ARJST
e-ISSN: 3048-9644
publications.ngmc.ac.in
https://doi.org/10.63300/ARJST.v5i07.1

Performance Analysis of Experimental Data for Heart Disease Prediction Using Unsupervised K-Means Clustering

Shabir Ahmad Najar[1,*]

[1] Assistant Professor, Higher Education Department, Jammu and Kashmir, India
***Corresponding author:** Shabir Ahmad Najar, shabirshaheen@gmail.com
ORCID iD: Shabir Ahmad Najar https://orcid.org/0009-0001-8148-2210

ARTICLE INFO

Article history
Received: 1 November 2026
Revised: 5 November 2026
Accepted: 15 November 2026

Article ID
ARJST-2026-0134

Peer review
Double-blind peer review

Keywords
Performance Analysis of Experimental Data for Heart Disease Prediction Using Unsupervised K-Means Clustering

ABSTRACT

Background: Heart disease is a leading cause of death worldwide, and medical databases hold large volumes of data from which diagnostic patterns could be extracted. This study develops a predictive system for heart disease using unsupervised K-means clustering and examines how the number of clusters (k) affects performance.

Methods: A heart disease dataset from a hospital in Srinagar, Jammu and Kashmir, was pre-processed using information-gain attribute evaluation with a ranker search. Redundant, missing and erroneous instances were removed, leaving 200 instances with 12 attributes. K-means models were built in Weka with k = 2, 4, 8 and 14. They were compared on squared error, time taken and number of iterations.

Results: Squared error fell as k increased, from 27.45% at k = 2 to 23.09% at k = 4, 14.23% at k = 8 and 3.04% at k = 14. Over the same range, build time rose from 0.0 to 0.05 seconds and the number of iterations from 3 to 9.

Conclusion: Larger numbers of clusters reduced error but increased computation time, iterations and model complexity. The eight-cluster model (k = 8) was selected as the most appropriate. A moderate number of clusters therefore appears best suited to heart disease prediction, and the optimal k depends on dataset size.

How to cite this article: Najar, S. A. (2026). Performance Analysis of Experimental Data for Heart Disease Prediction Using Unsupervised K-Means Clustering. *ARJST*, *5*(07), 1. https://doi.org/10.63300/ARJST.v5i07.1
Copyright and licence: © 2026 The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Data mining is an analytical process aimed at discovering useful patterns and significant knowledge from large and complex data warehouses in an efficient manner. The medical field generates substantial volumes of data relating to various critical diseases in diverse formats, including text, video, and images, which are continuously stored in backend databases in healthcare institutions. Thanks to the advancements made by the data mining community, it has become possible to extract valuable insights from these medical data repositories using an array of sophisticated mining techniques.

While the data encompasses various severe diseases, this study focuses specifically on heart disease datasets, aiming to develop a predictive model for timely diagnosis using the K-means clustering technique. Heart diseases are a leading cause of mortality globally, particularly in both developing and underdeveloped countries. In 2008, approximately 17 million individuals died worldwide due to cardiovascular diseases. The World Health Organization (WHO) projects that by 2030, around 24 million people will succumb to heart-related conditions [1].

K-means clustering is an unsupervised partition-based data mining technique that primarily aims to maximize intra-class similarity while minimizing inter-class similarity. The fundamental goal of the K-means algorithm is to categorize n observations into k clusters, where "k" denotes the number of clusters to be formed. For instance, k=2 indicates the formation of two clusters, and k=4 denotes four clusters. The choice of the number of clusters is crucial for developing an effective heart disease prediction model.

In this study, we utilize a processed heart disease dataset obtained from Jawaharlal Nehru Hospital, Srinagar, J&K,

comprising 250 instances with 12 attributes. These instances were classified into clusters to determine the optimal cluster size for enhancing predictive performance. We systematically varied the number of clusters (k=2, k=4, k=8, and k=14) to achieve acceptable predictive results using appropriate cluster sizes. High-quality clusters correlate with more efficient predictions, characterized by reduced ambiguity and increased certainty.

