# CLASSIFYING IRIS DATASET USING K-MEANS CLUSTERING

**Project-1 Submitted in partial fulfilment for the award of**

**B.Sc. Degree in Computer Science**

**Madurai Kamaraj University, Madurai.**


| | | |
|---|---|---|
| **NAME** | **:** | **M.SRINIVAS** |
| **ROLL.NO** | **:** | **22AUCS036** |
| **CLASS** | **:** | **III-B.Sc. COMPUTER SCIENCE** |
| **SUBMITTED ON** | **:** | **17/02/2025** |
| **PROJECT GUIDE** | **:** | **Dr.T.KATHIRVALAVAKUMAR** |




**Research Center in Computer Science**

**V.H.N.SENTHIKUMARA NADAR COLLEGE(Autonomous)**
**Virudhunagar.**


**FEBRUARY 2025**

# CLASSIFYING IRIS DATASET USING K-MEANS CLUSTERING

**Project-1 Submitted in partial fulfilment for the award of**

**B.Sc. Degree in Computer Science**

**Madurai Kamaraj University, Madurai.**


| | | |
|---|---|---|
| **NAME** | **:** | **M.SRINIVAS** |
| **ROLL.NO** | **:** | **22AUCS036** |
| **CLASS** | **:** | **III-B.Sc. COMPUTER SCIENCE** |
| **SUBMITTED ON** | **:** | **17/02/2025** |
| **PROJECT GUIDE** | **:** | **Dr.T.KATHIRVALAVAKUMAR** |


**Project Guide Signature** **HOD Signature**


**Research Center in Computer Science**

**V.H.N.SENTHIKUMARA NADAR COLLEGE(Autonomous)**
**Virudhunagar.**


**FEBRUARY 2025**

# CLASSIFYING IRIS DATASET USING K-MEANS CLUSTERING

## OBJECTIVES

The primary objectives of this project are to leverage the **K-means clustering** algorithm in Python to group data points based on their features, and to critically assess the algorithm's performance using accuracy metrics. This work aims to provide valuable insights into data organization and enhance the understanding of clustering techniques in data science.

## PROBLEM DESCRIPTION

Clustering is an unsupervised machine learning technique used to group similar data points into clusters. The K-means clustering algorithm is one of the most popular clustering algorithms, which partitions the data into **K clusters**, where each data point belongs to the cluster with the nearest mean. This project focuses on implementing the **K-means algorithm**, normalizing the data, and evaluating the clustering results on a dataset.

## PROCEDURE

1. **Data Preprocessing**

   - **Encoding** is to convert categorical string data points to numerical codes using the astype('category').cat.codes built in method.

   - **Normalization** is applied for scaling the features to a range of [0, 1] using the formula

$$x_{\text{normalized}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

*Where*

$x$ *is the original value.*

$x_{\text{min}}$ *is the minimum value in the dataset.*

$x_{\text{max}}$ *is the maximum value in the dataset.*

$x_{\text{normalized}}$ *is the normalized value.*

   - Use train_test_split() built in method from sklearn package for divide the data into 80% for training and 20% testing.

2. **Initialize Centroids**
    - Randomly select K unique points from the dataset as initial centroids.

3. **Euclidean Distance Calculation**
    - Compute the Euclidean distance between two points centroids and data points.
    - For each data point $x_i$ and each centroid $c_k$

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - c_{jk})^2}$$

4. **Cluster Assignment**
    - Assign each data point to the nearest centroid based on Euclidean distance.

5. **Update Centroids**
    - Compute the new centroids as the mean of all points assigned to each cluster.

6. **Label Assignment**
    - Assign predicted labels to each cluster based on the majority class of the original labels of the centroids.

7. **Iterate**
    - Repeat the clustering and updating steps until the centroids converge (i.e., no change in centroids).

8. **Evaluate Accuracy**
    - Calculate the accuracy of the predicted labels using accuracy_score() method from sklearn package.

## EXPERIMENTAL RESULTS

The K-means clustering algorithm was applied to the Iris dataset. Iris dataset has 4 features and 1 label. The label contains three species or class that are, Iris-setosa, Iris-versicolor, and Iris-virginica. Each species has 50 Instances.

Experiment is repeated ten times for each k value. The average values are given in the following table. The Accuracy is differed for each value of **K**, that is tabulated below:

**Table: Performance results of K-means**

| Value for K | Centroids | Members | Label Prediction | Accuracy |
|---|---|---|---|---|
| 4 | CENTROID 1 | 8 | Iris-versicolor | 77.62% |
| | CENTROID 2 | 8 | Iris-setosa | |
| | CENTROID 3 | 5 | Iris-setosa | |
| | CENTROID 4 | 9 | Iris-virginica | |
| 5 | CENTROID 1 | 7 | Iris-versicolor | 76.366% |
| | CENTROID 2 | 5 | Iris-virginica | |
| | CENTROID 3 | 4 | Iris-versicolor | |
| | CENTROID 4 | 5 | Iris-setosa | |
| | CENTROID 5 | 9 | Iris-versicolor | |

## CONCLUSION

The K-means clustering algorithm successfully partitioned the data points into distinct clusters based on their features. The results show that the algorithm can effectively group similar data points together, achieving an average accuracy is 76.993%. This project demonstrates the implementation of K-means clustering from scratch and highlights the importance of data preprocessing and normalization in improving the performance of clustering algorithms.