

## Enhanced Comparative Analysis of K-Means Clustering Using Euclidean and Manhattan Distance Metrics for Chronic Kidney Disease Prediction

**Dr. R.S. Walse<sup>1</sup>**

<sup>1</sup>Assistant Professor (Computer Science),  
Department of Dairy Business Management,  
College of Dairy Technology, Warud (Pusad),  
M.A.F.S.U, Nagpur, Maharashtra State, India  
rajeshwalse@gmail.com

### Abstract:

*Chronic Kidney Disease (CKD) is a progressive disorder that has become a significant healthcare challenge due to its increasing prevalence and the severe complications associated with delayed diagnosis. Recent advances in data mining and machine learning have enabled the development of intelligent frameworks capable of extracting meaningful information from medical datasets to support clinical decision-making. In this study, a comparative clustering-based approach is presented for the analysis and prediction of chronic kidney disease using the K-Means algorithm integrated with Euclidean and Manhattan distance measures. The experimental investigation was performed on the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository, consisting of 400 patient records and 25 clinical attributes. Data preprocessing and clustering operations were conducted using the WEKA environment. Multiple cluster configurations ranging from two to five clusters were examined under different evaluation settings, including training-set-based clustering and class-to-cluster evaluation. Performance assessment was carried out using within-cluster sum of squared errors to determine cluster compactness and effectiveness. Experimental results demonstrate that increasing the number of clusters improves clustering quality by reducing the error values and producing better partitioning of CKD and non-CKD instances. Comparative analysis reveals that both Euclidean and Manhattan distance metrics provide efficient clustering performance and are capable of generating reliable predictive patterns from clinical data. The proposed framework offers a simple and computationally efficient unsupervised learning methodology that may assist healthcare professionals in the early identification and management of chronic kidney disease.*

**Keywords**— Chronic Kidney Disease, K-Means Clustering, Euclidean Distance, Manhattan Distance, Data Mining, Machine Learning, WEKA, Unsupervised Learning.

### I. Introduction:

Chronic Kidney Disease (CKD) is recognized as one of the most prevalent and life-threatening disorders affecting the global population. The disease is characterized by the gradual deterioration of kidney function, which may ultimately lead to kidney failure and associated cardiovascular complications. Owing to changes in lifestyle, increasing incidence of diabetes and hypertension, and the aging population, the number of CKD

patients has increased considerably over the last few decades. Early diagnosis and timely intervention are essential to reduce disease progression and improve patient outcomes.

The rapid growth of healthcare data has created opportunities for the application of data mining and machine learning techniques in disease diagnosis and prediction. These computational approaches facilitate the extraction of hidden patterns and valuable knowledge from large volumes of clinical data, thereby assisting healthcare professionals in making informed decisions. In

particular, unsupervised learning techniques provide an effective mechanism for discovering similarities and relationships among data instances without relying on predefined class labels.

Table 1 Attributes and its units

| Sr. no. | Attributes |                        | Type      | Type of attributes and its unit |
|---------|------------|------------------------|-----------|---------------------------------|
| 1       | Age        | Age                    | Numerical | Age in years                    |
| 2       | bp         | Blood pressure         | Numerical | Bp in mm/Hg                     |
| 3       | sg         | Specific gravity       | Nominal   | 1.005,1010,1.015,1.020,1.025    |
| 4       | al         | Albumin                | Nominal   | 0,1,2,3,4,5                     |
| 5       | su         | Sugar                  | Nominal   | 0,1,2,3,4,5                     |
| 6       | bgr        | Blood glucose random   | Numerical | mg/dl                           |
| 7       | bu         | Blood urea             | Numerical | mg/dl                           |
| 8       | sc         | Serum creatinine       | Numerical | mg/dl                           |
| 9       | sod        | Sodium                 | Numerical | mEq/L                           |
| 10      | pot        | Potassium              | Numerical | mEq/L                           |
| 11      | Hemo       | Hemoglobin             | Numerical | gms                             |
| 12      | pcv        | Packed cell volume     | Numerical | Numerical                       |
| 13      | wc         | White blood cell count | Numerical | cells/cumm                      |
| 14      | rc         | Red blood cell count   | Numerical | millions/cmm                    |
| 15      | Class      | Class                  | Nominal   | ckd/notckd                      |

The processed data using Weka can be analyzed using simple k-means clustering algorithm with Euclidean in addition Manhattan distance function and to visualize the predicted value in the form of graphs and visualized into a two-dimensional graphical format, respectively, and we will get the result what will happen with our data in the form of visualization techniques.

