Journal of Renewable Energy, Electrical, and Computer Engineering
Volume 5, Number 1, March 2025. 111-115
eISSN 2776-0049
Research Original Article
DOI: https://doi.org/10.29103/jreece.v5i1.18165

Clustering Level of Cigarettes Addiction Among Malikussaleh University Students Using K-Means Method

Alvin Alvesaldy✉[1], Asrianda[2], Ar Razi[3]
[1]Department of Informatics Engineering, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24355, Indonesia, alvin.190170078@mhs.unimal.ac.id
[2]Department of Informatics Engineering, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24355, Indonesia, asrianda@unimal.ac.id
[3]Department of Informatics Engineering, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24355, Indonesia, ar.razi@unimal.ac.id
✉Corresponding Author: alvin.190170078@mhs.unimal.ac.id | **Phone: +6282361301862**

Received: August 07, 2024 **Revision: January 15, 2025** **Accepted: March 10, 2025**

Abstract
Cigarettes are a form of tobacco product produced by rolling dried tobacco leaves into small cylindrical sticks. Cigarettes are usually used for smoking, namely smoking and inhaling the smoke produced when tobacco leaves are burned. Cigarettes generally contain ingredients such as tobacco leaves, which can contain nicotine, an addictive substance that causes dependence. Apart from that, cigarettes also contain various other dangerous chemicals such as tar, carbon monoxide and formaldehyde. The smoke produced when a cigarette is burned creates more than 4,000 chemicals, of which about 70 are known to cause cancer. This research aims to help students at the Faculty of Engineering, Malikussaleh University to help students find out the level of their addiction to cigarettes. This research also gave birth to a grouping system that uses the Python programming language and MySQL as the database. The K-Means Clustering algorithm used in this grouping system states that out of 200 students at the Faculty of Engineering, Malikussaleh University, 28 people are smokers who have a low level of addiction (C1), 77 people have a moderate level of addiction (C2), 55 people have a heavy level of addiction. (C3), 40 people had a very severe level of addiction (C4). This system can be used to determine the level of cigarette addiction among students at the Faculty of Engineering, Malikussaleh University in the future.
Keywords: Cigarettes, Addiction, Clustering, K-Means

Introduction
Cigarettes, a type of addictive substance, are a serious threat to individuals and society when used. This product is made from tobacco that is processed and rolled in paper, leaf, or corn husk, about the size of a pinky with the or corn husk, is about the size of a pinky with a about 8-10 centimeters long, and is generally smoked after burning the tip. the end is burnt. The number of adult smokers in Indonesia has been steadily increasing in the last decade. According to the Global Adult Tobacco Survey (GATS) 2021 issued by the Ministry of Health (MOH), there has been an increase of 8.8 million people in the number of adult smokers from 2011 (60.3 million) to 2021 (69.1 million). Although smoking prevalence in Indonesia has slightly decreased from 1.8% to 1.6%, however, if cigarette consumption remains high every year, the number of deaths caused by smoking in Indonesia will also continue to grow. The focus of this research is students study at faculty of engineering in Malikussaleh University. Malikussaleh University. Many of the students and female students of Malikussaleh University do not know the level of their addiction to cigarettes. To help Malikussaleh University students know at what level of addiction to cigarettes cigarettes, a system is built using one of the data mining classifications according to how it it works, namely clustering (Seimahuira, 2021).

Clustering is the steps to group a set of data into several groups. data into several groups or clusters, where objects in the group have significant similarities, but are significantly different from objects in other groups. strikingly different from objects in other groups. The method applied in clustering research on cigarette addiction level that will be carried out is the k-means method. KMeans is an iterative clustering algorithm. This algorithm starts by randomly selecting K, which represents the desired number of clusters. Next, the value of K is randomly assigned as the center of the cluster, also known as the centroid, by using a formula to find the closest distance Each data is classified based on its proximity to the centroid. is classified based on its proximity to the centroid. This process is repeated until the centroid value does not change (Nugraha et al., 2014).

Data mining is a group of systems that explore the values of information and complex relationships stored in a data set. By performing information pattern analysis on data, the goal is to process data into new, more useful information through the process of extracting and discovering valuable or interesting patterns contained in the database(Utomo & Purba, 2019).

Based on that background, a topic "Clustering the Level of Cigarette Addiction in Malikussaleh University Students

Using the K-Means Method" is adopted.

