



ANALISIS KESEHATAN MENTAL BERDASARKAN PREFERENSI MUSIK MENGGUNAKAN K-MEANS CLUSTERING

Maryanti ^{1*}, Wardianto ²

^{1,2} Informatika, Fakultas Sains dan Teknologi, Universitas Nurul Huda, Indonesia

Article Info

Article History:

Received: 10-June-2026

Revised: 30-June-2026

Published: 12-July-2026

Keywords:

Health Studies,
K-Means Clustering,
Mental Health,
Music Preference.

ABSTRACT

A This study aims to identify patterns of mental health conditions based on music preferences using the K-Means Clustering algorithm. The research employed secondary data obtained from the Music & Mental Health Survey Results dataset available on Kaggle. The initial dataset consisted of 736 records and was processed through data preprocessing, including missing value removal, attribute selection, data encoding, and normalization. After preprocessing, 616 valid records with six variables (Fav Genre, Hours per Day, Anxiety, Depression, Insomnia, and Music Effects) were analyzed. The K-Means algorithm was implemented with three clusters to classify respondents according to similarities in their mental health characteristics. The results indicate that the algorithm successfully grouped respondents into three distinct clusters representing low, moderate, and high levels of mental health conditions. The second cluster showed the lowest average scores for anxiety, depression, and insomnia, while the third cluster exhibited the highest scores. These findings demonstrate that K-Means Clustering is capable of identifying patterns between music preferences, listening habits, and mental health conditions. The results may serve as preliminary information for future data-driven mental health studies.

Kata Kunci:

Kesehatan Mental,
Pengelompokan K-Means,
Preferensi Musik,
Studi Kesehatan.

ABSTRAK

Penelitian ini bertujuan untuk mengidentifikasi pola kondisi kesehatan mental berdasarkan preferensi musik menggunakan algoritma K-Means Clustering. Penelitian ini menggunakan data sekunder yang diperoleh dari dataset Music & Mental Health Survey Results yang tersedia di Kaggle. Dataset awal terdiri dari 736 catatan dan diproses melalui pra-pemrosesan data, termasuk penghapusan nilai yang hilang, pemilihan atribut, pengkodean data, dan normalisasi. Setelah pra-pemrosesan, 616 catatan valid dengan enam variabel (Genre Favorit, Jam per Hari, Kecemasan, Depresi, Insomnia, dan Efek Musik) dianalisis. Algoritma K-Means diimplementasikan dengan tiga kluster untuk mengklasifikasikan responden berdasarkan kesamaan karakteristik kesehatan mental mereka. Hasil menunjukkan bahwa algoritma berhasil mengelompokkan responden ke dalam tiga kluster berbeda yang mewakili tingkat kondisi kesehatan mental rendah, sedang, dan tinggi. Kluster kedua menunjukkan skor rata-rata terendah untuk kecemasan, depresi, dan insomnia, sedangkan kluster ketiga menunjukkan skor tertinggi. Temuan ini menunjukkan bahwa K-Means Clustering mampu mengidentifikasi pola antara preferensi musik, kebiasaan mendengarkan, dan kondisi kesehatan mental. Hasil ini dapat berfungsi sebagai informasi awal untuk studi kesehatan mental berbasis data di masa mendatang.

Corresponding Author:

Maryanti

Informatika, Fakultas Sains dan Teknologi, Indonesia

Universitas Nurul Huda

Email: maryanti.12@gmail.com

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license



1. INTRODUCTION

Current advances in information technology have driven increased data utilization in various sectors, including healthcare. The ever-increasing volume of data can be processed into valuable information through the application of data mining techniques [1]. This technique enables the storage of complex data, allowing it to be used as a basis for decision-making [2]. One widely used method in data mining is clustering, a technique for grouping data based on similar characteristics without using specific labels [3]. This method allows data to be classified into groups, making certain patterns easier to identify. One commonly used clustering algorithm is K-Means Clustering, due to its simple process and its ability to effectively group data based on the proximity of values between data points [4]. In healthcare, particularly mental health, data analysis plays a crucial role in understanding individual conditions. Mental health is one aspect that affects a person's quality of life, where disorders such as anxiety, depression, and insomnia can have a significant impact [5]. Therefore, a data-driven approach is needed to identify patterns and factors that influence mental health conditions [6].