Among various clustering techniques, the K-Means algorithm has gained significant attention because of its simplicity, computational efficiency, and capability to partition data into homogeneous groups. The clustering process is strongly influenced by the distance measure employed to determine the similarity between data points. Euclidean distance and Manhattan distance are two widely adopted metrics that affect cluster formation and overall clustering performance. A comparative investigation of these distance measures is therefore essential to identify their suitability for medical data analysis.

## Proposed Module:

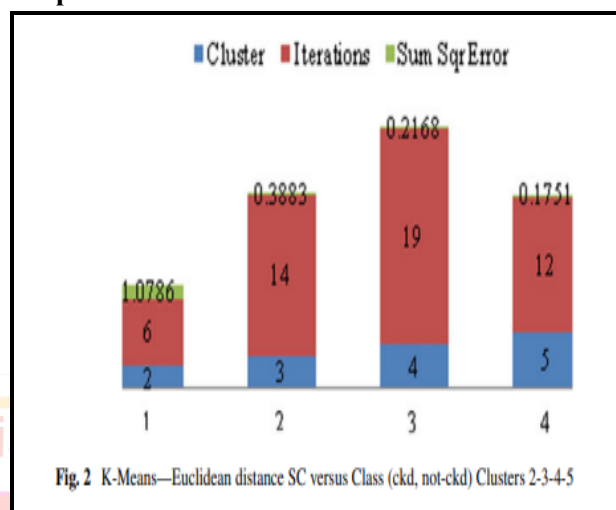


Fig. 2 K-Means—Euclidean distance SC versus Class (ckd, not-ckd) Clusters 2-3-4-5

Several researchers have utilized classification algorithms and predictive models for chronic kidney disease diagnosis. However, limited attention has been devoted to studying the influence of different distance metrics within clustering frameworks for CKD analysis. Furthermore, the effectiveness of K-Means clustering under varying numbers of clusters requires additional investigation to establish an optimal clustering strategy capable of generating meaningful partitions and improving predictive accuracy. In this research, a comparative framework based on the K-Means clustering algorithm employing Euclidean and Manhattan distance functions is proposed for chronic kidney disease analysis. The study utilizes the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository, comprising 400 instances and 25 attributes. Experimental analysis is conducted under different clustering configurations to evaluate the influence of cluster size and distance measures on clustering performance. The quality of clustering is assessed using within-cluster sum of squared errors, which serves as an indicator of cluster compactness and separation.

The primary contributions of this study are summarized as follows:

1. Development of a clustering-based framework for chronic kidney disease analysis using the K-Means algorithm.
2. Comparative evaluation of Euclidean and Manhattan distance metrics under different cluster configurations.
3. Investigation of the effect of increasing the number of clusters on clustering quality.

4. Assessment of clustering performance using within-cluster sum of squared errors.
5. Demonstration of the applicability of unsupervised learning techniques for supporting early CKD prediction and clinical decision-making.

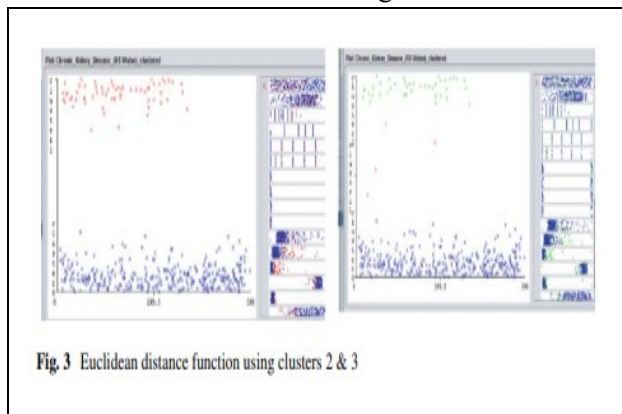


Fig. 3 Euclidean distance function using clusters 2 & 3

The remainder of the paper is organized as follows. Section II reviews related studies on chronic kidney disease prediction and clustering techniques. Section III describes the dataset, preprocessing procedures, and experimental methodology. Section IV presents the experimental results and comparative analysis. Section V discusses the findings and implications of the study. Finally, Section VI concludes the paper and highlights possible directions for future research.

