**Clustering and Fitting Report on Iris Data**

**Student Name:** Bhavanisankar Jammalamadugu **Student ID:** 23106158

**Repository Link:**

**Introduction**

I analyzed the Iris dataset using clustering and regression techniques to classify flowers based on their physical traits. The dataset consists of three species—Setosa, Versicolor, and Virginica—each with four features: sepal length, sepal width, petal length, and petal width. This study aims to explore feature distributions, apply K-Means clustering for classification, and use linear regression to establish relationships between variables. Additionally, I performed clustering prediction to classify unseen samples and fitting prediction to estimate petal length based on sepal length.

**Descriptive Statistics**

I calculated **mean, median, standard deviation, skewness, and kurtosis** for each feature (Figure 1). The results indicate that **sepal length is normally distributed, while petal length and width show high variability**. **Sepal width has a slight positive skew**, suggesting asymmetry. **Petal length and width have lower kurtosis**, meaning their values are more spread out. Understanding these patterns helps in classification and regression modeling.

A black text on a white background

AI-generated content may be incorrect.

Figure 1 Displays key statistical measures of the Iris dataset.

A screenshot of a graph

AI-generated content may be incorrect.**Feature Distributions Across Species**

I examined species-wise distributions using **boxplots** (Figure 2). **Setosa is distinct with smaller petal dimensions**, making it easy to classify. **Versicolor and Virginica overlap in sepal dimensions**, making them harder to distinguish. **Outliers exist, particularly in Setosa's sepal width**, indicating variability within species.

Figure 2 Illustrates species-wise distribution of features, highlighting key differences and overlaps.

A chart of different colors

AI-generated content may be incorrect.**Feature Relationships**

I used a **correlation heatmap** (Figure 3) to quantify feature relationships. The heatmap shows a **strong correlation (0.96) between petal length and petal width**, indicating redundancy in classification. **Sepal width has weak correlations**, making it less useful for clustering. The **species variable is strongly correlated with petal features**, confirming their importance in distinguishing different Iris types.

Figure 3 Illustrates the correlation strength between attributes, highlighting key relationships for clustering and classification

A graph with a line

AI-generated content may be incorrect.

**Clustering - Finding Optimal k**

To determine the best number of clusters, I applied the **Elbow Method** (Figure 4), which shows an optimal value of **k=3**, aligning with the actual number of species.

Figure 4 Identifies the best number of clusters by analyzing inertia values.

A diagram of a number of dots

AI-generated content may be incorrect.

**K-Means Clustering Results**

Using **k=3**, I applied **K-Means clustering** (Figure 6). The model successfully separated the three species, with **centroids aligning well with species means**. However, **Versicolor and Virginica overlap**, leading to minor misclassifications.

Figure 5 Displays the K-Means clusters with centroids.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Fitting Prediction (Regression Analysis)**

The **confidence interval plot** (Figure 6) shows the **linear regression model** used to predict **petal length from sepal length**. The regression line fits well, and the **confidence band** captures the uncertainty, ensuring most data points fall within an acceptable margin of error.

The **linear regression model plot** (Figure 7) further validates the fit, showing that the regression line effectively captures the **relationship between sepal length and petal length**. The strong alignment of data points confirms the model's reliability in predicting flower characteristics.

A graph with green dots and red line

AI-generated content may be incorrect.

Figure 6 Shows the regression line with confidence intervals to visualize prediction uncertainty.

A graph with green dots and a red line

AI-generated content may be incorrect.

Figure 7 Illustrates the accuracy of the fitted regression model for petal length estimation.

**Conclusion**

This study confirms that K-Means clustering effectively classifies the three Iris species, with petal features being the most distinguishing. The Elbow Method validated k=3 as the best clustering choice. Additionally, linear regression accurately predicted petal length, reinforcing feature relationships. Future improvements could involve Principal Component Analysis (PCA) for dimensionality reduction and Support Vector Machines (SVM) for enhanced classification. These findings highlight the effectiveness of clustering and regression in biological classification.