Journal of Information Systems and Informatics
Vol. 6, No. 3, September 2024 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalism.v6i3.841 Published By DRPM-UBD

Manhattan Metric in K-Means Clustering for Data Grouping

Mila Sari[1] , Armansyah[2]

[1,2] Department of Computer Science, State Islamic University of North Sumatra, Indonesia
Email: [1]milasari3601@gmail.com, [2]armansyah@uinsu.ac.id

Abstract

Clustering can be defined as a method commonly applied in data mining to group objects into clusters. Clusters consist of data objects that are similar to each other in a group but different from objects in other clusters. In this study, the data used is the data of KIP scholarship recipients for the 2016-2023 period. Various clustering metric measurement techniques have been frequently used by researchers, especially those focusing on distance and similarity metrics, such as Euclidean Distance, Manhattan, and Minkowski. In general, K-Means is an unsupervised learning method used in the clustering process to group data based on similarity. The elbow method is used to determine the optimal number of clusters, so that the clustering results obtained can be maximized to achieve better results. This study aims to analyze the use of Manhattan technique in K-Means clustering for data grouping. The research problem is how to analyze the Manhattan metric technique in K-Means clustering for effective data grouping. Applying the K-Means method shows that the existing data is successfully divided into four specified clusters. After determining the correct number of clusters, the K-Means method is used to sort the data in the dataset. From 3172 data, the final results obtained cluster 0 as many as 774 data, cluster 1 as many as 417 data, cluster 2 as many as 1244 data, and cluster 3 as many as 737 data. The results of the clustering process obtained a davies-bouldin index value of 1.4568.

Keywords : K-Means, Manhattan Distance, Elbow Method, Davies Bouldin Index

1. INTRODUCTION

Baskoro stated that clustering is a technique in data mining that is used to group objects into several clusters. In one cluster there is a group of data objects that have similarities, but are different from objects in other clusters [1] . Clustering can be interpreted as identifying groups of objects that have similarities with other data. By using this clustering technique, the density and distance between areas in the object space can be determined, and the overall distribution pattern and correlation between attributes can be determined. Samuel may be among the first to define machine learning (ML) as a branch of artificial intelligence that studies algorithm design methods that can "learn" from data without being explicitly programmed [2] . Partitional clustering is a method that divides data into several

clusters without any hierarchical relationship between the clusters. In this approach, each cluster has one center point (centroid), and the main goal is to reduce the distance (difference) between all data points and the corresponding cluster center. Examples of methods used include K-Means, K-Medoids, Fuzzy K-Means, and Mixed Modeling [3] . Thus, there are several clustering approach algorithms, one of which is the K-Means method.

Basically, K-Means is one of the unsupervised learning methods in the clustering process that groups data based on similarity, meaning that data that has similar characteristics will be placed in one cluster, while data that has different characteristics will be placed in different clusters [4] . This is influenced by the clustering metric measurement technique in forming clusters.

Several clustering metric measurement techniques that are often used by previous researchers, especially those focusing on distance and similarity metrics, include Euclidean Distance, Manhattan, and Minkowski. Euclidean distance is used to assess the level of similarity between data based on distance. Manhattan is used as a distance measurement method that is generated based on the number of differences between two data objects [5] . Minkowski is a metric in vector space that is considered a generalization of Euclidean Distance and Manhattan Distance [6] . There are various methods used in clustering, one of which is the elbow method which is used to determine the number of clusters. The elbow method is used to determine the optimal number of clusters, so that the clustering results obtained can be maximized to achieve better results [7] .

The purpose of this study is to conduct Manhattan technique analysis in K-Means clustering for data grouping. This is to find out the right metric value with the Manhattan technique approach and K-Means clustering for data grouping. And to find out the cluster of scholarship recipients based on Manhattan distance and K-Means clustering. Many studies use the Manhattan technique in K-Means clustering, including: based on research conducted by [8] entitled "Implementation of the Manhattan Distance-Based K-Means Clustering Algorithm for Student Field Concentration Clustering". The purpose of this study is to help students choose a concentration that suits their interests based on grades using the K-Means clustering method. This study discusses the grouping of grades from students in concentration courses to find out how many groups are formed from the existing data. The data needed are NIM, student name, courses and grades according to the provisions of the field concentration determined by the study program. The research conducted by Waskito Wahyu Pribadi, et al. aims to compare the calculation of K-Means Euclidean and Manhattan Distance using standard deviation and silhouette coefficient which are expected to be able to determine clustering in the Covid-19 zone in Malang district. The distance metrics used are Euclidean Distance and Manhattan Distance. The test results show that

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

Euclidean Distance is superior to Manhattan Distance, with a comparison of 0.71> 0.64 [9] . Further research has been conducted by Puspitasari et al. entitled "Grouping of Pepper Producing Areas Using the Application of the K-Means Algorithm". This study aims to group pepper planting areas through the application of the K-Means algorithm and three types of distance measurements, namely Euclidean Distance, Manhattan Distance, and Minkowski. The parameters that affect the grouping of pepper planting areas include area, production volume, average production, and number of plantation workers (TKP). Based on the accuracy test using the silhouette coordinate factor (SC) method, it was found that for this case study the best distance measurement method was the Manhattan Distance method [10] .