## II. Research Methodology:

The application of data mining and machine learning techniques in healthcare has received considerable attention because of their ability to extract meaningful information from complex medical datasets. Numerous studies have explored computational approaches for disease diagnosis, prediction, and clinical decision support systems.

Chen et al. [1] introduced an improved clustering approach for determining cluster centers efficiently. The proposed methodology enhanced clustering performance and reduced computational complexity, thereby providing better cluster formation compared with conventional methods.

Arasu and Selvi [2] presented a novel prediction methodology for kidney disease analysis using data mining techniques. Their study emphasized the role of intelligent models in identifying disease patterns and improving prediction accuracy. Similarly, Kunwar et al. [4] investigated various classification algorithms for chronic kidney disease diagnosis and demonstrated

the effectiveness of machine learning techniques in medical data analysis.

Khanna [6] analyzed the economic burden associated with dialysis treatment in India and emphasized the necessity of early disease detection to reduce healthcare expenses. Varma [8] investigated the prevalence of chronic kidney disease in India and reported a continuous increase in the number of affected individuals. The study highlighted the urgent need for advanced predictive frameworks capable of supporting early diagnosis and preventive healthcare measures.

Although previous studies have extensively explored classification algorithms and disease prediction models, comparatively fewer investigations have focused on clustering-based analysis of chronic kidney disease. In addition, limited research has examined the impact of different distance measures on clustering performance. Most existing approaches primarily emphasize classification accuracy while overlooking the influence of cluster formation and similarity metrics. Therefore, a comprehensive comparative analysis of K-Means clustering using Euclidean and Manhattan distance functions is necessary to understand their effectiveness in identifying CKD and non-CKD patterns. The present study addresses this research gap by investigating the performance of K-Means clustering under different cluster configurations and distance metrics. The proposed framework aims to provide a computationally efficient and reliable approach for chronic kidney disease analysis using unsupervised learning techniques.

## III. Materials and Methods:

### A. Dataset Description:

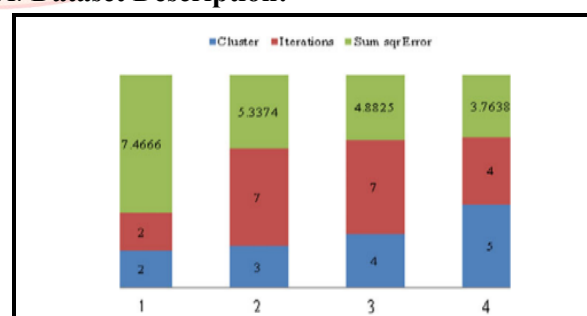


Fig. 4 K-Means—Manhattan distance SC versus Class (ckd, not-ckd) Clusters 2-3-4-5 Classes to cluster evaluation + ignoring attributes

The experimental analysis presented in this study was performed using the Chronic Kidney Disease (CKD) dataset obtained from the UCI Machine Learning Repository. The dataset contains 400 patient records characterized by 25 clinical attributes associated with kidney function and related physiological conditions. These attributes consist of both numerical and categorical variables, providing comprehensive information required for disease analysis.