Several previous studies have applied the K-Means Clustering method to health data analysis. The results show that this algorithm is capable of grouping data based on specific indicators such as stress levels, depression, and sleep quality, thus helping to understand an individual's psychological state in a more structured manner [7]. Furthermore, other studies have also shown that clustering techniques are effective in identifying patterns in complex health data [8]. However, these studies generally focus on clinical aspects and have not considered behavioral factors, such as music preferences. Music not only serves as a means of entertainment but also influences a person's psychological state, helping to reduce stress, improve mood, and provide a relaxing effect. Differences in music preferences among individuals can reflect different emotional states, thus offering potential use as an indicator in mental health analysis [9].

Although various studies have applied the K-Means Clustering method to analyze mental health data, most of them focus on psychological aspects and clinical indicators, while research linking music preferences to mental health conditions through a clustering approach is relatively limited. Therefore, this study aims to identify mental health patterns based on music preferences using the K-Means Clustering method by utilizing the variables Favorite Genre, Hours per Day, Anxiety, Depression, Insomnia, and Music Effects. Through this data clustering, it is hoped that information can be obtained regarding the characteristics of respondents' mental health based on their habits and preferences in listening to music.

2. RESEARCH METHOD

This research uses a data mining approach with the K-Means Clustering method. K-Means is an algorithm used to group data based on the level of similarity of characteristics into several clusters without any initial labels (unsupervised learning)[10]. This method is used because it is able to find hidden patterns in data simply and effectively.

2.1. Data Collection and Preprocessing

Describe the chronology of the research, including the research design, research procedures (in the form of algorithms, pseudocode, or other), testing methods, and data acquisition [1], [3], [6]. The description of the research process must be supported by references so that the explanation is scientifically acceptable [2]. The research methods for completing this research consist of concept design, material collection, production and rendering, testing and revision, and presentation and publication.

The data used in this study is secondary data obtained from the Kaggle platform, entitled "Music & Mental Health Survey Results." This dataset contains information on respondents' music listening habits and mental health conditions. This data was used to analyze the relationship between music preferences and mental health using the K-Means Clustering method [11].

Before the clustering process, data preprocessing was performed to improve data quality and consistency. The preprocessing process included checking for missing values, removing blank data, removing duplicate data, and selecting attributes relevant to the study. Preprocessing is a crucial step in data mining to improve data quality before analysis [12]. Furthermore, data transformation was performed to enable use in the K-Means algorithm. This involved converting categorical data into numerical form (encoding) and normalizing the numerical data to maintain a uniform scale and prevent oversampling the clustering results [13].

2.2. Research Stages

This research was conducted systematically and structured, consisting of several main stages: data sources, data processing, data transformation, data analysis, and reporting of results. Each stage was carried out sequentially to obtain analytical results that align with the research objectives using the K-Means Clustering method.

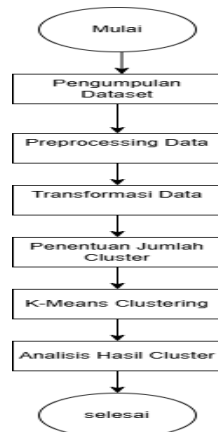


Figure 1. Research Stages

a. Data Source

This research uses secondary data obtained from the Kaggle platform with the title Music & Mental Health Survey Results. The dataset can be accessed via <https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results>. This data is widely used in data mining research because it is open and easy to access for analysis

b. Data Process

This stage begins with understanding the structure of the dataset which consists of 736 data with 33 attributes. Next, missing values were checked, then incomplete data was deleted and attributes were selected that were relevant to the research. After the cleaning process, the amount of data became 616 with 6 main variables, namely Fav genre, Hours per day, Anxiety, Depression, Insomnia, and Music effects. This process is carried out using Python with the pandas and numpy libraries. Data cleaning was carried out to improve data quality before analysis

c. Data Transformation

After the data is clean, data transformation is carried out by changing categorical variables to numeric using an encoding process, as well as normalizing them so that all variables have the same scale. This is done so that the distance calculation in the K-Means algorithm is not biased towards certain variables. Data transformation is needed so that the algorithm can work optimally

d. Data Analysis

At this stage a clustering process is carried out using the K-Means method. The steps taken include determining the number of clusters (K), initializing the initial centroid, calculating the distance of each data to the centroid, grouping data based on the closest distance, and iterating until the centroid does not change significantly. This method is used to group data based on similarities in patterns of music listening habits and the mental health conditions of respondents