Based on references related to previous research, the Manhattan technique in K-Means clustering is seen as one of the methods that can be used to group KIP scholarship recipient data. The data used is discrete data. Therefore, the more appropriate technique to use is the Manhattan metric. Furthermore, this study is expected to explain how the Manhattan technique in the K-Means clustering method can be applied to analyze data grouping.

In conducting this research, the researcher used clustering analysis technique with K-Means method using Manhattan metric with each variable attribute to analyze Manhattan technique on KIP scholarship recipient data at State Islamic University of North Sumatra. By conducting Manhattan metric analysis, we can group data based on distance. Based on the background description, the formulation of this research problem is how to analyze Manhattan metric technique in K-Means clustering for data grouping. And problem limitations were carried out using Manhattan metric technique analysis using the K-Means algorithm for grouping data on KIP scholarship recipients for 2016-2023. by analyzing student data at the State Islamic University of North Sumatra, Medan.

2. METHOD

The research methodology describes the procedure for applying the K-Means algorithm to classify data into groups and identify attribute patterns in each cluster. This process includes data processing, determining the initial point of the cluster center (centroid), calculating the distance of the data to the cluster center using the Manhattan metric, grouping the data based on the smallest distance to the cluster center, and repeating the process if the object moves. The iteration is stopped when no objects move, and the results for each cluster are obtained. Furthermore, clustering evaluation is carried out using the Davies Bouldin Index. Figure 1 shows the flowchart of the K-Means algorithm.

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

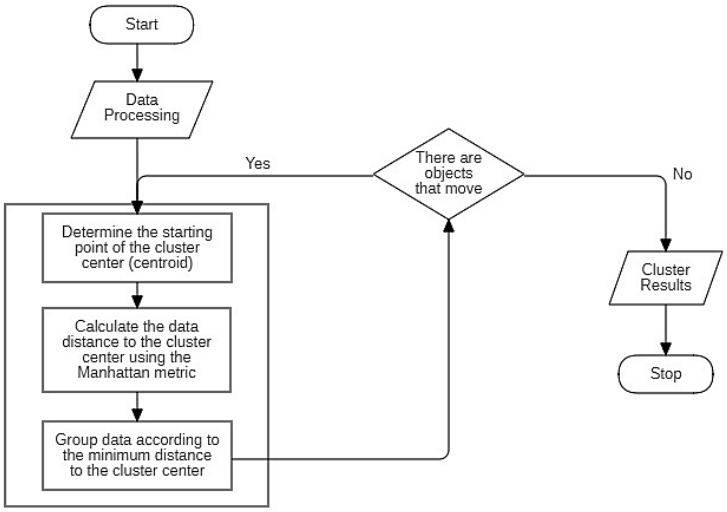

Figure 1. Flowchart algorithm k-means

2.1 Research Framework

The research framework includes a series of steps taken to find a solution to a problem. In this study, the method applied is applied research with a quantitative approach. Quantitative data refers to data that can be measured directly as a numeric variable. In this study, the data used is secondary data obtained from PUSTIPADA, State Islamic University of North Sumatra, Medan, then analyzed using Manhattan metric analysis in K-Means clustering. The following is the research framework that was carried out, which can be seen in Figure 2.

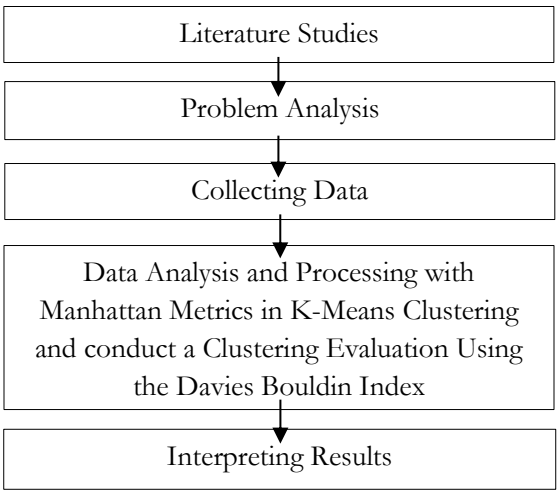

Figure 2. Research framework

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

Based on Figure 2 the details as follow.

