Joinforcs: Journal of informatics and information system
Volume 1, (1) 2026, p. 38-43
E-ISSN : XXXX-XXXX

Clustering Analysis of Kepler Exoplanet Data Using the Simple K-Means Algorithm

Nanda Rahma Anggyta [a,1,*], Indi Najwa Alifia [a,2], Putri Alfiya [a,3]
[a] STMIK AMIKOM Surakarta, Kartasura, Sukoharjo, 57163, Indonesia
[1] nanda.10449@mhs.amikomsolo.ac.id , [2] indi.10456@mhs.amikomsolo.ac.id , [3] putri.10459@mhs.amikomsolo.ac.id
* corresponding author

Article Info

Article History:

Received Dec 30, 2025
Revised Jan 23, 2026
Accepted Feb 06, 2026

Keywords :
Clustering, Data Mining, Eksoplanet, Simple K-Means, Teleskop Kepler, WEKA

Abstract

The discovery of exoplanets has generated a massive volume of astronomical data, requiring efficient analytical methods to detect hidden patterns and features. One approach that can be implemented is clustering using data mining techniques. This study aims to examine the clustering process of exoplanet data obtained from the Kepler Telescope using the Simple K-Means algorithm. The data analyzed in this research were sourced from the Kepler exoplanet dataset and processed using the WEKA software. The research stages included data selection, preprocessing, clustering implementation, and evaluation of clustering results. The clustering process was carried out using the Simple K-Means algorithm with several predefined clusters to identify similarities in exoplanet characteristics. The results indicate that the algorithm can effectively group exoplanet data based on similar characteristics and produce satisfactory cluster distributions with a minimal error rate. In addition, the clustering visualization successfully illustrated the distribution patterns of exoplanet data within each cluster. These findings suggest that Simple K-Means can be efficiently applied for exoplanet data exploration and support the process of astronomical data analysis. This study is expected to contribute to the development of data mining applications in the field of astronomy, particularly in exoplanet data analysis.

1. Introduction

Advances in contemporary astronomical technology have increased the volume of space observation data, particularly in the study of exoplanets. Exoplanets are planets located outside our solar system and are a major focus of research in astronomy because they can provide insights into the characteristics of planetary systems in the universe [1]. One of the largest projects in the search for exoplanets is NASA's Kepler Telescope, which has successfully collected massive amounts of astronomical data through continuous observations [2]. The sheer volume of data generated poses challenges in the analysis process because the data is complex and high-dimensional, requiring effective and efficient data processing methods.

A commonly used approach for analyzing astronomical data is data mining. Data mining is the process of extracting information and hidden patterns from large volumes of data using specific computational techniques [3]. In the field of astronomy, data mining can help researchers identify distribution patterns, the properties of celestial objects, and relationships between data that are difficult to identify manually [4]. One common technique in data mining is clustering. Clustering is a method of grouping data based on similar characteristics, so that data within a single group share higher similarities compared to data in other groups[5][6].

One popular and widely used clustering technique is K-Means Clustering. It works by dividing the data into a number of clusters based on the shortest distance to each cluster's centroid [7]. The implementation often chosen is Simple K-Means because its computational process is simple, fast, and efficient for clustering high-dimensional data [8]. Furthermore, this algorithm can generate data scatter plots that facilitate the exploration and interpretation of patterns in astronomical data [9].

Joinforcs: Journal of informatics and information system
Volume 1, (1) 2026, p. 38-43
E-ISSN : XXXX-XXXX

Various previous studies indicate that clustering techniques can be efficiently applied in many fields, such as healthcare, business, and astronomical data analysis. Kurniawan et al. [10] found that the K-Means algorithm can produce satisfactory data clustering based on the similarity of data characteristics. Research by Rahman and Putra [11] notes that clustering is effective for analyzing large datasets because it can automatically detect hidden patterns. Additionally, the international study by Mishra et al. (2023) [12] demonstrates that clustering algorithms perform well in high-dimensional astronomical data analysis. Alkhateeb et al. (2024) [13] explain that the application of data mining techniques to astronomical datasets helps identify patterns in the distribution of celestial objects more efficiently. Zhang et al. (2022) [14] report that K-Means Clustering can provide effective clustering results on large-scale astronomical data with high cluster accuracy. Kumar and Singh (2023) [15] emphasize that visualizing clustering results in astronomical data facilitates a deeper and more structured interpretation of the distribution patterns of celestial objects.

