Read Me File from joint Github

# Iris_Analysis_Final

### Andrea Castro, Cyril Sambrano, Colin Leung, Sheila Brear

## Project Overview

This project aims to test a null hypothesis: That it is not possible to determine different Iris species by comparing measurements of different parts of the flower. The iris plant is a species of flowering plant, known for its variety of colors.

The iris dataset is a series of measurements of the length and width of the sepals and petals in three different species of iris: the Iris-setosa, the Iris-versicolor and the Iris-virginica. There are fifty samples of each of two measurements (petal and sepal) of each of the three species in the one dataset.

The original dataset describes 2 measurements, width and length, made on both:

Sepals: This is an outer structure of the iris bud, that protects the bloom before it opens. The shape and color of the sepal attracts pollinating insects to the plant.

Petals: This structure develops inside the sepal, and protects the reproductive, pollen-receptive reproductive structures of the plant.

The following questions will be answered in the data analysis that this group will perform.

1. Exploratory data analysis: is there a difference in the length of the petals of the three different species?
2. Scatterplot: Do the different species display different lengths and/or width of sepals and petals?
3. Decision tree: Can the different attributes of the sepals and petals predict the species?

### Data Overview

The Iris dataset contains the length and width of petals and sepals in three different species of irises. There are fifty measurements of each species in the four different categories (petal width and length, and sepal width and length).

### Machine Learning Model: Method and Results

After viewing the dataset for dimension, lack of absent or null values (4-10), the initial exploratory data analysis took place.

KDE plots were created of the total data so that the distribution of the measurements could be visualized (11-14).

Initial scatter plots were used to visualize any patterns in the distribution of the data (16-17).

A correlation matrix was then used to view distinct differences in sizes between the different species, comparing petal width and petal length across the species. It was clear that the species Iris Setosa had significantly smaller petall width and length than the other two species. The same process was repeated for sepal length and width (19)

A small portion of the data was trained, to see if it the petal and sepal measurements could be used to predict the species. This correlation showed how the variables related to one another, and also showed that the measurements of the petals and sepals can predict the idenitifcation of the Iris species(20). A random forrest classifier was then used to classify the data; this process showed that using the measurements to predict the Iris species has an accuracy of over 95% (21-22). A decision tree was used to see if the data could be classified by species, and it was found that using this machine learning tool showed an accuracy of over 86%. A K-Means algorithm was used to group the data into K clusters; the algorithm used in this process determines which cluster a species belongs to based on the measurements. The K means algorithm can be viewed in a scatterplot. (28). A process was then used to Fit the Model, which is a means to measure how closely the model can generalize data on which it was trained (29-33). sklearn was used to predict the response: It was found the accuracy was very high at 0.9 (34-36). Scatterplots were created to visualize the predictions, the final of which is a 3D scatter plot, that clearly shows the relationship between measurements of petal and sepal width and length and the Iris species. Conclusion: The null hypothesis was rejected.

### Database Integration

AWS was used to store and manipulate the data: very little needed to be done to this data set.

### Dashboard

A dashboard was created to provide a clear visualization of the data, clearly showing the relationship between petal and sepal measurements and the Iris species.

![dash](dash.png)

### link to dashboard:

[link to dashboard](https://public.tableau.com/app/profile/cyril.sambrano/viz/IrisAnalysis_16594059191210/SepalLengthY)

### link to presentation slides:

[link to google slides](https://docs.google.com/presentation/d/1TLXAITWZCAdblGrwjQDEPZtHc_QX7gqjuL4Oc_xWLVk/edit#slide=id.p)