1. library, the initial research activities carried out are to first examine the problems to be studied including the problem topic, objectives and solutions. Then, theories related to the research topic are studied from various sources such as journals and books. The theories studied in the literature are distance measurement, normalization, K-Means algorithm and other related theories [11] .
2. problem, at this stage the actual problem point will be determined and what attributes are needed to solve the problem in the data grouping process.
3. Data collection, In this study, data were collected through direct observation. The data used in this study were obtained from PUSTIPADA, State Islamic University of North Sumatra, Medan. The data collected include NIM, name, class, gender, study program, faculty, admission path, student status and nominal UKT.
4. Data analysis and processing using Manhattan metrics in K-Means clustering as shown in Equation 1, then the existing data is normalized using the z-score model, the aim is to ensure that there is a balance between the criteria attributes and that there is no mutual reduction between the criteria in the data used.

$$Z=\frac{(x-\mu)}{\sigma} \quad (1)$$

The analyzed data is then processed by applying the Manhattan Metric method to K-Means for clustering. In this study, Manhattan Metric is used in the K-Means method to find solutions to research problems. After going through the normalization and analysis stages, Manhattan Metric is re-applied in the K-Means algorithm for data clustering. Data processing is carried out using the K-Means approach where the distance calculation uses the Manhattan method, then a cluster evaluation is carried out using the Davies Bouldin Index calculation as shown in Equation 2 and 3.

$$d(p,q)=|p1-q1|+|p2-q2|+\ldots+|pn-qn|=\sum_{i=1}^{n}|pi-qi| \quad (2)$$

$$DBI=\frac{1}{k}\sum_{i=1}^{k} max_{i\neq j}(R_{ij}) \quad (3)$$

5. Interpretation of the results is done through a testing process to determine whether the resulting model or algorithm is valid and in accordance with the calculations that have been carried out. The final results are presented in the form of graphs, tables, and scatter plots.

3. RESULTS AND DISCUSSION

In this chapter, several discussions will be presented, including data analysis, data representation, and testing. The explanation is as follows.

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

3.1. Data analysis

In this study, student academic data has been collected and prepared for analysis by applying the K-Means clustering algorithm. The data used is the data of KIP scholarship recipients for the 2016-2023 period. From these data, the attributes analyzed include class, pathway, gender, study program, faculty, GPA, student status, and the amount of UKT. The K-Means method is applied by determining the appropriate number of clusters and iteratively updating the centroid position to categorize students according to similar attributes. After clustering is complete, an analysis is carried out to identify profiles and patterns in each cluster, as well as to provide in-depth interpretation of the results obtained. The results of this study are expected to provide important insights into student characteristics based on selected attributes, thereby supporting decision making in academic management and administration in higher education.

The following is an application of the K-Means method using the Manhattan Metric technique, with input parameters including the number of datasets as many as n data and the initial centroid value set at k = 4 according to the research results. This study uses the K-Means method to analyze data from 3172 students. The dataset description is in Figure 3.

```
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   NIM            3172 non-null   int64
 1   NAMA           3172 non-null   object
 2   ANGKATAN       3172 non-null   int64
 3   JALUR          3172 non-null   int64
 4   JENIS KELAMIN  3172 non-null   int64
 5   PRODI          3172 non-null   int64
 6   FAKULTAS       3172 non-null   int64
 7   IPK            3172 non-null   float64
 8   STATUS         3172 non-null   int64
 9   UKT NOMINAL    3172 non-null   int64
dtypes: float64(1), int64(8), object(1)
memory usage: 247.9+ KB
None
```

Figure 3. Data set description

After the data is obtained, the stages in the K-Means method are as follows:

1. Prepare the data set.
2. Determination of the number of clusters assigned (K) is done using the elbow technique.
3. Select a centroid point randomly.
4. Group the data to form K clusters, each of which has a centroid point determined using the following Equation 2.
5. update centroid center value using Equation 4.

$$\mu k = \frac{1}{N_k} \sum_{i=1}^{N_k} xi \quad (4)$$

Where:
μk is the centroid point of the K-th cluster.
Nk indicates the number of data in the Kth cluster.
Xi is the i data in the K cluster.

6. Repeat steps 3 through 5 until the midpoint value of the oid no longer changes.

3.2. Data Representation

The research conducted involved 3172 data to be grouped. The dataset has 8 attributes that will affect the grouping process, namely class, pathway, gender, study program, faculty, GPA, student status, and UKT amount. Furthermore, the data representation process is carried out with the aim of facilitating the grouping process. The intended dataset that will be represented for all data can be found in Table 1 below:

Table 1. Student representation data

No.	Variables	Representation		
1.	Track	SNBP		1
		SNBT		2
		SPAN-PTKIN		3
		UM-PTKIN		4
		Independent		5
2.	Gender	Man		1
		Woman		2
		Islamic Belief and Philosophy	0401	1
		Study of Religions	0402	2
		Science of the Quran and Interpretation	0403	3
		Islamic Political Thought	0404	4
		Science of Hadith	0406	6
		Islamic Economics	0501	1
		Sharia Accounting	0502	2
		Islamic Banking	0503	3
		Sharia Insurance	0505	5
		Management	0506	6
		Islamic education	0301	1
		Arabic Language Education	0302	2
		Islamic Education Counseling Guidance	0303	3
		English Teaching	0304	4
		Mathematics Education	0305	5
		Elementary Madrasah Teacher Education	0306	6
		Islamic Education Management	0307	7
		Early Childhood Islamic Education	0308	8