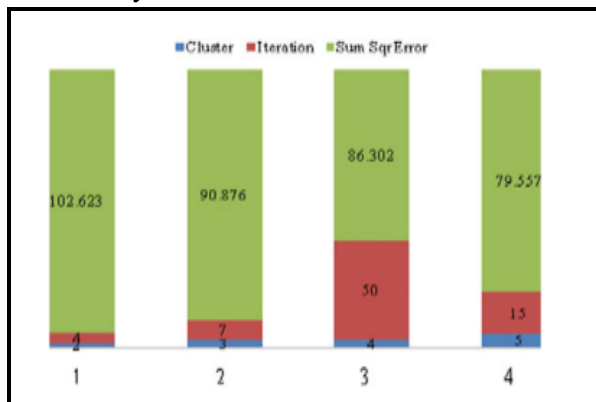


**Fig. 5** K-means—Euclidean distance 15 Attributes (Full) versus Class (ckd, not-ckd) Clusters 2-3-4-5 Classes to cluster evaluation + ignoring attributes

The numerical attributes include age, blood pressure, blood glucose random level, blood urea, serum creatinine, sodium concentration, potassium concentration, hemoglobin level, packed cell volume, white blood cell count, and red blood cell count. The categorical attributes include specific gravity, albumin, sugar, red blood cell status, pus cell condition, bacteria, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia, and the corresponding class label representing CKD and non-CKD instances.

Among the total observations, approximately 63% belong to chronic kidney disease cases, whereas the remaining 37% correspond to non-CKD patients. The diversity of clinical variables makes the dataset suitable for evaluating clustering-based prediction frameworks.

### B. Data Preprocessing:

Data preprocessing represents a crucial stage in knowledge discovery because the quality of input data significantly influences the performance of machine learning algorithms. Prior to clustering, the dataset was subjected to cleaning procedures to eliminate inconsistencies and handle missing values.

The preprocessing operations were carried out using the WEKA data mining environment.

Numerical and categorical attributes were properly organized to ensure compatibility with the clustering algorithm. Missing values present in the dataset were addressed during preprocessing to improve data quality and reduce the possibility of biased cluster formation. The resulting processed dataset was subsequently utilized for experimental analysis.

### C. K-Means Clustering Algorithm:

K-Means is an unsupervised partitioning algorithm that divides a dataset into K distinct clusters based on the similarity among data points. The algorithm aims to minimize intra-cluster variation while maximizing inter-cluster separation. Initially, K cluster centroids are selected, and each data instance is assigned to the nearest centroid according to a predefined distance metric. The cluster centers are then updated iteratively until convergence is achieved.

The objective function of K-Means can be expressed as

$$[J = \sum_{i=1}^K \sum_{x_j \in C_i} \|x_j - \mu_i\|^2]$$

where:

- (K) denotes the number of clusters,
- ( $C_i$ ) represents the set of observations belonging to the ( $i^{th}$ ) cluster,
- ( $x_j$ ) indicates an individual data point,
- ( $\mu_i$ ) corresponds to the centroid of cluster ( $i$ ).

The optimization process continues until the within-cluster sum of squared errors reaches a minimum value.

### D. Euclidean Distance Measure:

Euclidean distance is one of the most commonly employed similarity measures in clustering analysis. It determines the straight-line distance between two points in multidimensional space and is mathematically represented as

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

where ( $x_i$ ) and ( $y_i$ ) represent the attribute values of two observations and ( $n$ ) denotes the total number of features.

The Euclidean metric emphasizes geometric proximity and is widely used because of its

computational simplicity and effectiveness in measuring similarity among numerical attributes.

#### E. Manhattan Distance Measure:

Manhattan distance, also known as city-block distance, calculates the absolute difference between corresponding coordinates of two data points. The mathematical expression is given by

$$[d(x,y)=\sum_{i=1}^n|x_i-y_i|]$$

Unlike Euclidean distance, Manhattan distance measures the path traversed along orthogonal directions and is less sensitive to outliers. This characteristic makes it suitable for certain medical datasets where extreme values may influence cluster formation.

#### F. Experimental Configuration

Experimental investigations were conducted using the WEKA software platform. The K-Means clustering algorithm was executed with a maximum iteration limit of 500. Two distance measures, namely Euclidean distance and Manhattan distance, were employed to evaluate clustering performance.

Three experimental scenarios were considered:

1. Training-set-based clustering using serum creatinine and class attributes.
2. Class-to-cluster evaluation using serum creatinine and class information.
3. Clustering using all fifteen numerical attributes available in the dataset.