Literature Review

Data Mining

Data mining is an attempt to generate significant models from large-scale data, which can be stored in various forms of data storage such as databases, data warehouses, or other information stores. such as databases, data warehouses, or other information stores. The field of data mining is closely related to other disciplines, including database systems, statistics, machine learning, information retrieval, and computational computing, data warehouse, statistics, machine learning, information retrieval, and high-level computing. Data mining is used as an implementation method to patterns and models that can be used to make predictions based on historical data over a period of time based on historical data over a period of time. It is a form of data mining that is used to explore knowledge in large data sets (Wahyudi et al., 2022).

Clustering

Clustering is a component of data mining, which is the process of extraction of interesting patterns from large data. Clustering can be explained as a technique of grouping data into several clusters so that the data in one cluster has maximum similarity. There are various approaches used to explain clustering methods. The two main approaches are partition-based clustering and hierarchical clustering. Partition-based clustering, also known as partition-based clustering, is a data clustering process in which the data under study is placed into several existing clusters, without considering the hierarchy of the data. In the partition clustering method, each cluster has a centroid, and the general goal is to minimize the dissimilarity of each cluster. (dissimilarity) of each data to its respective cluster center (Dewi et al., 2022).

K-Means

K-Means is a non-hierarchical clustering method that seeks to partition data into one or more groups so that data with the characteristics of more groups so that data with are grouped together. Data that has different characteristics are grouped into different groups. This is one of the distance-based clustering algorithm that works with numerical data (Kurniawan et al., 2021).

Stages of the k-means algorithm in this study are as follows:

- Determine the number of clusters K that you want to design using the elbow method.
- Select random data from the existing data as much as the value of K to be used as the initial centroid.
- Calculate the distance between each object and the centroid of each cluster. To measure the distance between object and centroid, the distance formula can be used manhattan

$$d(x, y) = \sum_{i=1}^{n} |x_i - yi| \quad (1)$$

Description :
d = distance between x and y
x = cluster center data
y = data on the attribute
i = each data
n = number of data
xi = data at the cluster center to i
yi = data on every ith data

Min-Max Normalization

Min-Max normalization is a normalization technique technique that involves a linear transformation of the original data to create an equilibrium comparison value between the data before and after the process. Min-max Normalization is a data standardization technique that aims to place data in a smaller range, such as [-1,1] or [0,1], thus preventing the data from having values that are too large or too small which can hinder analysis. This method converts data values into a scale that is determined based on the desired minimum and maximum values (Nasution et al., 2019).

$$X^{l} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

Description :
X^1 = Attribute data to be normalized
X_{min} = The smallest value of the attribute
X_{max} = The largest value of the attribute

Python

Python is a programming language that executes a number of multipurpose instructions directly, with an object-oriented approach. The language provides commands that make theory testing fast and simple. With a variety of easy-to-use libraries, users can select and enjoy the functionality provided with a simple interface. provided with a simple interface. This makes it easy for users to understand basic concepts such as repetition and branching. Nonetheless, the end result in using Python largely depends on one's motivation and willingness to learn programming and science. to learn programming and science (Rizal et al., 2021).

Materials & Methods

Research on Clustering the Level of Cigarette Addiction at the Faculty of Engineering Malikussaleh University Students Using the K-Means Method began in the even semester of March 2023 to July 2023. The location of this research is at the Faculty of Engineering, Malikussaleh University.

In the initial process of the K-Means method, the first step that must be taken is to input the questionnaire data on the level of cigarette addiction in students of the Faculty of Engineering, Malikussaleh University, determine the number of clusters, weight rank, maximum iteration, and the smallest desired error, determine the initial centroid based on the number of clusters to be formed, calculate the centroid value with the Manhattan formula, grouping based on the closest distance, if there is data transferred back to the initial centroid determination process. Determine the initial centroid, but if there is none, it will display the clustering results.

Results and Discussion

The data used is data on Malikussaleh University Faculty of Engineering students who filled out the questionnaire as many as 200 people. In the questionnaire there are 5 questions containing the reasons for smoking, the number of cigarettes, smoking time lag, smoking stages, and smoking cessation efforts.

After all the required data is collected, the author then performs the calculation process applying the K-Means algorithm using the Manhattan distance formula showing the results that cluster 1 has 28 data, cluster 2 has 77 data, cluster 3 has 55 data, and cluster 4 has 40 data.