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

No.	Variables	Representation		
		IPS Education	0309	9
		Biology Education	0310	10
		Indonesian Language Teaching	0314	14
		Islamic Communication and Broadcasting	0101	1
		Islamic Counseling Guidance	0102	2
		Development of Islamic Society	0103	3
		Da'wah Management	0104	4
		Computer Science	0701	1
		Information Systems	0702	2
		Mathematics	0703	3
		Biology	0704	4
		Physics	0705	5
		Library Science	0601	1
		History of Islamic Civilization	0602	2
		Sociology of Religion	0604	4
		Communication Studies	0105	5
		Public Health Science	0801	1
		Nutrition	0802	2
		Family Law (Akhwal Syaksiyah)	0201	1
		Comparison of Madzhabs	0202	2
		Constitutional Law (Siyasah)	0203	3
		Sharia Economic Law (Muamalah)	0204	4
		Islamic Criminal Law (Jinayah)	0205	5
		Law	0206	6
		Sociology of Religion	0604	4
		Communication Studies	0105	5
		Public Health Science	0801	1
		Nutrition	0802	2
		Family Law (Akhwal Syaksiyah)	0201	1
		Comparison of Madzhabs	0202	2
		Constitutional Law (Siyasah)	0203	3
		Sharia Economic Law (Muamalah)	0204	4
		Islamic Criminal Law (Jinayah)	0205	5
		Law	0206	6
		Preaching and Communication	0101	1
		Sharia and Law	0201	2
		Education and Teaching Science	0301	3
4.	Faculty	Usuluddin and Islamic Studies	0401	4
		Islamic Economics and Business	0501	5
		Social Sciences	0601	6
		Science and Technology	0701	7
		Public health	0801	8
		Passed	1	
		Active	2	
5.	Status	Leave	3	
		Non-active	4	
		Exit/Dropout	5	

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

So that student data that has been in the representation stage will be continued to the normalization process. The data normalization process is shown in Table 2.

Table 2. Data normalization

No.	Nim	Force	Track	Gender	Study Program	Faculty	GPA	Status	UKT Nominal
1.	162018	2016	4	1	3	4	3.68	1	1,500,000
2.	162023	2016	4	2	3	4	3.68	1	1,500,000
3.	162027	2016	4	1	3	4	3.87	1	1,500,000
4.	162036	2016	4	2	3	4	3.66	1	1,500,000
5.	163051	2016	5	2	3	4	3.36	4	1,500,000
6.	172085	2017	5	1	3	4	3.81	1	1,787,000
7.	173086	2017	5	1	3	4	3.48	2	1,500,000
8.	173090	2017	5	2	3	4	3.72	1	1,785,000
9.	173111	2017	5	2	3	4	3.43	1	1,785,000
10.	181005	2018	3	1	3	4	3.64	1	3,316,000
-------	----------	----------	-------	----------	-------	----------	-------	----------	----------
3165	231120	2023	1	1	6	2	3.7	2	2,710,000
3166	232066	2023	2	1	6	2	3.9	2	1,708,000
3167	232117	2023	2	2	6	2	3.3	2	2,710,000
3168	233137	2023	5	1	6	2	4	2	2,710,000
3169	233154	2023	5	1	6	2	3.7	2	1,708,000
3170	233161	2023	5	1	6	2	3.9	2	1,708,000
3171	233168	2023	5	2	6	2	3.9	2	2,710,000
3172	233180	2023	5	1	6	2	3.7	2	1,708,000

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

1. Data pre-processing, there are several stages that need to be done to prepare student data before being processed by applying the K-Means clustering method. The stages in question are as follows:

1) Data cleansing, the data processing stage in cleansing to ensure the data used is error-free and ready for analysis. The steps taken include checking for duplicate data and deleting data that does not have attribute values.
2) Data normalization, data normalization is needed to change data into a format which is suitable for K-Means clustering analysis to optimize the clustering process. The normalization technique used in this study is z-score so that all attribute values in the data are in the same value range. The attributes used for the calculation are the 8 attributes mentioned earlier. The following are the results of data normalization using z-score can be seen in table 3 below.

