APPLIED DATA SCIENCE ASSIGNMENT 2

Iris Species Classification through K-Means Clustering Analysis

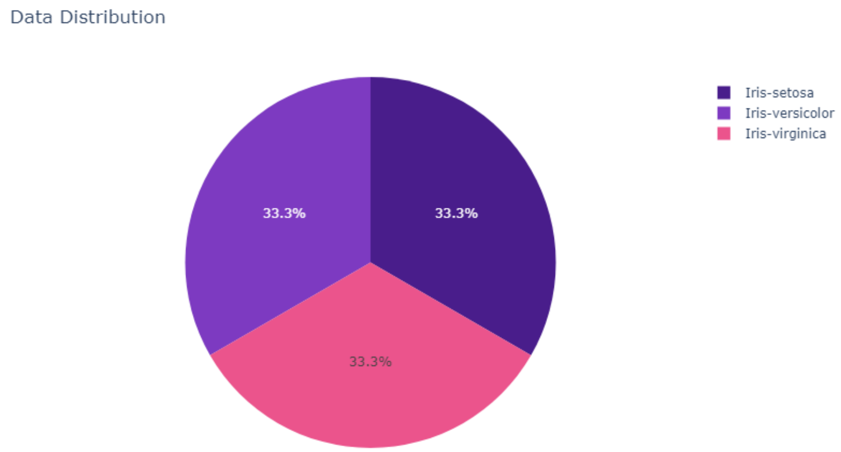
Name: Nitesh Reddy
Student ID Number: 23038663
University: University of Hertfordshire

Dataset: https://www.kaggle.com/datasets/uciml/iris
Github_link:

Abstract:

A well-known dataset for teaching machine learning, data visualization, and basic classification is the Iris dataset. It is very straightforward: four numerical properties about each of the three classes (different Iris species)—petal length, petal width, sepal length, and sepal width—are given for each of the 50 samples in each class. We will be doing the fitting as well as clustering and get to see in which the species will classify.

Graphs:



Pie Chart:

By analyzing the pie chart, we can conclude that the data is perfectly balanced. The dataset is comparatively balanced as the pie chart illustrates, with roughly one-third of the total observations falling into each class.

Fig 1: Pie chart

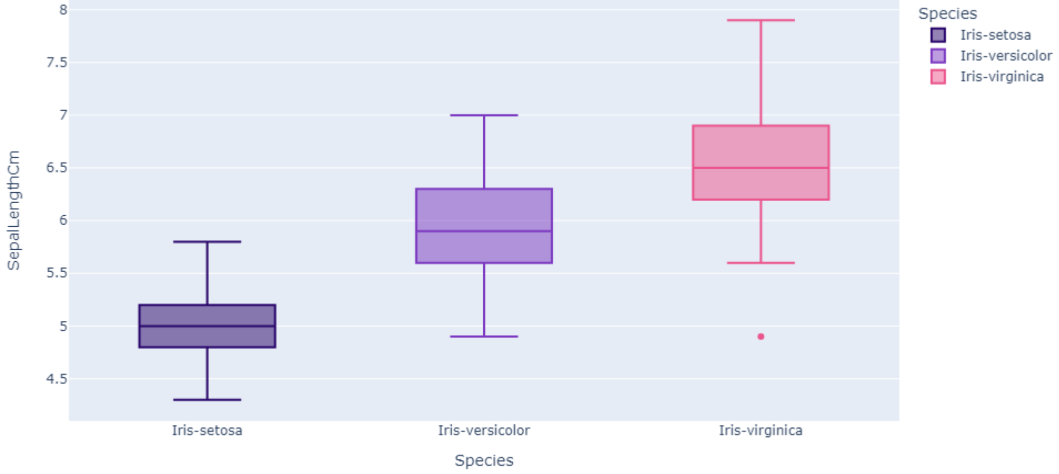


Fig 2: BOX PLOT

Box Plot:

With the Virginica and Versicolor classes having longer petals than the Setosa class, the box plot clearly illustrates the disparities in petal length between the classes. From these plots we can conclude that the Setosa has much smaller Sepallength than the other two classes and we can see that Virginica contains an outlier.