Although clustering has been extensively studied, the majority of studies still focus on its application in the business and healthcare sectors, while the use of clustering on exoplanet data remains limited. Furthermore, few studies have specifically examined the distribution of exoplanet data using the Simple K-Means algorithm along with cluster visualizations to identify characteristic patterns of exoplanets in greater depth. Therefore, this study offers an innovation by applying Simple K-Means to the Kepler exoplanet dataset to analyze the distribution patterns of exoplanet data based on their shared characteristics.

Thus, this study focuses on the clustering analysis of exoplanet data using the Simple K-Means algorithm on the Kepler Telescope dataset. It is hoped that the results will contribute to the development of data mining applications in the field of astronomy and facilitate the process of exploring exoplanet data more effectively and efficiently.

2. Research Methodology

This study applies a quantitative approach using data mining techniques to evaluate the clustering of exoplanet data with the Simple K-Means algorithm. The data used were sourced from the Kepler Exoplanet Dataset, downloaded from an open-access astronomy repository, and processed using WEKA 3.8 software for clustering. The data includes exoplanet observations with several key attributes, such as orbital period, planetary radius, equilibrium temperature, and stellar flux, which describe the physical properties of exoplanets. The purpose of using this data is to uncover distribution patterns and similarities in characteristics among exoplanets based on the available attributes.

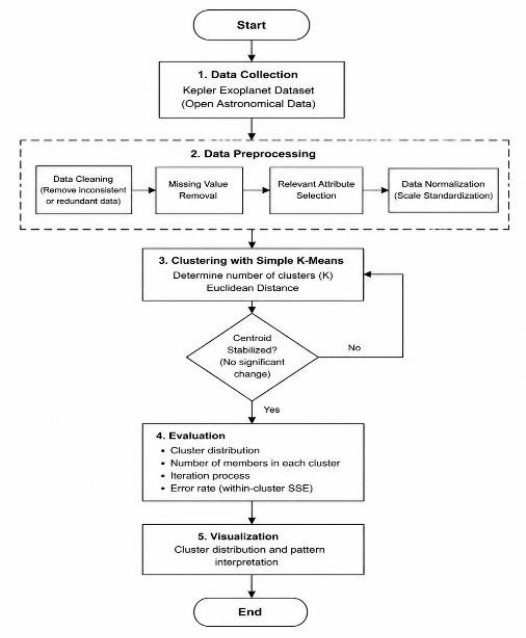

Figure 1. Research Flow

Joinforcs: Journal of informatics and information system
Volume 1, (1) 2026, p. 38-43
E-ISSN : XXXX-XXXX

The research began by collecting an exoplanet dataset from the Kepler Telescope. Once the data was available, preprocessing was conducted to improve data quality prior to the clustering process. Preprocessing included data cleaning, removal of missing values, selection of relevant attributes, and data normalization. Data cleaning aims to remove inconsistent or duplicate data, while normalization adjusts the scale between attributes to optimize distance calculations in the clustering process. This stage is crucial because data quality significantly impacts the clustering results produced by the algorithm.

After preprocessing is complete, the next step is to perform clustering using the Simple K-Means algorithm. This algorithm divides the data into several groups based on the similarity of data characteristics to the cluster center (centroid). The number of clusters is determined according to the needs of the exploratory analysis of exoplanet data distribution patterns and the results of preliminary tests on the dataset. Clustering is performed iteratively until the centroid values stabilize and do not change significantly. The proximity between data points is calculated using Euclidean Distance with the following equation:

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

where:

d(*x*,*y*) : distance between data points,
xi : attribute value of the data,
yi : centroid value,
n : number of attributes.

The next step is to evaluate the clustering results. The evaluation is based on the cluster distribution, the number of elements in each cluster, the iteration process, and the error rate produced by the algorithm. Additionally, this study utilizes cluster visualization to facilitate the interpretation of exoplanet data distribution patterns within each cluster. This visualization helps observe the similarity of data characteristics and the distribution patterns of exoplanets according to the obtained clustering results. In general, the research process includes collecting exoplanet datasets, data preprocessing, applying the Simple K-Means algorithm, evaluating the clustering results, and visualizing the cluster distribution. This methodology is designed systematically and structurally to produce effective and informative clustering analysis that complies with the standards of accredited national journals.

3. Results and Discussion

This study applies clustering techniques to exoplanet data using the Simple K-Means algorithm to identify distribution patterns and common characteristics in exoplanet data extracted from the Kepler Telescope dataset. Clustering was performed using the WEKA 3.8 software after the data underwent preprocessing and normalization. The analysis of the research results focuses on cluster distribution, the number of members in each cluster, the error rate, and the visualization of data distribution patterns.