Table 3. Normalized student data

No.	Nim	Force	Track	Gender	Study Program	Faculty	GPA	Status	UKT Nominal
1.	162018	-2.102198275	0.166351321	-1.397436662	-0.204394993	-0.019973778	0.203804121	-1.625867278	-0.702862529
2.	162023	-2.102198275	0.166351321	0.715370341	-0.204394993	-0.019973778	0.203804121	-1.625867278	-0.702862529
3.	162027	-2.102198275	0.166351321	-1.397436662	-0.204394993	-0.019973778	0.8758069	-1.625867278	-0.702862529
4.	162036	-2.102198275	0.166351321	0.715370341	-0.204394993	-0.019973778	0.133066987	-1.625867278	-0.702862529
5.	163051	-2.102198275	0.957455355	0.715370341	-0.204394993	-0.019973778	-0.927990032	4.443867135	-0.702862529
6.	172085	-1.635598812	0.957455355	-1.397436662	-0.204394993	-0.019973778	0.663595496	-1.625867278	-0.302184018
7.	173086	-1.635598812	0.957455355	-1.397436662	-0.204394993	-0.019973778	-0.503567225	0.397377526	-0.702862529
8.	173090	-1.635598812	0.957455355	0.715370341	-0.204394993	-0.019973778	0.345278391	-1.625867278	-0.304976203
9.	173111	-1.635598812	0.957455355	0.715370341	-0.204394993	-0.019973778	-0.680410061	-1.625867278	-0.304976203
10.	181005	-1.168999349	-0.624752713	-1.397436662	-0.204394993	-0.019973778	0.062329852	-1.625867278	1.83244122
-------	----------	----------	----------	----------	----------	----------	----------	----------	----------
3165	231120	1.163997967	-2.206960781	-1.397436662	1.051268503				
3166	232066	1.163997967	-1.415856747	-1.397436662	1.051268503				
3167	232117	1.163997967	-1.415856747	0.715370341	1.051268503				
3168	233137	1.163997967	0.957455355	-1.397436662	1.051268503				
3169	233154	1.163997967	0.957455355	-1.397436662	1.051268503				
3170	233161	1.163997967	0.957455355	-1.397436662	1.051268503				
3171	233168	1.163997967	0.957455355	0.715370341	1.051268503				
3172	233180	1.163997967	0.957455355	-1.397436662	1.051268503				

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

Faculty	GPA	Status	UKT Nominal
-1.009924125	0.274541256	0.397377526	0.986409242
-1.009924125	0.981912602	0.397377526	-0.412475316
-1.009924125	-1.140201436	0.397377526	0.986409242
-1.009924125	1.335598275	0.397377526	0.986409242
-1.009924125	0.274541256	0.397377526	-0.412475316
-1.009924125	0.981912602	0.397377526	-0.412475316
-1.009924125	0.981912602	0.397377526	0.986409242
-1.009924125	0.274541256	0.397377526	-0.412475316

3.3. Testing

The K-Means clustering test with 3,172 data aims to assess the effectiveness of the algorithm in grouping data into appropriate clusters. This process includes several standard methods, such as the elbow method to determine the most optimal number of clusters by finding the "elbow" point on the SSE (sum of squared errors) graph. In addition, the Davies-Bouldin index is used to assess the quality of clustering by comparing the distance between clusters. between groups with the size of the group itself, where lower values indicate better clustering. Through this test, it can be ascertained that K-Means clustering produces accurate and meaningful results in data analysis with these 3,172 data. In this study, the author used Jupyter Notebook as a dataset testing tool. The testing stages carried out include:

1) Importing libraries and loading datasets, the author started by importing the necessary libraries such as pandas, numpy, and scikit-learn, matplotlip, and loaded the dataset totaling 3,172 data into Jupyter Notebook. It can be seen in Figure 4.

```
In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        from sklearn.cluster import KMeans
        from sklearn.metrics import davies_bouldin_score
        from scipy.spatial.distance import cdist
        from sklearn.preprocessing import StandardScaler

        file_path = 'Data Mahasiswa.xlsx'
        data = pd.read_excel(file_path)

        # Memilih atribut yang relevan
        selected_columns = ['ANGKATAN', 'JALUR', 'JENIS KELAMIN', 'PRODI', 'FAKULTAS', 'IPK', 'STATUS', 'UKT NOMINAL']
        data_selected = data[selected_columns]
```

Figure 4. Importing libraries and loading datasets

2) Data preprocessing, this step includes data normalization to ensure all features are on the same scale, so that K-Means clustering can run more effectively. The following are the results of data normalization that can be seen in Figure 5.

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

	ANGKATAN	JALUR	JENIS KELAMIN	PRODI	FAKULTAS	IPK	STATUS	UKT NOMINAL
0	-2.10253	0.166378	-1.397657	-0.204427	-0.019977	0.203836	-1.626124	-0.702973
1	-2.10253	0.166378	0.715483	-0.204427	-0.019977	0.203836	-1.626124	-0.702973
2	-2.10253	0.166378	-1.397657	-0.204427	-0.019977	0.875945	-1.626124	-0.702973
3	-2.10253	0.166378	0.715483	-0.204427	-0.019977	0.133088	-1.626124	-0.702973
4	-2.10253	0.957606	0.715483	-0.204427	-0.019977	-0.928136	4.444568	-0.702973

Figure 5. Normalization results

The output of data normalization shows that all features are on the same scale. This normalization process ensures that features such as class, track, gender, study program, faculty, GPA, status and nominal UKT, which have different value ranges, are adjusted into a uniform range. Thus, the K-Means clustering algorithm can work more effectively in grouping data without the influence caused by differences in scale between features.