2.3. K-Means Clustering Method

K-Means Clustering is a method in data mining that is used to group data into several clusters based on the level of similarity between data. This method is included in the unsupervised learning technique because it does not use labels on the data [14]. The process begins by determining the number of clusters (K), then selecting the initial centroid as the cluster center. Next, each data is calculated for its distance to the centroid using Euclidean Distance and placed in the closest cluster. This process is repeated until the centroid does not change significantly [15]. Euclidean Distance Formula:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Information:

$d(x, y)$ = distance between data and centroid

x_i = data value in attribute i

y_i = centroid value on the i-th attribute

n = number of attributes

This study used a K value of 3 to generate three groups of mental health levels: low, medium, and high. The clustering process used the attributes Favorite Genre, Hours per Day, Anxiety, Depression, Insomnia, and Music Effects to identify mental health patterns based on respondents' music preferences. This section also explains the research stages, along with the results obtained at each stage.

3. RESULTS AND ANALYSIS

3.1. Data Preprocessing Results

The preprocessing stage was carried out to ensure data quality before clustering modeling was performed. The initial dataset consisting of 736 entries with 33 attributes was processed through the elimination of incomplete data (missing values) and the selection of relevant attributes [16]. The final cleaning result produced 616 valid entries with 6 main variables: Favorite genre, Hours per day, Anxiety, Depression, Insomnia, and Music effects. This entire process was implemented using the Python programming language with the pandas and numpy libraries. To optimize the algorithm's performance, categorical variables were transformed through label encoding, while numeric data were normalized using the Min-Max Scaler into a scale range of 0–1 to avoid bias in the Euclidean distance calculation.

3.2. Results of K-Means Clustering Implementation

This study set the number of clusters at K=3 to categorize respondents' mental health profiles into low, medium, and high levels. The algorithm iterated until the cluster centroids reached a stable point of convergence. The average values of the variables in each group are presented in Table 3.1 below:

Table 1. Average Cluster Results

Variabel	Cluster 0 (Sedang)	Cluster 1 (Rendah)	Cluster 2 (Tinggi)
Fav Genre (Encoded)	2.86	10.22	12.49
Jam per Hari	4.08	3.36	3.82
Anxiety	6.76	3.76	7.45
Depression	6.17	2.14	6.85
Insomnia	4.53	1.81	5.33
Music Effects (Encoded)	0.17	0.30	0.30

3.3. Discussion

The results of applying the K-Means Clustering algorithm with the number of clusters (K) = 3 resulted in grouping respondents based on six variables: Favorite Genre (Encoded), Hours per Day, Anxiety, Depression, Insomnia, and Music Effects (Encoded). This grouping demonstrated variation in characteristics between clusters based on the average values of each analyzed variable.

Cluster 1 was characterized by relatively higher average values for Anxiety (6.76), Depression (6.17), and Insomnia (4.53) compared to Cluster 2, and a music listening intensity of 4.08 hours per day, the highest among all clusters. This indicates that respondents in this group have a relatively high music listening activity, but still have varying mental health conditions based on the indicators used. Furthermore, the low Music Effects value (0.17) indicates that the perceived influence of music in this group tends to be limited.

Cluster 2 showed the lowest average scores for Anxiety (3.76), Depression (2.14), and Insomnia (1.81) compared to the other clusters. This indicates that respondents in this group tended to have more stable mental health. The frequency of music listening was 3.36 hours per day, with a higher Music Effects score (0.30) than Cluster 1, indicating that music had a more positive influence on emotional well-being in this group.

Meanwhile, Cluster 3 showed the highest average scores for Anxiety (7.45), Depression (6.85), and Insomnia (5.33) among all clusters. This indicates that this group had relatively lower mental health compared to the other groups. Despite the relatively high frequency of music listening (3.82 hours per day), the Music Effects score (0.30) was the same as Cluster 2, indicating that the influence of music did not significantly affect emotional well-

being in this group. Differences in Favorite Genre (Encoded) scores also indicate variations in music preferences within the clusters.