The primary objective of the K-means algorithm is to group data objects that exhibit similarity. During our experimental phase, various models were developed, and a comparative analysis of performance results was conducted to identify the most suitable model among the devised alternatives. The model that exhibited superior performance based on the selected cluster sizes was adopted for this research. It is important to note that both advantages and limitations are associated with increasing the number of clusters. For the experimental work, the Weka simulation tool was employed to develop models and identify the most effective heart disease prediction system. The flowchart in Figure 1 illustrates the proposed methodology for the development and diagnosis of heart disease prediction models.

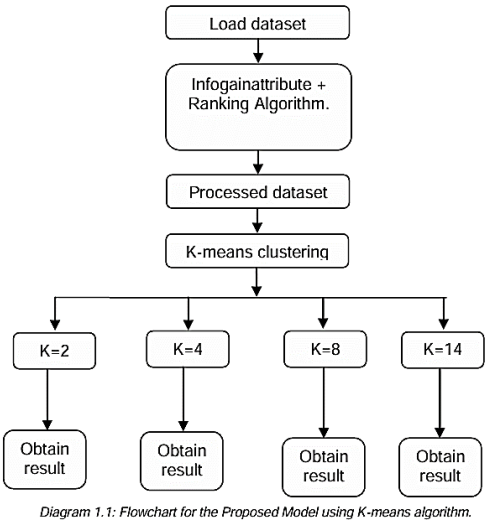


Figure 1. Flowchart of the proposed model using the K-means algorithm.

2. Review of Literature

The prediction of cardiovascular diseases (CVDs) has garnered significant attention in recent research, with various methodologies being employed for model development. For instance, a comprehensive model employing neural networks (NN), K-means clustering, and fuzzy inference systems (FIS) for the prediction of cardiovascular diseases was presented by [2]. This study utilized a set of 13 attributes to enhance the diagnosis of heart diseases, demonstrating the importance of attribute selection in predictive modeling.

In a comparative study conducted by Pandey et al., a novel approach utilizing clustering algorithms was proposed to analyze cardiovascular datasets. Their investigation into various clustering methods revealed that a density-based clustering algorithm achieved an accuracy of 85.8086%, identifying it as the most adaptable solution for heart disease prediction [3].

Singh et al. [4] evaluated the efficiency of two clustering techniques: K-means and hierarchical clustering. They found that K-means clustering exhibited superior performance, demonstrating greater efficiency and faster convergence when applied to large datasets. In contrast, hierarchical clustering operates by either merging smaller clusters into larger ones or subdividing large clusters into smaller units. The Weka simulation tool was employed to compare both techniques in terms of accuracy and runtime performance.

Further contributions to heart disease prediction include research conducted by Shekar et al., who developed an algorithm designed to extract association rules from hospital databases using a digital sequence mechanism. Their study emphasized the use of clustering to segment extensive databases, facilitating the grouping of data points into manageable clusters for more efficient heart disease predictions. This method notably reduces memory load by allowing the use of smaller clusters, thereby enhancing scalability [5].

Additionally, K-means clustering has been leveraged to uncover useful data patterns within medical databases for heart disease prediction. The Maximal Frequent Itemset Algorithm (MAFIA) was instrumental in calculating the weightage of these patterns, aiding in the diagnosis of CVDs [6].

Another study [7] implemented a hybrid approach utilizing clustering and association rule mining to predict and diagnose heart diseases, developed with C programming. This method effectively minimized main memory requirements by generating smaller clusters of data, which are more manageable.

Moreover, research presented in [8] explored the integration of K-means clustering with decision tree algorithms to enhance heart disease prediction models. The authors proposed multiple centroid selection methods to improve model efficiency and accuracy, achieving an accuracy rate of 83.9% through the application of their enabling method using a two-cluster model.

In summary, these studies highlight the diverse methodologies and techniques that can be employed for predicting cardiovascular diseases, demonstrating the evolving landscape of computational approaches in medical diagnostics. The ongoing exploration of clustering algorithms, attribute selection, and hybrid modeling techniques underscores their potential in effectively diagnosing heart conditions.