3) The author applies the elbow method to find the most appropriate number of clusters. This method involves plotting the SSE (sum of squared errors) against various cluster numbers and finding the "elbow" point where the SSE decline begins to slow down. The following is a graph showing the results of determining the number of clusters using the elbow method. It can be seen in Figure 6.

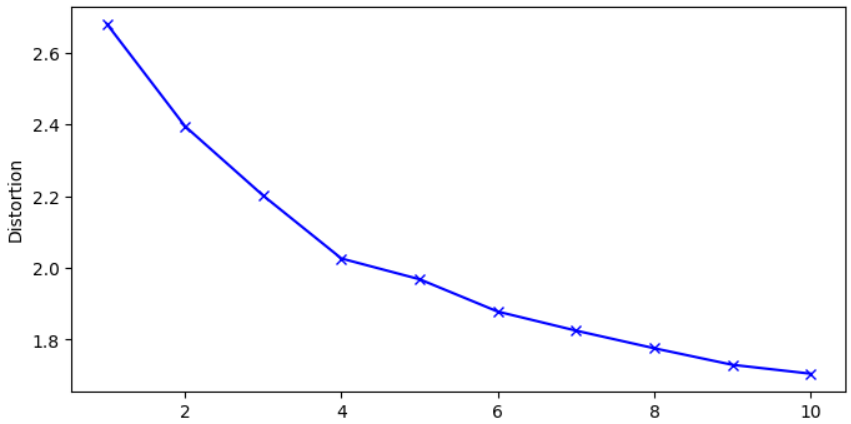

Figure 6. Elbow method graph

From the output above, the line experiences a break that forms an elbow when k=4 and the optimal k is obtained when k-4 with a value of 2.02632.

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

4) After the most appropriate number of clusters is determined, the K-Means method is used to categorize the data in a dataset into predetermined clusters. The output of the data clustering can be found in Figure 7.

	NIM	ANGKATAN	JALUR	JENIS KELAMIN	PRODI	FAKULTAS	IPK	STATUS	UKT NOMINAL	Cluster
0	403162018	2016	4	1	3	4	3.68	1	1500000	3
1	403162023	2016	4	2	3	4	3.68	1	1500000	3
2	403162027	2016	4	1	3	4	3.87	1	1500000	3
3	403162036	2016	4	2	3	4	3.66	1	1500000	3
4	403163051	2016	5	2	3	4	3.36	4	1500000	2
...	...	...	...	...	...	...	...	...	...	...
3167	206233137	2023	5	1	6	2	4.00	2	2710000	0
3168	206233154	2023	5	1	6	2	3.70	2	1708000	0
3169	206233161	2023	5	1	6	2	3.90	2	1708000	0
3170	206233168	2023	5	2	6	2	3.90	2	2710000	2
3171	206233180	2023	5	1	6	2	3.70	2	1708000	0

3172 rows × 10 columns

Figure 7. Results of data grouping using k-means

The application of the K-Means method shows that the data has been successfully grouped into four predetermined clusters. After determining the right number of clusters, the K-Means method is used to categorize the data in a dataset. From 3172 data, the final results obtained are cluster 0 with 774 data, cluster 1 with 417 data, cluster 2 with 1244 data and cluster 3 with 737 data.

Further analysis, the data grouping process obtained the results that from all the data analyzed, based on the clustering results there were 774 data classified as cluster 0. Judging from the student status, it was found that students in cluster 0 tend to have the potential to graduate late and this cluster is KIP students with a similarity of medium GPA. Late graduation ranges from 14 or 12 semesters from what should be 7 to 8 semesters. Especially students who have the potential to graduate late come from the 2016 class . Furthermore, there are 417 data classified as cluster 1. Members of this cluster show a different pattern compared to other clusters. It was found that those in cluster 1 tend to graduate on time and have a similarity of high GPA and most students who graduate on time come from the 2017 class. On the other hand, there are 1244 data classified as cluster 2. This cluster is the largest. It was found that those in cluster 2 tend to graduate late and have a similarity of low GPA. Late graduation ranges from 12 semesters from what should be 7 to 8 semesters. Specifically, the number of students who graduated late came from the 2019 class of 40 students . Finally, there are 737 data that are classified as cluster 3. It was found that cluster 3 tends to have the potential to graduate late and has a similarity of moderate GPA. Especially students who have the potential to graduate late come from the 2016 class.