Overall, the clustering results demonstrate that the K-Means Clustering method is capable of grouping data based on similar characteristics of the analyzed variables. This finding indicates a pattern of association between music preferences, music listening intensity, and respondents' mental health conditions. However, these results are exploratory and do not indicate a causal relationship, but rather a pattern of association between the variables.

3.4. Evaluation of Clustering Results

The clustering results were evaluated to determine the quality of the clusters generated by the K-Means Clustering algorithm. This study used three metrics: the Silhouette Score, the Davies-Bouldin Index (DBI), and the Calinski-Harabasz Index. The results yielded a Silhouette Score of 0.229, a Davies-Bouldin Index of 1.451, and a Calinski-Harabasz Index of 181.898. These three values are used as indicators to assess the level of compactness of the data within each cluster and the degree of separation between the clusters.

A Silhouette Score of 0.229 indicates that the clustering results have formed three fairly representative clusters, although some data still share characteristics with other clusters. Furthermore, a Davies-Bouldin Index of 1.451 indicates that the quality of separation between clusters is considered quite good, with each cluster exhibiting a sufficient level of internal similarity. On the other hand, the Calinski-Harabasz Index value of 181.898 indicates that the variation between clusters is greater than the variation within clusters, indicating that the resulting clustering structure is able to distinguish the characteristics of each respondent group.

Overall, the evaluation results indicate that the K-Means Clustering algorithm can be used to group respondent data based on similar characteristics. These results support the clustering obtained in this study, providing a picture of mental health patterns based on music preferences. However, cluster quality can still be improved by selecting a more optimal number of clusters or by adding other relevant variables to improve clustering results in future research.

4. CONCLUSION

Based on the research results, the K-Means Clustering method was successfully applied to group respondents' mental health conditions based on music preferences using the variables Favorite Genre, Hours per Day, Anxiety, Depression, Insomnia, and Music Effects. The clustering process produced three groups with different characteristics, thus providing an overview of variations in mental health conditions based on music listening habits. The evaluation results using the Silhouette Score of 0.229, Davies-Bouldin Index of 1.451, and Calinski-Harabasz Index of 181.898 indicate that the quality of the resulting clustering is in the fairly good category and is able to represent the characteristics of the data. Thus, the K-Means Clustering method can be used as an approach to identify patterns of relationships between music preferences and mental health conditions, although further research is recommended to optimize the number of clusters and add other variables to improve the quality of the clustering results.

REFERENCE

- [1] F. Sari, M. Kuntari, W. Yati, H. Khaulasari, dan M. Hafiyusholeh, "Implementasi K-Means Clustering Melalui Pemanfaatan Sampling Kombinasi Pada Pengelompokan Pola Kesehatan Mental Mahasiswa Sains dan Teknologi," *J. Nas. Teknol. Dan Sist. Inf.*, vol. 11, no. 1, hlm. 9–16, Apr 2025, doi: 10.25077/TEKNOSI.v11i01.2025.9-16.
- [2] D. Ayu Ramadhanty, Renita Syafitri, Errissya Rasywir, dan Despita Meisak, "Implementasi Data Mining Untuk Menentukan Persediaan Stok Obat Di Apotek K-24 Menggunakan Metode K-Means Clustering," *J. Inform. Dan Rekayasa KomputerJAKAKOM*, vol. 2, no. 1, hlm. 155–160, Apr 2022, doi: 10.33998/jakakom.2022.2.1.31.
- [3] E. D. Putra, M. H. Rifqo, D. Deslianti, dan K. Krismiyan, "Analysis of The Theme Clustering Algorithm Using K-Means Method," *J. Komput. Inf. Dan Teknol.*, vol. 2, no. 2, hlm. 431–442, Des 2022, doi: 10.53697/jkomitek.v2i2.884.
- [4] "Analisis Cluster Kualitas Pemuda di Indonesia pada Tahun 2022 dengan Agglomerative Hierarchical dan K-Means | Komputika : Jurnal Sistem Komputer." Diakses: 2 Mei 2026. [Daring]. Tersedia pada: <https://ojs.unikom.ac.id/index.php/komputika/article/view/10348>
- [5] "Implementasi K-Means Clustering Melalui Pemanfaatan Sampling Kombinasi Pada Pengelompokan Pola Kesehatan Mental Mahasiswa Sains dan Teknologi | Jurnal Nasional Teknologi dan Sistem Informasi." Diakses: 2 Mei 2026. [Daring]. Tersedia pada: https://teknosi.fti.unand.ac.id/index.php/teknosi/article/view/2884?utm_source=chatgpt.com