2026 The Author(s) https://creativecommons.org/licenses/by/4.0/ DOI: 10.63300/ARJST.v5i07.1

3. Methodology

3.1 Materials Used

The datasets employed for the experimental analysis of heart disease predictions were sourced from the Jawaharlal Nehru Hospital in Srinagar, Jammu and Kashmir. The dataset comprises a total of 1,600 instances, each characterized by 14 attributes. Table 1 provides a detailed overview of the attributes included in the dataset. It is noteworthy that the dataset encompasses instances from both male and female patients, thereby ensuring a comprehensive representation of gender in the analysis. This diverse dataset will serve as a critical foundation for applying various predictive modeling techniques in the subsequent phases of the research.

Table 1. Attributes of the heart disease dataset.

Attribute	Description
Patient Id	Dummy values are used.
Gender	Defines the gender of a patient: 1 = male, 0 = female.
Age	Categorized as child, young, or old (in years).
Exang.	Indicates exercise-induced angina.
Old peak	Represents depression induced angina.
Heart rate	Frequency of heart rate (maximum achieved).
Chol.	Levels of various types of cholesterol, including LDL and HDL.
fbs	Indicates whether fasting blood sugar exceeds the risk level, i.e., >120 mg/dl increases risk.
rbs	Resting blood sugar level.
trestbps	Systolic blood pressure to check its range.
Obesity	Indicates if a person is obese, calculated using BMI.
restecg	Normal, ST-T wave abnormality, or left ventricular hypertrophy.
smoking	Indicates whether the patient is a smoker (Yes or No).
Thal.	Measures blood flow in the heart.

3.2 Pre-processing

This study prioritizes the attainment of high-quality and clean datasets to develop an efficient predictive model for heart disease. The K-means clustering algorithm, which is employed in the analysis, is particularly sensitive to outliers. These outliers are observations that lie considerably far from the cluster's centroid, and their presence can significantly distort the mean value of the cluster, leading to unreliable results.

Pre-processing serves as a critical foundational step in the data mining process, establishing the groundwork for achieving successful outcomes. The careful and systematic selection of attributes plays a vital role in enhancing the likelihood of developing a robust diagnostic model, as it reduces the misclassification error rate across various scenarios.

In this research, we utilized the information gain method, specifically the infogainattribute approach combined with a ranker search, to assess the relevance of each attribute. This method measures the information gain concerning the class and ranks the attributes based on their individual evaluations. Through the pre-processing phase, we eliminated redundant, missing, and erroneous instances, thereby improving the overall quality of the dataset. As a result, we retained 200 instances characterized by 12 attributes for the subsequent experimental work.

3.3 K-means Clustering Algorithm

Clustering is a technique used to group objects that exhibit similarity, with the assumption that items within the same cluster are more alike than those in different clusters. The K-means clustering algorithm, first introduced by Macqueen in 1967, is a well-known and widely applied method in data mining. This technique falls under the category of partition clustering algorithms and operates as an unsupervised learning algorithm designed for unlabeled data.

The primary goal of clustering through the K-means algorithm is to partition the dataset into distinct groups such that objects within each group demonstrate a high degree of similarity. Each group formed by the algorithm is represented by a variable denoted as K, which refers to the number of clusters. The outcomes of the K-means clustering algorithm include:

- Identification of centroids for the K formed clusters, which are utilized to label new data points.
- Assignment of each data point from the dataset to its closest cluster based on distance.

The resulting clusters are characterized by the features of their respective centroids, allowing for qualitative interpretations of the formed groups. For example, let $X = \{x_1, x_2, x_3, \ldots, x_n\}$ represent the set of data points, and $V = \{v_1, v_2, v_3, \ldots, v_c\}$ represent the set of cluster centers. The equation utilized in the K-means algorithm to compute the new cluster centers is given by:

$$V_i = \frac{1}{C_i} \sum_{j=1}^{C_i} x_j$$

2026 The Author(s) https://creativecommons.org/licenses/by/4.0/ DOI: 10.63300/ARJST.v5i07.1

where C_i denotes the number of data points in the i-th cluster.