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

5) To evaluate clustering performance, the authors use the Davies-Bouldin index evaluation metric to measure clustering quality. The Davies-Bouldin index measures clustering quality by comparing the distance between clusters with the distance within the cluster. The following is a graph of the Davies Bouldin Index which can be seen in Figure 8.

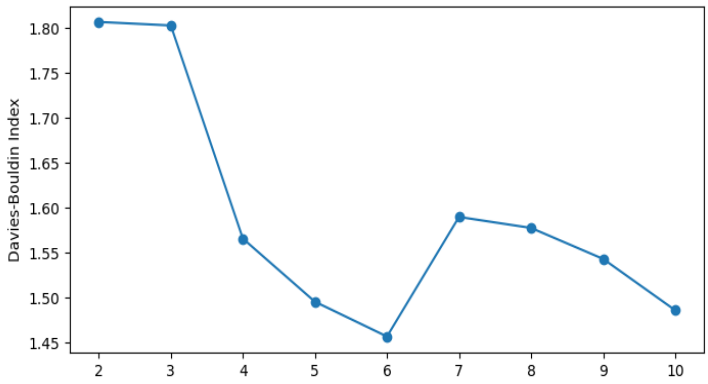

Figure 8. Davies bouldin index chart

The results of using the Davies-Bouldin Index evaluation metric show that the clustering quality is 1.4568, which is shown at k = 6. A lower Davies-Bouldin Index value indicates better clustering, where the distance between clusters is greater than the size within the cluster. With this DBI value, the author can evaluate how good the clustering results are, with a lower DBI value indicating more optimal clustering results.

6) To display the clustering results, the clustering results are visualized using scatterplots and other graphs to understand the distribution of data in each cluster and evaluate the clustering results visually. The results of the clustering visualization can be seen in Figure 9.

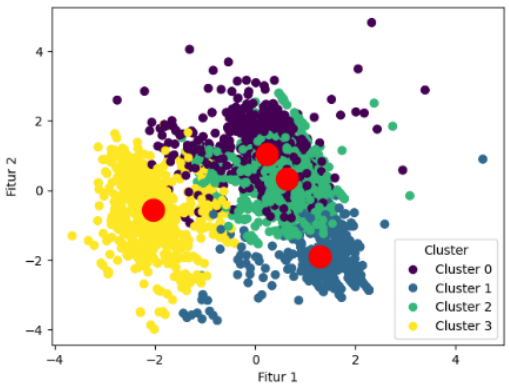

Figure 9. The clustering results are visualized using scatter plots

ISSN: **2656-5935** http://journal-isi.org/index.php/isi ISSN: **2656-4882**

After the data was divided into 4 clusters, the number of data in each cluster was 774 for cluster 0, 417 for cluster 1, 1244 for cluster 2, and 737 for cluster 3. All data analyzed in detail has been mapped into the appropriate cluster.

3.4 Discussion

The application of the K-Means clustering algorithm in this study has yielded significant insights into the academic characteristics of KIP scholarship recipients from 2016 to 2023. By analyzing 3,172 student records across eight attributes—class, pathway, gender, study program, faculty, GPA, student status, and UKT amount—the study successfully categorized the students into four distinct clusters. These clusters reveal patterns that can inform academic management and support interventions. For instance, students in Cluster 0 and Cluster 3, who tend to have moderate GPAs and are at risk of delayed graduation, may benefit from targeted academic support. In contrast, Cluster 1 includes students with high GPAs who are on track for timely graduation, suggesting that they may require different types of engagement or challenges to maintain their performance.

The process of data representation played a crucial role in ensuring the effectiveness of the clustering analysis. By converting categorical data into numerical values, the study facilitated a more accurate application of the K-Means algorithm. This step, combined with data normalization, allowed the algorithm to treat all attributes on a comparable scale, thus preventing any single attribute from disproportionately influencing the clustering results. The successful separation of clusters demonstrates the importance of these preparatory steps in achieving meaningful and interpretable outcomes.

Testing and validation were integral to assessing the quality of the clustering results. The use of the elbow method to determine the optimal number of clusters and the Davies-Bouldin Index (DBI) to evaluate clustering quality ensured that the results were both accurate and reliable. The study identified four clusters as the optimal solution, with a DBI value of 1.4568 at k=6, indicating well-separated and compact clusters. These quantitative measures not only validate the clustering but also provide a benchmark for comparing the results with other potential clustering methods or configurations.

The study's findings have practical implications for academic management, particularly in identifying students who may be at risk of delayed graduation or lower academic performance. By understanding the profiles of students within each cluster, educational institutions can design targeted interventions that address the specific needs of each group. For example, students in Cluster 2, who have low GPAs and are likely to graduate late, could benefit from academic advising, tutoring, or mentoring programs designed to improve their academic outcomes.