- [6] M. B. Fajri dan S. D. Purnamasari, "Klasterisasi Pola Penyebaran Penyakit Pasien Berdasarkan Usia Pasien Menggunakan K-Means Clustering," *J. Inf. Technol. Ampera*, vol. 3, no. 3, hlm. 317–334, Des 2022, doi: 10.51519/journalita.volume3.issue3.year2022.page317-334.
- [7] J. Parhusip, J. Feranita, M. P. Mentari, T. Apriliani, dan A. M. Jessend, "Implementasi Algoritma K-Means Clustering Untuk Segmentasi Tingkat Kesehatan Mental Mahasiswa Berdasarkan Faktor Stress Dan Akademik," *J. Media Inform.*, vol. 7, no. 1, hlm. 449–458, Feb 2026, doi: 10.55338/jumin.v7i1.7856.
- [8] A. Nugroho dan R. A. Putri, "Clustering YouTube Comments on Mental Health in Indonesia Using the K-Means Algorithm," *J. Ris. Komput. JURIKOM*, vol. 13, no. 2, hlm. 537–544, Apr 2026, doi: 10.30865/jurikom.v13i2.9552.
- [9] R. D. Fitri, "PENGARUH MUSIK TERHADAP MOOD DAN EMOSI PERAN MUSIK DALAM KESEHATAN MENTAL," *Circ. Arch.*, vol. 1, no. 4, Mei 2024, Diakses: 2 Mei 2026. [Daring]. Tersedia pada: <https://circle-archive.com/index.php/carc/article/view/103>
- [10] B. Hartono, S. Eniyati, dan K. Hadiono, "Perbandingan Metode Perhitungan Jarak pada Nilai Centroid dan Pengelompokan Data Menggunakan K-Means Clustering," *J. Sist. Komput. Dan Inform. JSON*, vol. 4, no. 3, hlm. 503, Mar 2023, doi: 10.30865/json.v4i3.6021.
- [11] Qomariyah dan M. U. Siregar, "Comparative Study of K-Means Clustering Algorithm and K-Medoids Clustering in Student Data Clustering," *JISKA J. Inform. Sunan Kalijaga*, vol. 7, no. 2, hlm. 91–99, Mei 2022, doi: 10.14421/jiska.2022.7.2.91-99.
- [12] Y. M. Geasela dkk., "Analysis of Student Mental Health Dataset Using Mining Techniques," *J. Comput. Sci.*, vol. 20, no. 1, hlm. 121–128, Jan 2024, doi: 10.3844/jcssp.2024.121.128.
- [13] K. Wang, S. Gao, dan J. Huang, "Learning About Your Mental Health From Your Playlist? Investigating the Correlation Between Music Preference and Mental Health of College Students," *Front. Psychol.*, vol. 13, hlm. 824789, Apr 2022, doi: 10.3389/fpsyg.2022.824789.
- [14] N. A. Maori dan E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Simetris J. Tek. Mesin Elektro Dan Ilmu Komput.*, vol. 14, no. 2, hlm. 277–288, Nov 2023, doi: 10.24176/simet.v14i2.9630.
- [15] L. R. Iyohu, Ismail Djakaria, dan La Ode Nashar, "Perbandingan Metode K-Means Clustering dengan Self-Organizing Maps (SOM) untuk Pengelompokan Provinsi di Indonesia Berdasarkan Data Potensi Desa," *J. Stat. Dan Apl.*, vol. 7, no. 2, hlm. 195–206, Des 2023, doi: 10.21009/JSA.07208.
- [16] F. A. Fitrony and F. Pereira, "Multiclass Classification of Brain Tumors in MRI Images Based on Deep Learning," *J. Kesehat. dan Sci.*, vol. 2, no. 1, pp. 38–46, 2026.