For each scenario, the number of clusters was varied from two to five to examine the effect of cluster size on clustering quality. A 70% data split was adopted during experimentation. The effectiveness of cluster formation was evaluated using the within-cluster sum of squared errors (SSE). Lower SSE values indicate improved cluster compactness and better clustering performance.

The comparative analysis between Euclidean and Manhattan distance metrics enables the identification of the most suitable similarity measure for chronic kidney disease data analysis and prediction.

### IV. Results And Performance Analysis :

#### A. Experimental Evaluation :

The proposed clustering framework was implemented in the WEKA environment to investigate the effectiveness of K-Means clustering using Euclidean and Manhattan distance metrics. The analysis was performed on the Chronic Kidney

Disease dataset comprising 400 instances and 25 attributes. A maximum iteration limit of 500 was considered, and the experiments were conducted using a 70% data partition.

To examine the influence of cluster size on clustering performance, the number of clusters was gradually increased from two to five. Three different experimental scenarios were investigated to evaluate the behavior of the clustering algorithm under varying attribute configurations and evaluation strategies.

#### B. Analysis Using Serum Creatinine Attribute :

Initially, clustering experiments were performed by considering the relationship between serum creatinine (SC) and the class label representing CKD and non-CKD instances. The clustering process was executed using both Euclidean and Manhattan distance measures under the training-set evaluation mode.

The experimental observations indicated that increasing the number of clusters from two to five resulted in a gradual reduction in the within-cluster sum of squared errors (SSE). A lower SSE value represents improved cluster compactness and indicates better grouping of similar instances. Consequently, the clustering quality improved with an increase in the number of clusters.

Both Euclidean and Manhattan distance functions exhibited similar behavior and generated meaningful cluster structures capable of separating CKD and non-CKD patient records effectively.

#### C. Class-to-Cluster Evaluation

A second set of experiments was conducted using class-to-cluster evaluation while preserving the relationship between serum creatinine and disease class. This approach enabled the investigation of how clusters correspond to the underlying disease categories.

The results demonstrated that increasing the cluster count enhanced cluster separation and reduced the SSE values. The iterative optimization process produced more compact clusters as the number of partitions increased. Similar trends were observed for both distance metrics, indicating that Euclidean and Manhattan measures are equally effective for identifying patterns within CKD data.

The obtained clusters showed a reasonable distribution of CKD and non-CKD instances,

thereby confirming the capability of K-Means clustering to discover hidden structures in medical datasets.

#### D. Analysis Using Complete Numerical Attributes

To further investigate the robustness of the proposed framework, clustering experiments were extended by incorporating all fifteen numerical attributes available in the dataset. These attributes included age, blood pressure, blood glucose random level, blood urea, serum creatinine, sodium concentration, potassium concentration, hemoglobin level, packed cell volume, white blood cell count, and red blood cell count, among others.

Experimental observations revealed a pattern similar to the previous analyses. Increasing the number of clusters consistently reduced the within-cluster sum of squared errors and improved cluster compactness. The results indicate that incorporating multiple clinical parameters enhances the capability of the clustering algorithm to capture relationships among patient records. Both Euclidean and Manhattan distance measures provided stable performance and produced effective partitions of CKD and non-CKD instances. The generated clusters exhibited satisfactory separation characteristics, demonstrating the suitability of K-Means clustering for chronic kidney disease analysis.

#### E. Comparative Analysis of Distance Metrics

The comparative study revealed that both distance measures are capable of generating reliable clustering structures. Euclidean distance emphasizes geometric proximity among observations, whereas Manhattan distance evaluates the cumulative absolute difference between attribute values. Despite their computational differences, both metrics produced comparable clustering behavior across all experimental settings. A reduction in the within-cluster sum of squared errors was observed as the number of clusters increased, indicating improved clustering performance. The findings suggest that cluster quality is influenced not only by the distance metric but also by the selection of an appropriate number of clusters.

