. Clustering Results

f. Save data
After obtaining the clustering results from the meanshift algorithm, the clustering data is then saved in .xlsx format.

```
# Simpan hasil clustering ke dalam file Excel
output_file = 'Hasil Clustering.xlsx'
df.to_excel(output_file, index=False, engine='openpyxl')
print(f"\nHasil clustering disimpan ke dalam file: {output_file}")
```

Figure 12. Save Data

With the code above, the data will be saved to the local directory of the computer. The following is a display of the clustering results after using the mean shift algorithm.

Table 4. Clustering Results

No	Name	Date / time	Passenger Category	Payment Method	Cost	*Cluster*
1	Nadia	2024-03-15	Student	QRIS	2000	0
2	Yusuf	2024-07-07	Student	QRIS	2000	0
3	Anita	2024-04-01	Student	QRIS	2000	0
4	Rizki	2024-01-18	Student	QRIS	2000	0
5	Siska	2024-04-01	Student	QRIS	2000	0
6	Arif	2024-05-21	General	Electronic card	4300	2
7	Nurul	2024-05-31	Elderly	Electronic card	2000	1
8	Fahmi	2024-01-02	Elderly	Electronic card	2000	1
9	Risma	2024-01-02	Elderly	Electronic card	2000	1
..	..	..	..	..	..	
1096	Dewi	2024-05-28	Student	QRIS	2000	1
1097	Yuda	2024-05-14	Student	QRIS	2000	0
1098	Dini	2024-06-20	Student	Electronic card	2000	0
1099	Ari	2024-07-20	Elderly	QRIS	2000	0
1100	Andri	2024-04-16	Elderly	Electronic card	2000	1

From the clustering results obtained, it is known that the number of data classified in cluster 0 is 393 data, which is classified in cluster 1 is 367 data and which is classified in cluster 2 is 340 data. By observing the members in each cluster specifically for the fare value, it can be concluded that cluster 0 is a student group, then cluster 1 is classified in the elderly category and cluster 2 is included in the general category.

g. Clustering evaluation
In the analysis of Trans Metro Deli Medan bus fare clustering using the mean-shift clustering method, evaluation of clustering results is an important step to measure the quality and effectiveness of the clustering carried out. One method that is commonly used to evaluate the quality of clustering results is the silhouette score [25].

Silhouette score is a metric used to assess how well a data point is in its cluster compared to other clusters. This score provides an overview of how consistent an object is with its own cluster and how different the object is from other clusters. Silhouette scores range from -1 to 1, where

1. A score approaching 1 indicates that the data point is in the correct cluster and far from other clusters.
2. A score close to 0 indicates that the data point is near the boundary between two clusters.
3. A score close to -1 indicates that the data point may be grouped into the wrong cluster.
4. To find the value of the silhouette score, the program code below is used.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

4. CONCLUSION

Based on the research that has been conducted, there are several things that can be concluded, including that Trans Metro Deli bus data can be grouped effectively using the mean shift algorithm based on several attributes, namely passenger category, payment method and fare. Cluster technique is a well-known clustering technique, which aims to group data into clusters so that each cluster contains data that is as similar as possible. Mean shift belongs to the category of clustering algorithms with unsupervised learning that assigns data points to clusters iteratively by shifting the points towards the mode (mode is the highest density of data points in the region in the context of mean shift). Mean shift does not require determining the number of clusters in advance The attributes used in the clustering process, namely passenger category, payment method and fare can properly create a hyperplane between clusters, thus creating significant differences from each cluster, as evidenced by the silhouette score obtained by 0.64. By conducting this analysis, it is expected to find a more efficient and fair fare clustering pattern, and provide practical recommendations for management in setting fares that are more in line with passenger needs. In addition, this research also aims to evaluate the effectiveness of mean shift clustering in the context of transportation fare analysis.