The steps of the K-means clustering algorithm, which are executed iteratively as the value of K changes, are outlined below:

Algorithmic steps for K-means clustering:

1. Select the dataset.
2. Apply preprocessing methods to obtain a cleaned and processed dataset.
3. Randomly select K (or c) initial cluster centers (input the intended number of clusters).
4. Use the Euclidean distance function to measure the distance between each data point and the cluster centers.
5. Assign each data point to the cluster whose center is closest.
6. Calculate new cluster centers by determining the mean of the points assigned to each cluster.
7. Repeat steps 2 through 6 iteratively until the assignment of data points to clusters stabilizes.
8. Repeat the process, starting from step 1, for different cluster sizes as needed.

In conclusion, the K-means algorithm aims to minimize the squared error function, defined as:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\| x_i - v_j \| \right)^2$$

Where,

'||xi - vj||' is the Euclidean distance function amid xi and vj
'ci' is the data points number in ith cluster.
'c' is the figure of cluster centers.

The EKA simulation tool has been developed to address the challenges associated with the K-means clustering algorithm, particularly its sensitivity to outliers, which can significantly distort expected outcomes. To mitigate this issue, it is essential to apply feature selection and ranking algorithms to the dataset prior to analysis. This preprocessing step enhances the clarity and efficiency of the dataset, resulting in a cleaner and more unambiguous data set.

Subsequently, we employ the refined dataset to develop various models utilizing the K-means clustering data mining algorithm. In this process, we systematically vary the number of clusters, analyzing configurations with k values of 2, 4, 8, and 14. The performance of each model is meticulously recorded to facilitate a comparative study of their efficacy.

4. Results

The experimental evaluation is conducted on the proposed model utilizing the JRLN hospital dataset to validate the effectiveness and reliability of the models for diagnosing heart diseases. A practical and comparative analysis is performed, examining the outcomes produced by the proposed models based on different cluster sizes. The experimental results obtained from the selection of various clusters using the K-means clustering technique are presented in Tables 2 and Figures 2 and 3.

Table 2. Comparative results across various cluster configurations (K-means clustering).

Performance measure	K=2	K=4	K=8	K=14
Squared error (%)	27.45	23.09	14.23	3.04
Time taken (s)	0.0	0.02	0.04	0.05
Number of iterations	3	6	7	9

This table summarizes the comparative performance results obtained from the K-means clustering technique across different cluster numbers (k values). The metrics evaluated include the squared error, time taken for computation, and the number of iterations required for convergence.

Figures 2 and 3 illustrates the performance measures achieved by the K-means clustering technique, depicting the results from various cluster configurations in a graphical format.



Figure 2. Performance measures for K-means clustering across different values of K.

2026 The Author(s) https://creativecommons.org/licenses/by/4.0/ DOI: 10.63300/ARJST.v5i07.1

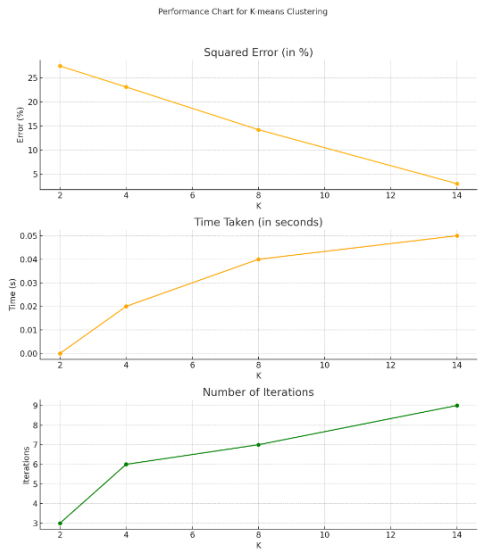


Figure 3. Squared error, time taken and number of iterations versus K.

5. Discussion

In this experimental study, we developed predictive models utilizing a clustering technique with a focus on varying cluster sizes through a detached approach. The formation of appropriate clusters is critical for the success of the predictive modeling process. Our work emphasizes several key aspects when employing K-means clustering to create these models:

- Utilization of processed datasets.
- Selection of an appropriate number of clusters.
- Assignment of data points to the cluster with the minimum distance from the cluster center, aimed at minimizing the squared error function.
- Time required to build the model.
- Number of iterations needed for convergence.

We explored various cluster sizes, specifically K=2, K=4, K=8, and K=14, to achieve satisfactory results based on the aforementioned criteria, employing comparative analysis between the models. During the modeling process, we systematically adjusted the cluster sizes to assess their impact on performance.