Similarly, the insights gained from Cluster 1 can help institutions recognize and support high-achieving students in maintaining their trajectory.

This study demonstrates the utility of K-Means clustering in analyzing student academic data and identifying meaningful patterns that can inform decision-making in higher education. The careful application of data representation, normalization, and clustering validation techniques has produced robust results that offer valuable insights into the characteristics of KIP scholarship recipients. These findings not only contribute to the academic literature but also provide a foundation for practical strategies to support student success in higher education.

4. CONCLUSION

Drawing from the data analysis and discussion, it can be concluded that student data can be effectively grouped using the Manhattan Metric technique in K-Means clustering by following key steps, including data preprocessing, determining the optimal number of clusters, executing the clustering process, and evaluating the results. The analysis identified four distinct clusters: Cluster 0, consisting of 774 students with a tendency for late graduation and moderate GPAs, predominantly from the 2016 cohort; Cluster 1, with 417 students, characterized by on-time graduation and high GPAs, mostly from the 2017 cohort; Cluster 2, the largest cluster with 1,244 students, marked by late graduation and low GPAs, particularly from the 2019 cohort; and Cluster 3, comprising 737 students with moderate GPAs and a potential for delayed graduation, also primarily from the 2016 cohort. The clustering process yielded a Davies-Bouldin Index of 1.4568, indicating reasonable clustering quality, although exploring alternative methods such as hierarchical clustering or DBSCAN, as well as different distance metrics like Euclidean, Minkowski, Chebyshev, Cosine, or Hamming, may provide improved results. Additionally, incorporating other attributes, such as more refined academic or demographic data, could further enhance the clustering outcomes.

REFERENCE

[1] M.J. Nurul Rohmawati and S. Defiyanti, "Implementation of K-Means Algorithm in Clustering Scholarship Applicants," *Jitter*, vol. 1, no. 2, pp. 62–68, 2015.

[2] Y. Heryadi and E. Irwansyah, *Deep Learning: Its Application in Geospatial Field*, PT. Artifisia Wahana Informa Teknologi, 2020.

[3] M.A. Hakim and I. N. Alam, "A Comparative Study of Clustering Techniques in Data Mining: K-Means, DBSCAN, and Hierarchical Clustering," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 12, pp. 275–283, 2018, doi: 10.14569/IJACSA.2018.091236.

[4] S.A. Rahmah and J. Antares, "Clusterization of Student Selection Candidates for Scholarship Recipients from the Foundation Using K-

Means Clustering," *INFORM.*, vol. 13, no. 2, p. 25, 2022, doi: 10.36723/juri.v13i2.282.

[5] M.S. Pangestu and M.A. Fitriani, "Comparison of Euclidean Distance, Manhattan Distance, and Cosine Similarity Calculations in Rice Seed Data Clustering Using the K-Means Algorithm," *Sainteks*, vol. 19, no. 2, p. 141, 2022, doi: 10.30595/sainteks.v19i2.14495.

[6] M. Nishom, "Comparison of Accuracy of Euclidean Distance, Minkowski Distance, and Manhattan Distance on Chi-Square Based K-Means Clustering Algorithm," *J. Inform.: J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019, doi: 10.30591/jpit.v4i1.1253.

[7] N. Syahfitri, E. Budianita, A. Nazir, and I. Afrianty, "Product Grouping Based on Inventory Data Using the Elbow and K-Medoid Methods," *KLIK Kajian Ilm. Inform. Komput.*, vol. 4, no. 3, pp. 1668–1675, 2023, doi: 10.30865/klik.v4i3.1525.

[8] R.F.N. Alifah and A.C. Fauzan, "Implementation of K-Means Clustering Algorithm Based on Manhattan Distance for Clustering of Student Field Concentration," *Ilk. J. Comput. Sci. Appl. Inform.*, vol. 5, no. 1, pp. 31–41, 2023, doi: 10.28926/ilkomnika.v5i1.542.

[9] W. Wahyu Pribadi, A. Yunus, and A.S. Wiguna, "Comparison of K-Means Euclidean Distance and Manhattan Distance Methods in Determining Covid-19 Zoning in Malang Regency," *JATI*, vol. 6, no. 2, pp. 493–500, 2022, doi: 10.36040/jati.v6i2.4808.

[10] N. Puspitasari, H. Haviluddin, and F.U.J. Helmi Puadi, "Clusterization of Pepper Plant Producing Areas Using the K-Means Algorithm," *Indones. J. Comput. Sci.*, vol. 11, no. 3, pp. 1001–1013, 2022, doi: 10.33022/ijcs.v11i3.3104.

[11] R. Syahputra and E. Bu'ulolo, "Determination of Students Receiving Single Tuition Assistance with K-NN," *J. Eksplora Inform.*, vol. 13, no. 1, pp. 46–54, 2023, doi: 10.30864/eksplora.v13i1.797.