Table 1. Dataset

No	Name	Major	P1	P2	P3	P4	P5
1	Aprian G. S	T. Informatika	2	7	1	1	1
2	M. Akbar	T. Informatika	2	20	5	4	4
3	Rahmad. R.R	Akuntansi	5	16	2	2	2
4	Ananda	T. Informatika	2	15	5	4	1
	……	…….	……	…...	….	……	……
200	Melly	T. Informatika	3	2	1	2	1

Description :

P1 = Reason for Smoking
P2 = Number of Cigarettes
P3 = Time Pause
P4 = Smoking Stages
P5 = Smoking Cessation Efforts

Testing Data

Determination of the center (centroid) of the initial cluster as test data as many as 4 centroids done randomly can be seen in the table.

Table 2. Testing Data

No	Name	Major	P1	P2	P3	P4	P5
25	Afif A. R.	T. Elektro	4	32	5	4	1
50	M. Qusyairi	T. Informatika	2	5	2	2	2
100	Aridho P	T. Informatika	2	24	1	4	1
150	Rifqi F. D.	T. Elektro	2	24	5	4	1

Implementation of Min-Max Normalization

The results of the application of min-max are as follows normalization in this study.

Table 3. Implementation of Min-Max Normalization

No	Name	P1	P2	P3	P4	P5
1	Aprian G. S	0,25	0,1935483	0	0	0
2	M. Akbar	0,25	0,6129032	1	1	1
3	Rahmad. R.R	1	0,4838709	0,25	0,333	0,333
4	Ananda	0,25	0,4516129	1	1	0
	……	……	…...	….	……	……
200	Melly	0,5	0,0322580	0	0.333	0

Implementation Elbow Method

Elbow method is a technique used in cluster analysis, especially in the K-Means algorithm, to determine the optimal number of clusters. The aim is to find the sweet spot between increasing the number of clusters and the complexity of the resulting model.

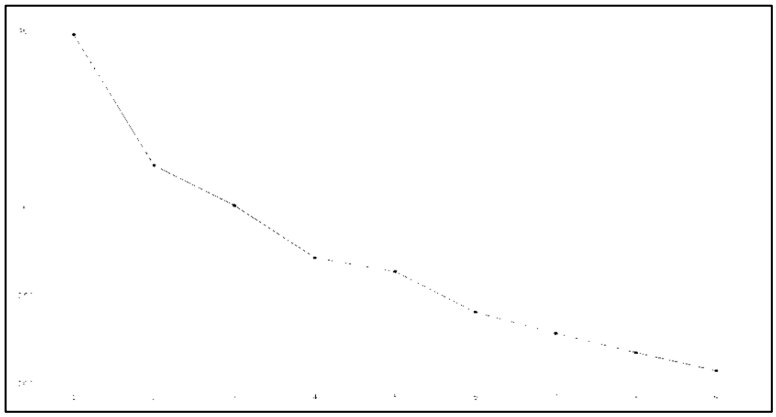

Figure 1. Elbow Method Graphics

Based on the graph above, for an explanation of the graph above can be seen in the following table.

Table 4. Result Table

Cluster	WCSS
1	252.8365053763442
2	198.15646398811978
3	182.42851992817677
4	161.2957006702975
5	157.06031656201424
6	145.4250570360228
7	138.83644358936246
8	134.4939117487307
9	131.34702818404006

Based on the results of the table above, it can be seen that the cluster suitable for this case is 4 clusters, because the results of the graph on the value of cluster 4 have an insignificant decrease.

K-Means Clustering Implementation

After calculating using the k-means clustering method, the author gets the results by iterating 10 times. Here are the results of calculations using the k-means method clustering method.

Table 5. K-Means Clustering Implementation

No	Name	C1	C2	C3	C4	Cluster
1	Aprian G. S	2,327956	0,653087	1,345843	2,348886	C2
2	M. Akbar	2,127112	3,113041	2,102357	1,165706	C4
3	Rahmad. R.R	0,998079	1,451957	1,764991	1,905721	C1
4	Ananda	1,584869	2,088502	1,351323	0,721774	C4
	……	……	…..	….	……	……
200	Melly	1,905913	0,452646	1,394954	2,426843	C2

The results of the tenth calculation show that the centroid values are the same and the calculation calculation is finished here.