Our findings indicated that when the cluster size was set to K=2, we observed a higher mean squared error; however, this configuration resulted in minimal processing time and fewer iterations. As we increased the cluster sizes to K=4, K=8, and K=14, the mean squared error decreased, but we noted a corresponding increase in model complexity, time taken to build the model, and the number of iterations required for convergence.

6. Conclusion

Heart disease remains one of the leading causes of mortality worldwide, making early detection crucial for effective intervention. Modern predictive systems for heart disease play a vital role in assisting healthcare professionals in identifying these conditions with enhanced accuracy. In recent years, researchers have developed a variety of predictive models for heart disease using diverse data mining techniques, achieving notable success.

This research paper employed the k-means clustering technique to create several heart disease prediction models, varying the number of clusters to determine the most effective configuration. Four distinct models were developed based on cluster sizes of 2, 4, 8, and 14 (i.e., k=2, k=4, k=8, and k=14). Among these, the model utilizing eight clusters (k=8) was selected as the most appropriate based on performance metrics.

While the model with a higher number of clusters demonstrated improved accuracy, it also required more iterations and processing time, resulting in increased ambiguity. This study suggests that selecting a moderate number of clusters is more suitable for developing a robust heart disease prediction model. Furthermore, the optimal number of clusters is influenced by the size of the dataset, indicating that careful consideration of both cluster size and dataset characteristics is essential for effective predictive modeling.

Acknowledgements

I would like to express my sincere gratitude to Assistant Professor Dr. Irshad Ahmad Mir from the Department of Computer Applications at GDC, Charisherief and Dr. M.Saleem Mir. Their invaluable support and guidance were instrumental in the successful completion of this research paper.

Funding

This research received no external funding.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

the dataset was obtained from Government Jawaharlal Nehru Hospital, Srinagar J&K, India

References

References

Mujtaba Ashraf, Azad Kumar, “Development and Evaluation of a collection of Heart Disease Prediction Models using Decision Tree Algorithm based on the various Feature Selection Methods” RRIJM, vol. 4, March 2019.

Mujtaba Ashraf et. al., “Comparative Study of Existing Techniques for Heart Diseases Prediction Using Data Mining

2026 The Author(s) https://creativecommons.org/licenses/by/4.0/ DOI: 10.63300/ARJST.v5i07.1

Approach." Asian Journal of Computer Science and Information Technology, 3-July - 2017.

Dangare, C. S., & Apte, S. S. (2012). A data mining approach for prediction of heart disease using neural networks. International Journal of Computer Engineering and Technology.

D'Souza, A. (2015). Heart Disease Prediction Using Data Mining Techniques. International Journal of Research in Engineering and Science (IJRES), 3(3), 74-77. ISSN (Online): 2320-9364.

Shafranovich. (2011). Common Format and MIME Type for Comma-Separated Values (CSV) Files. Retrieved September 12, 2011, from http://tools.ietf.org/html/rfc4180

Palaniappan, S., & Awang, R. (2008). Intelligent Heart Disease Prediction System using Data Mining Techniques. International Journal of Computer Science and Network Security, 8(8), 1-6.

Shekar, K., Deepika, N., & Sujatha, D. (2011). Association rule for classification of heart-attack patients. International Journal of Advanced Engineering Sciences and Technologies, 11(2), 253-257.

Patil, S. B., & Kumaraswamy, Y. S. (2009). Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction. International Journal of Computer Science and Network Security (IJCSNS), 9(2), 228-235.

Jabbar, M., Chandra, P., & Deekshatulu, B. (2011). Cluster Based Association Rule Mining for Heart Disease Prediction. Journal of Theoretical & Applied Information Technology, 32(2), 196–201.

Shouman, M., Turner, T., & Stocker, R. (2012). Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients. In Proceedings of the International Conference on Data Mining.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

Masethe, H. D., & Masethe, M. A. (2014). Prediction of heart disease using classification algorithms. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 2, pp. 22-24).

2026 The Author(s) https://creativecommons.org/licenses/by/4.0/ DOI: 10.63300/ARJST.v5i07.1