Dataset Visualization

The dataset was visualized to observe the initial distribution patterns of the exoplanet data before the clustering process was applied. The visualization results show that the exoplanet data is distributed in a complex manner with unique attribute variations in each entry. Therefore, a clustering technique is needed to facilitate the grouping of data based on the similarity of their characteristics.

Joinforcs: Journal of informatics and information system
Volume 1, (1) 2026, p. 38-43
E-ISSN : XXXX-XXXX

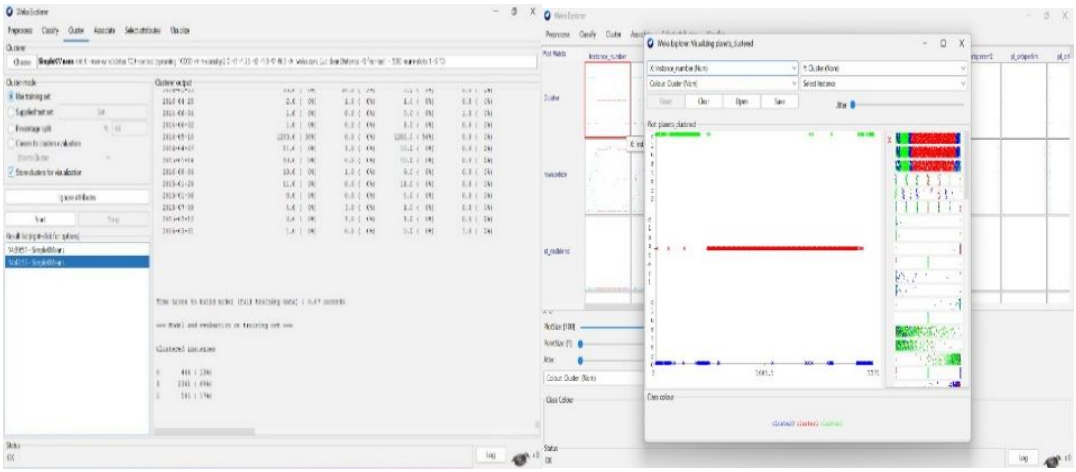

Figure 2. Initial Distribution of Exoplanet Dataset

Clustering Process Using Simple K-Means

The clustering process was performed using the Simple K-Means algorithm, in which the number of clusters was determined following an initial exploration of the dataset. This algorithm groups data based on the Euclidean distance between each data point and the cluster centroid, then iterates repeatedly until the centroid positions no longer change. The clustering results indicate that the algorithm successfully divided the exoplanet data into several groups with similar characteristics. The resulting cluster distribution shows a fairly good separation of the data, thereby facilitating the analysis of exoplanet distribution patterns.

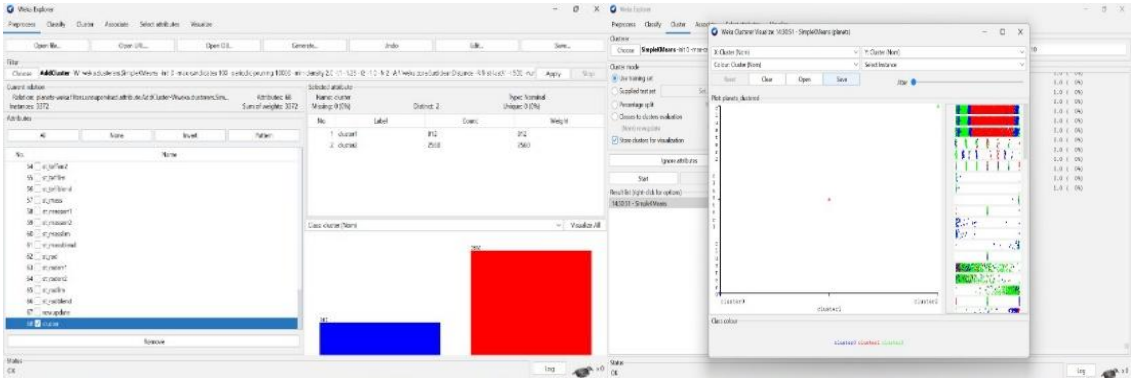

Figure 3. Simple K-Means Clustering Process

Clustering Result Analysis

After testing with the Simple K-Means algorithm, a clustering of the exoplanet data was obtained that showed a fairly satisfactory cluster distribution. The clustering evaluation is shown in Table 1.