The results of the new centroid calculation can be in the following table:

Table 6. Final Centroid

Centroid	P1	P2	P3	P4	P5
C1	0,767857143	0,426267281	0,410714286	0,857142857	0,30952381
C2	0,350649351	0,118977796	0,11038961	0,294372294	0,069264069
C3	0,263636364	0,272140762	0,145454545	0,896969697	0,212121212
C4	0,2	0,494354839	0,84375	0,866666667	0,325

Conclusions

The design of the clustering system for the level of cigarette addiction in students of the Faculty of Engineering, Malikussaleh University can be made in the form of use case diagrams, class diagrams, and activity diagrams. The results of the clustering system for the level of cigarette addiction in students of the Faculty of Engineering, Malikussaleh University can be implemented into a system that uses Python as a programming language and MySQL as a database.

Data from 200 Malikussaleh University students, grouped into 4 clusters, where cluster 1 amounted to 28 data, cluster 2 amounted to 77 data, cluster 3 amounted to 55 data, cluster 4 amounted to 40 data.

- In cluster 1, it can be seen that students who have a level of cigarette addiction "low" as much as 28 data.
- In cluster 2, it can be seen that students who have a "medium" level of cigarette addiction as much as 77 data.
- In cluster 3, it can be seen that students who have a "heavy" cigarette addiction level as much as 55 data.
- In cluster 4, it can be seen that students who have a cigarette addiction level of "very heavy" as much as 40 data.

Acknowledgments

Daddy and Mommy, Faisal Malay and Vera Yanti Yusnita, thank you for your prayers, encouragement, motivation, sacrifice, advice, and love that never stops until now, Partner in life Manisha Oktavia Anggraini S., Thank you for your presence that makes me very happy, my best friends for life, Muhammad Ody Alwy Siregar, Gazza Mahardika Hartoyo, thank you for making my life more colorful, my overseas brothers, M. Reza Fahlevi, M. Akbar Al Ariq, Afif Aulia Rauf, Yoga Harlis Irawan, thank you for teaching me the true meaning of bloodless brother, Late Brother, Syibbran Mulaesyi, Auza Aulia, Wira Yudha Raharja, Dimas Sari Afriza Lubis, Khairul Muzakkir Alwy Lubis, Ananda, Poor Muhammad Fajar, Taufik Habib Ansyari Siregar, Muhammad Iqbal, Andre Maulana, M. Akbar Husein Siregar, M. Bayu Juhri thank you for your knowledge and funny behavior which is very useful.

References

Dewi, F. P., Aryni, P. S., & Umaidah, Y. (2022). Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *7*(2), 111–121.

Kurniawan, R., Suhada, S., & Dewi, R. (2021). Penerapan Algoritma K-Means Clustering Dalam Persentase Merokok Pada Penduduk Umur Di Atas 15 Tahun Menurut Provinsi. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *2*(2), 178–186.

Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan normalisasi data untuk klasifikasi wine menggunakan algoritma K-NN. *Comput. Eng. Sci. Syst. J*, *4*(1), 78.

Nugraha, D. D. C., Naimah, Z., Fahmi, M., & Setiani, N. (2014). Klasterisasi Judul Buku dengan Menggun a kan Metode K-Means. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*.

Rizal, A. A., Kharisma, L. P. I., & Fahrurrozi, F. (2021). Peningkatan Efektifitas Programming dengan Pelatihan Python for Data Science Bagi Komunitas Programming Pondok Pesantren Nahdlatul Wathan Anjani. *Jurnal Widya Laksmi: Jurnal Pengabdian Kepada Masyarakat*, *1*(1), 13–19.

Seimahuira, S. (2021). Implementasi datamining dalam menentukan destinasi unggulan berdasarkan online reviews tripadvisor menggunakan algoritma K-Means. *Technologia: Jurnal Ilmiah*, *12*(1), 53–58.

Utomo, D. P., & Purba, B. (2019). Penerapan datamining pada data gempa bumi terhadap potensi tsunami di Indonesia. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, *1*, 846–853.

Wahyudi, A. K., Azizah, N., & Saputro, H. (2022). Data Mining Klasifikasi Kepribadian Siswa SMP Negeri 5 Jepara Menggunakan Metode Decision Tree Algoritma C4. 5. *Journal of Information System and Computer*, *2*(2), 8–13.