#### F. Discussion of Findings

The experimental results confirm that K-Means clustering can effectively identify patterns associated with chronic kidney disease. The

progressive reduction in SSE values demonstrates that increasing the number of clusters improves cluster compactness and enhances the discrimination between CKD and non-CKD instances. Furthermore, the study indicates that both Euclidean and Manhattan distance measures provide satisfactory performance for medical data analysis. The visualization and clustering outcomes obtained from the WEKA platform suggest that unsupervised learning techniques can serve as valuable tools for assisting healthcare professionals in understanding disease characteristics and supporting early diagnosis. Overall, the proposed framework demonstrates the feasibility of applying clustering-based approaches to chronic kidney disease prediction and highlights the significance of distance metrics in determining clustering quality.

#### V. Conclusion and Future Scope:

This study presented a comparative analysis of the K-Means clustering algorithm employing Euclidean and Manhattan distance measures for chronic kidney disease prediction. The experiments were conducted using the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository and implemented within the WEKA environment. Different clustering configurations ranging from two to five clusters were investigated under multiple evaluation settings to analyze their impact on clustering quality.

The experimental findings demonstrated that the within-cluster sum of squared errors decreased as the number of clusters increased, indicating improved cluster compactness and enhanced clustering performance. Both Euclidean and Manhattan distance metrics exhibited satisfactory behavior and produced reliable clustering structures capable of distinguishing between CKD and non-CKD patient instances. The comparative analysis revealed that both distance measures are effective for clustering medical datasets and can contribute to the development of intelligent decision-support systems for healthcare applications. The results obtained from this investigation emphasize the potential of unsupervised learning techniques in extracting meaningful information from clinical data. The proposed framework offers a computationally efficient approach for identifying disease-related

patterns and may assist healthcare professionals in early diagnosis, treatment planning, and disease management. Furthermore, the use of clustering techniques can reduce the dependency on extensive laboratory examinations and support cost-effective healthcare practices. Although the present study focused on K-Means clustering with Euclidean and Manhattan distance functions, several opportunities exist for extending this research. Future investigations may consider the application of advanced clustering techniques such as Fuzzy C-Means, Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixture Models. In addition, hybrid machine learning approaches integrating clustering and classification algorithms may be explored to improve predictive accuracy.

Further research may also incorporate larger and more diverse clinical datasets to enhance the generalization capability of the proposed framework. Feature selection methods, dimensionality reduction techniques, and deep learning-based architectures can be investigated to improve computational efficiency and prediction performance. Moreover, the development of real-time clinical decision support systems based on intelligent data mining methodologies represents a promising direction for future healthcare applications.

Overall, the findings of this study demonstrate that K-Means clustering combined with Euclidean and Manhattan distance metrics provides an effective and reliable framework for chronic kidney disease analysis and can serve as a foundation for developing advanced predictive systems in the medical domain.

## References

1. J. Chen et al., "A novel cluster center fast determination clustering algorithm," *Applied Soft Computing*, vol. 57, pp. 539–555, Aug. 2017.
2. S. Arasu and R. D. Thirumalaiselvi, "A novel imputation method for effective prediction of kidney disease," in *Proc. 2nd Int. Conf. Computing and Communication Technologies (ICCCCT)*, Chennai, India, Feb. 2017, pp. 127–136.
3. V. Kunwar et al., "Chronic kidney disease analysis using data mining classification techniques," in *Proc. 6th Int. Conf. Cloud System and Big Data Engineering (Confluence)*, Noida, India, Jan. 2016.
4. T. Due and A. Luong, "A K-means approach to clustering disease progressions," Dept. Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA, Tech. Rep., Sept. 2017.
5. N. Kumar and S. Khatri, "Implementing WEKA for medical data classification and early disease prediction," in *Proc. 3rd IEEE Int. Conf. Computational Intelligence and Communication Technology (CICT)*, Ghaziabad, India, Feb. 2017.
6. U. Khanna, "The economics of dialysis in India," *Indian Journal of Nephrology*, vol. 19, no. 1, pp. 1–4, Jan. 2009.
7. P. P. Varma, "Prevalence of chronic kidney disease in India: Where are we heading?," *Indian Journal of Nephrology*, vol. 25, no. 3, pp. 133–135, May 2015.