REFERENCES

[1] M. R. Pratama, "Tinjauan Lokasi Halte Bus Trans Metro Deli Di Koridor 5 Medan Lapangan Merdeka – Tembung Terhadap Naik Turun Penumpang Bus Trans Metro Deli," 2021.
[2] G. N. Aulia and E. Patriya, "IMPLEMENTASI LEXICON BASED DAN NAIVE BAYES PADA ANALISIS SENTIMEN PENGGUNA TWITTER TOPIK PEMILIHAN PRESIDEN 2019," *infokom*, vol. 24, no. 2, pp. 140–153, 2019, doi: 10.35760/ik.2019.v24i2.2369.
[3] M. Rifadh and R. S. M. Sihombing, "KAPASITAS DINAS PERHUBUNGAN KOTA MEDAN DALAM PENGENDALIAN SARANA DAN PRASARANA BUS TRANS METRO DELI DI KOTA MEDAN," *JSSR*, vol. 6, no. 1, p. 174, Feb. 2023, doi: 10.54314/jssr.v6i1.1187.
[4] M. Mustopa, I. Junaedi, and A. Z. Sianipar, "SISTEM INFORMASI PENJUALAN DAN PENGENDALIAN STOCK BARANG BANGUNAN PADA TOKO BANGUNAN DELIMA," *JMIJayakarta*, vol. 1, no. 2, p. 105, Apr. 2021, doi: 10.52362/jmijayakarta.v1i2.447.
[5] A. R. Harahap, "PROGRAM STUDI TEKNIK INDUSTRI FAKULTAS TEKNIK UNIVERSITAS MEDAN AREA," 2022.
[6] S. I. Nurhafida and F. Sembiring, "ANALISIS TEXT CLUSTERING MASYARAKAT DI TWITER MENGENAI MCDONALD'SXBTS MENGGUNAKAN ORANGE DATA MINING," 2021.
[7] A. A. Arif, M. Firdaus, and Y. Maruhawa, "Comparison of Data Mining Methods for Predition of Rainfall with C4.5, Naïve Bayes, and KNN Algorithm," 2022.
[8] M. Rosadi, D. Aulia Nurhasanah, and M. Siddik Hasibuan, "Clustering Panjang Ruas Jalan di BBPJN Sumut Menggunakan Algoritma K-Means," *CoSIE*, pp. 29–38, Jan. 2023, doi: 10.55537/cosie.v2i1.567.
[9] R. Kurniawan R and I. Zufria, "Penerapan Text Mining Pada Sistem Penyeleksian Judul Skripsi Menggunakan Algoritma Latent Dirichlet Allocation(LDA)," *ijcs*, vol. 11, no. 3, Dec. 2022, doi: 10.33022/ijcs.v11i3.3120.
[10] Vrantika Br Samosir, Agung Mulyo Widodo, Nizirwan Anwar, Binastya Anggara Sekti, and Nixon Erzed, "Identifikasi Outlier Menggunakan Teknik Data Mining Clustering Untuk Analisis Data Tracer Study Pada Fakultas Ilmu Komputer Universitas Esa Unggul," *ikraith-informatika*, vol. 8, no. 1, pp. 162–174, Mar. 2024, doi: 10.37817/ikraith-informatika.v8i1.3211.
[11] R. Rizuan, E. Haerani, J. Jasril, and L. Oktavia, "Penerapan Algoritma Mean-Shift Pada Clustering Penerimaan Bantuan Pangan Non Tunai," *JoSYC*, vol. 4, no. 4, pp. 1019–1027, Aug. 2023, doi: 10.47065/josyc.v4i4.3876.
[12] M. W. Talakua, Z. A. Leleury, and A. W. Taluta, "ANALISIS CLUSTER DENGAN MENGGUNAKAN METODE K-MEANS UNTUK PENGELOMPOKKAN KABUPATEN/KOTA DI PROVINSI MALUKU BERDASARKAN INDIKATOR INDEKS PEMBANGUNAN MANUSIA TAHUN 2014," *BAREKENG: J. Mat. & Ter.*, vol. 11, no. 2, pp. 119–128, Dec. 2017, doi: 10.30598/barekengvol11iss2pp119-128.
[13] R. Dinata Kesuma and N. Hasdina, *Machine Learning*. Lhokseumawe: UNIMAL Press, 2022.
[14] D. Apriyanty, A. M. Putra, and R. A. Caesar, "Implementasi Mean Shift Clustering Dalam Mengelompokkan Pelanggan Retribusi Alat Pemadam Kebakaran Pada Dinas Pemadam Kebakaran dan Penanggulangan Bencana Kota Palembang," *Jurnal Sistem Informasi*, vol. 4, no. 1, 2023.
[15] S. Wirma, "Data Mining Dengan Metode Naïves Bayes Classifer dalam Memprediksi Tingkat Kepuasan Pelayanan Dokumen Kependudukan," *INFEB*, pp. 156–160, Sep. 2022, doi: 10.37034/infeb.v4i3.155.
[16] I. S. Tinendung and I. Zufria, "Pengelompokan Status Stunting Pada Anak Menggunakan Metode K-Means Clustering," *mib*, vol. 7, no. 4, p. 2014, Oct. 2023, doi: 10.30865/mib.v7i4.6908.
[17] R. Bahtiar, "Implementasi Data Mining Untuk Prediksi Penjualan Kusen Terlaris Menggunakan Metode K-Nearest Neighbor," 2023.
[18] A. Trisnawati, D. Surani, and A. Fidriyanto, "ANALYSIS OF MATERIAL UNDERSTANDING USING GOOGLE COLABORATORY IN CLASS X INFORMATICS SUBJECTS AT SMAN 5 SERANG," vol. 8, 2024.
[19] A. Ma'rif, *Buku Ajar Pemrograman Lanjut Bahasa Pemrograman Python*. Yogyakarta: Universitas Ahmad Dahlan, 2020.
[20] A. I. Putri and M. Furqan, "Application of Data Mining to Predict Birth Rates in Medan City Using the K-Nearest Neighbor Method," *Journal of Computer Science*, vol. 5, no. 1, 2024
[21] M. Ikhsan and R. Kurniawan, "Penerapan Text Mining pada Sistem Rekomendasi Pembimbing Skripsi Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier," *Indonesian Journal of Computer Science*, vol. 12, no. 6, 2023.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 832-841
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5728

[22] G. I. E. Soen, M. Marlina, and R. Renny, "Implementasi Cloud Computing dengan Google Colaboratory pada Aplikasi Pengolah Data Zoom Participants," *jitu*, vol. 6, no. 1, pp. 24–30, Jun. 2022, doi: 10.36596/jitu.v6i1.781.
[23] A. Yuniarti, A. Yasin, and A. N. Yohannes, "Efektifitas Algoritma Data Mining dalam Menentukan Pendonor Darah Potensial," *Syntax : Jurnal Informatika*, vol. 11, no. 01, pp. 12–22, 2022.
[24] R. Kurniawan, A. Halim, and H. Melisa, "Prediksi Hasil Panen Pertanian Salak di Daerah Tapanuli Selatan Menggunakan Algoritma SVM (Support Vector Machine)" 2023
[25] M. A. Ramdhani, "Gunung Djati Converence Series," vol. 3, 2021.