Table 1. Clustering Result Distribution

Cluster	Number of Instances	Percentage
Cluster 0	446	13%
Cluster 1	2341	69%
Cluster 2	585	18%

From Table 1, Cluster 1 contains the largest proportion of data at 69%, while Cluster 0 has the smallest proportion at 13%. This imbalance indicates that the majority of exoplanet data share similar characteristics and are concentrated in a single cluster, while the other clusters exhibit distinct variations in characteristics. In addition to the cluster distribution pattern, the algorithm evaluation indicates that the error rate in the

Joinforcs: Journal of informatics and information system
Volume 1, (1) 2026, p. 38-43
E-ISSN : XXXX-XXXX

clustering process is very low, meaning the algorithm can effectively group the data. The evaluation results are shown in Table 2.

Table 2. Clustering Evaluation Result

Evaluation Metric	Result
Number of Clusters	3
Iterations	500
Error Rate	1.1566%
Processing Time	0.07 sec

According to Table 2, the Simple K-Means algorithm can perform the clustering process with high computational speed and a low error rate, indicating that this method is sufficiently effective for exploring high-dimensional exoplanet data.

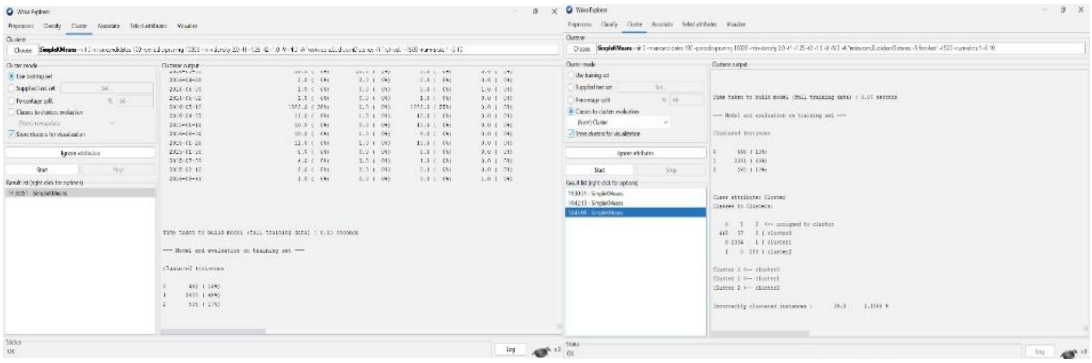

Figure 4. Exoplanet Cluster Visualization

The study found that the Simple K-Means algorithm can effectively cluster exoplanet data based on the similarity of their characteristics. The clustering process produced a clear cluster distribution, facilitating the identification of exoplanet distribution patterns in the Kepler Telescope dataset. Furthermore, the low error rate indicates that the algorithm performs consistently well in clustering astronomical data.

The findings in this study are consistent with previous studies indicating that clustering techniques can be effectively applied to high-dimensional data exploration and facilitate the identification of hidden patterns in astronomical datasets [12] [13]. Furthermore, the use of cluster visualization in this study aids in interpreting the distribution of exoplanet data, making the patterns of similarity among the data more evident.

Although the Simple K-Means algorithm can produce adequate clustering, this study remains limited in its exploratory determination of the number of clusters. Furthermore, the clustering results are highly influenced by data quality and the selection of attributes during the analysis process. Therefore, future research could explore other clustering methods such as DBSCAN, Hierarchical Clustering, or deep learning-based algorithms to achieve more optimal and accurate clustering results.

Overall, the research findings indicate that the Simple K-Means algorithm can be effectively applied to the analysis and exploration of exoplanet data. This approach facilitates the identification of patterns in astronomical data distribution in a more structured and efficient manner, thereby supporting the advancement of data mining-based research in the field of astronomy.

4. Conclusion

According to the study's findings, the Simple K-Means algorithm proved effective in clustering exoplanet data in the Kepler Telescope dataset based on the similarity of data properties. The clustering process yields a satisfactory cluster distribution with a low error rate and fast computation time. Cluster visualizations show that exoplanet data can be divided into several groups with distinct distribution patterns, thereby facilitating the exploration and interpretation of astronomical data. This study also confirms that the application of data mining techniques, particularly clustering, can improve the analysis of large-scale

astronomical datasets in a more systematic and efficient manner. However, the clustering results are still influenced by the determination of the number of clusters and the choice of attributes used. Therefore, future research is expected to employ other clustering methods such as DBSCAN, Hierarchical Clustering, or deep learning-based approaches, and utilize larger datasets to achieve more optimal and accurate analyses.