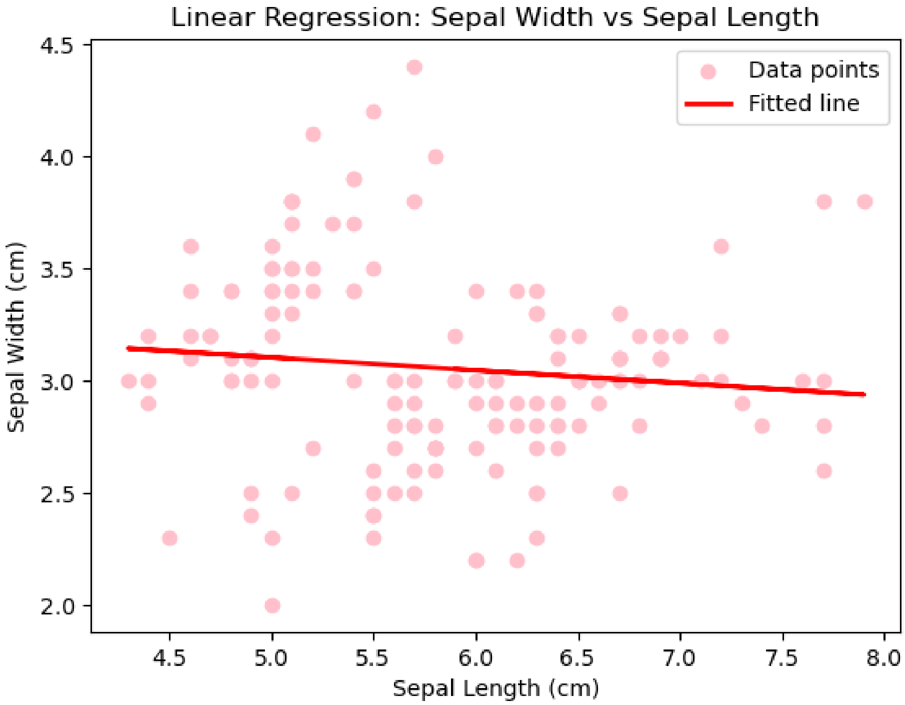


Fig 3: Scatter plot

The scatter plot shows that the length and width of the sepals on iris blooms is modeled in the graph using linear regression. The Iris dataset's Sepal Length and Sepal Width data are extracted, a linear regression model is fitted to the data, and the dataset is plotted alongside the fitted regression line. Understanding the two variables' linear relationship and evaluating how well the model captures it are made easier by the depiction.

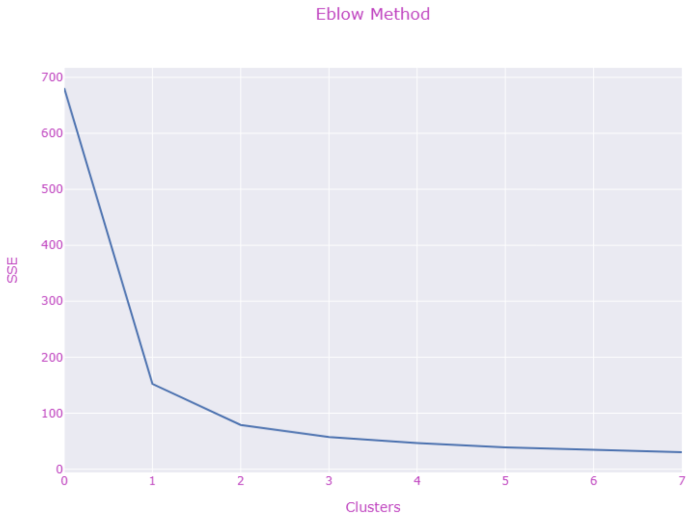


Fig 4: Elbow plot

The elbow graph shows a sharp decrease in distortion up to 3 components, indicating that 3 components are sufficient to capture most of the variation in the data. Beyond 3 components, the decrease in distortion becomes less significant, suggesting that adding more components may not provide much additional information. This will provide a good balance between reducing dimensionality and preserving the structure of the data.

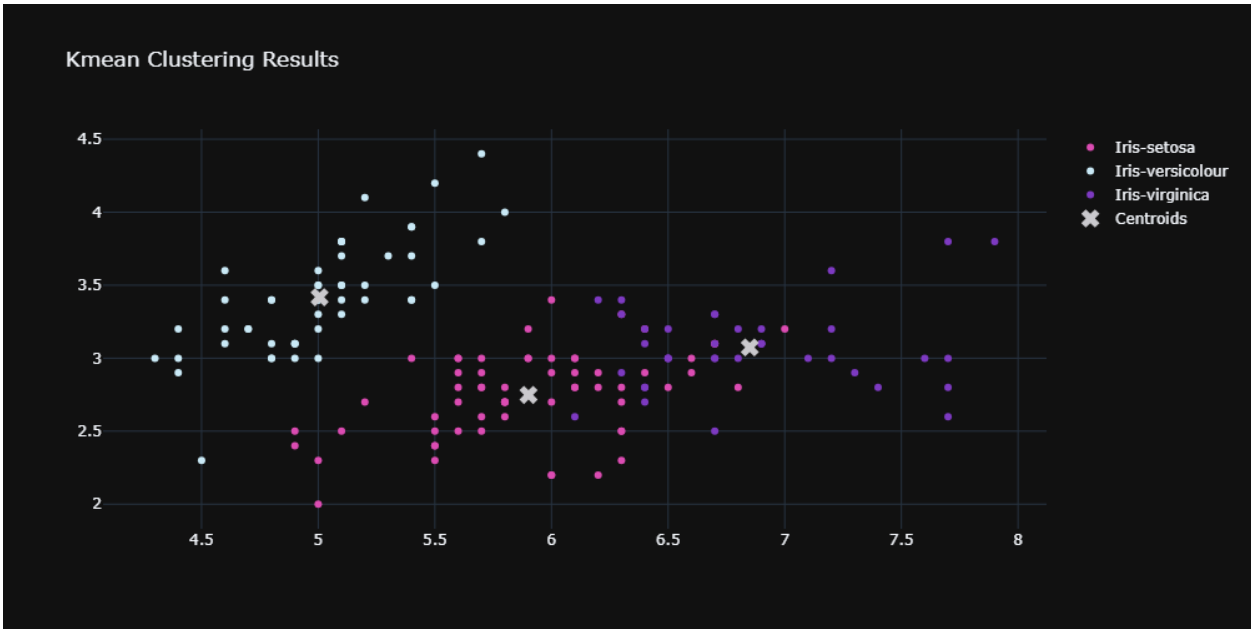


Fig 5: K-Means clustering of Species

Overall, the K-means clustering of the Iris dataset was successful in identifying the three species of iris flowers based on their sepal and petal measurements. The results demonstrate the potential of clustering algorithms for unsupervised learning and data exploration.