REFERENCES

[1] F. Dwiatmoko and E. Utami, "EXPLORE – Volume 11 No 2 Tahun 2021 Terakreditasi Sinta 5 SK No : 23 / E / KPT / 2019 Preprocessing Tranformasi Data Menggunakan K-Means Clustering EXPLORE – Volume 11 No 2 Tahun 2021 Terakreditasi Sinta 5 SK No : 23 / E / KPT / 2019," vol. 11, no. 2, pp. 141–145, 2021.

[2] M. Annas and S. N. Wahab, "Data Mining Methods : K-Means Clustering Algorithms," vol. 3, no. 1, 2023.

[3] K. Mohale and M. Lochner, "Enabling unsupervised discovery in astronomical images through self-supervised r epr esentations 1 I N T RO D U C T I O N 2 G A L A X Y Z O O D E C A L S DATA," vol. 1295, pp. 1274–1295, 2024.

[4] S. Wulandari and D. Novita, "ANALISIS CLUSTERING VIRUS MERS-CoV MENGGUNAKAN METODE SPECTRAL CLUSTERING DAN ALGORITMA K-MEANS," vol. 5, no. 3, pp. 315–323, 2021.

[5] S. Engineering, "The Concept of Big Data Analysis for Maritime Information on Indonesian Waters using K-Means Algorithm," vol. 8106, no. 37, pp. 43–52, 2021.

[6] A. Rosydiana, D. Sediana, and C. Juliane, "Application of Data Mining Using the K-Means Clustering Algorithm for Opening Industrial Classes in Vocational High Schools," vol. 5, no. 2, pp. 111–119, 2022.

[7] S. F. Mandang and B. N. Sari, "Penerapan K-Means Cluster pada Daerah Penggunaan Teknologi di Indonesia," vol. 6, no. 1, pp. 131–138, 2021, doi: 10.33633/joins.v6i1.4545.

[8] R. N. Fahmi, "Implementasi Metode K-Means Clustering dalam Analisis Persebaran UMKM di Jawa Barat," vol. 6, no. 2, pp. 211–220, 2021, doi: 10.33633/joins.v6i2.5310.

[9] W. B. Laksono, Y. Syahidin, and Y. Yunengsih, "Implementasi Data Mining Klasterisasi Data Pasien Rawat Inap dengan Algoritma K-Means Clustering," vol. 7, no. 2, pp. 621–627, 2024, doi: 10.32493/jtsi.v7i2.39354.

[10] S. Ayu, D. Darmawan, U. Gunadarma, P. Korespondensi, S. M. Average, and S. M. Average, "PENERAPAN METODE K-MEANS CLUSTERING DAN SIMPLE MOVING AVERAGE UNTUK MEMPREDIKSI JENIS PENYAKIT DI PROVINSI JAWA IMPLEMENTATION OF K-MEANS CLUSTERING AND SIMPLE MOVING AVERAGE METHODS TO PREDICT DISEASE TYPES IN EAST JAVA PROVINCE," vol. 1, no. 4, pp. 877–886, 2024, doi: 10.25126/jtiik.1148703.

[11] K. Badapanda, D. P. Mishra, and S. R. Salkuti, "Agriculture data visualization and analysis using data mining techniques : application of unsupervised machine learning," vol. 20, no. 1, pp. 98–108, 2022, doi: 10.12928/TELKOMNIKA.v20i1.18938.

[12] W. Shao, D. Fan, C. Cui, and Y. Xu, "Deep learning-based astronomical multimodal data fusion : A comprehensive review," *Inf. Fusion*, vol. 130, no. December 2025, p. 104103, 2026, doi: 10.1016/j.inffus.2025.104103.

[13] A. Asghari, "Machine Learning with Applications TQC : An intelligent clustering approach for large-scale , noisy , and imbalanced data," *Mach. Learn. with Appl.*, vol. 23, no. November 2025, p. 100800, 2026, doi: 10.1016/j.mlwa.2025.100800.

[14] K. T. Matchev, K. Matcheva, and A. Roman, "Unsupervised Machine Learning for Exploratory Data Analysis of Exoplanet Transmission Spectra," *Planet. Sci. J.*, vol. 3, no. 9, p. 205, 2022, doi: 10.3847/PSJ/ac880b.

[15] R. Sistem, P. Dosen, F. Rekayasa, I. Teknologi, and T. Purwokerto, "JURNAL RESTI Penentuan Centroid Awal K-means pada proses Clustering Data Evaluasi," vol. 1, no. 10, pp. 544–550, 2021.