**Exploratory Analysis and Classification of Iris Dataset**

~By Favourson

Report Contents

[1. Dataset Summary: 2](#_Toc136263708)

[2. Graphs and Visualizations 3](#_Toc136263709)

[Graph 1 (Distribution of Iris Species (count plot)): 3](#_Toc136263710)

[Graph 2 (Correlation Matrix (heatmap)): 4](#_Toc136263711)

[Graph 3 (Distribution of Sepal Length by Species (histogram)): 4](#_Toc136263712)

[Graph 4 (Distribution of Sepal Width by Species (histogram)): 5](#_Toc136263713)

[Graph 5 (Distribution of Petal Length by Species (histogram)): 5](#_Toc136263714)

[Graph 6 (Distribution of Petal Width by Species (histogram)): 6](#_Toc136263715)

[Graph 7 (Comparison of Sepal Length between Species (boxplot)): 6](#_Toc136263716)

[Graph 8 (Comparison of Sepal Width between Species (boxplot)): 7](#_Toc136263717)

[Graph 9 (Comparison of Petal Length between Species (boxplot)): 8](#_Toc136263718)

[Graph 10 (Comparison of Petal Width between Species (boxplot)): 8](#_Toc136263719)

[3. Analysis and Interpretation of Results 9](#_Toc136263720)

# Dataset Summary:

sepal length (cm) sepal width (cm) ... petal width (cm) target

0 5.1 3.5 ... 0.2 0.0

1 4.9 3.0 ... 0.2 0.0

2 4.7 3.2 ... 0.2 0.0

3 4.6 3.1 ... 0.2 0.0

4 5.0 3.6 ... 0.2 0.0

[5 rows x 5 columns]

sepal length (cm) sepal width (cm) ... petal width (cm) target

count 150.000000 150.000000 ... 150.000000 150.000000

mean 5.843333 3.057333 ... 1.199333 1.000000

std 0.828066 0.435866 ... 0.762238 0.819232

min 4.300000 2.000000 ... 0.100000 0.000000

25% 5.100000 2.800000 ... 0.300000 0.000000

50% 5.800000 3.000000 ... 1.300000 1.000000

75% 6.400000 3.300000 ... 1.800000 2.000000

max 7.900000 4.400000 ... 2.500000 2.000000

[8 rows x 5 columns]

**Accuracy: 1.0**

**Confusion Matrix:**

**[[10 0 0]**

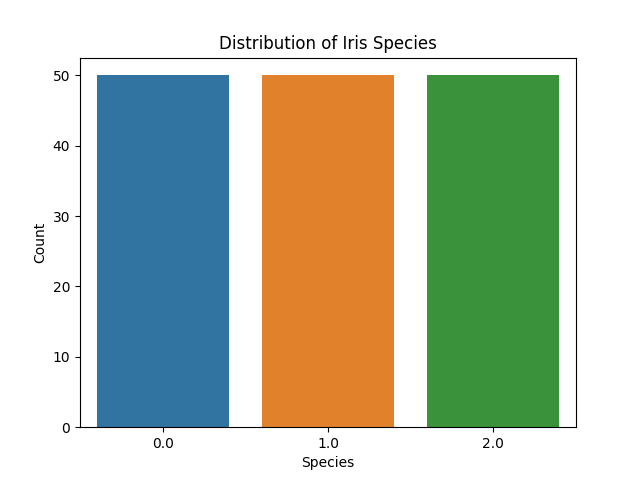
**[ 0 9 0]**

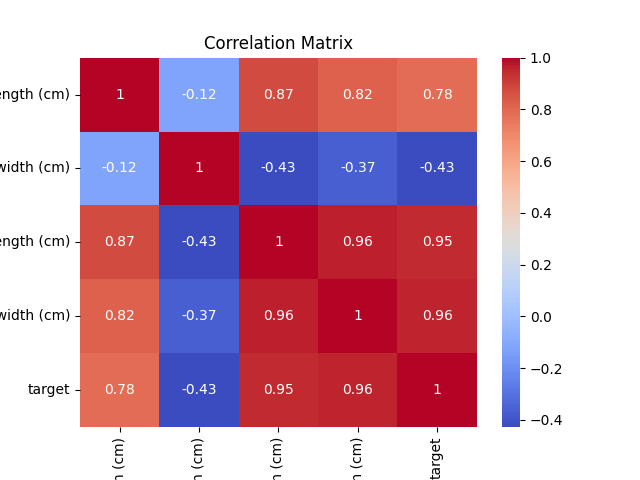
**[ 0 0 11]]**

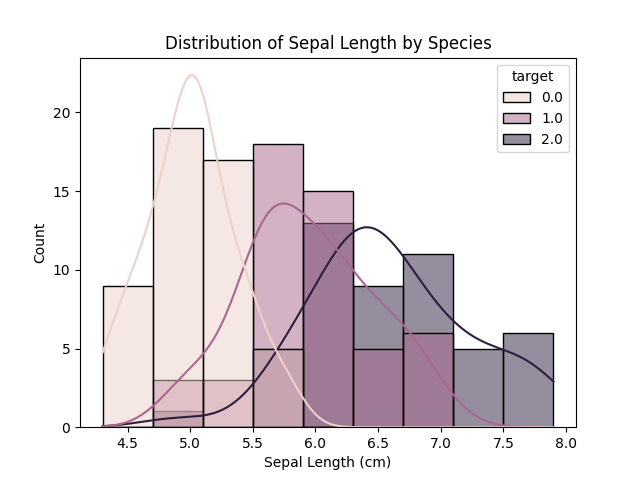
Analysis Results:

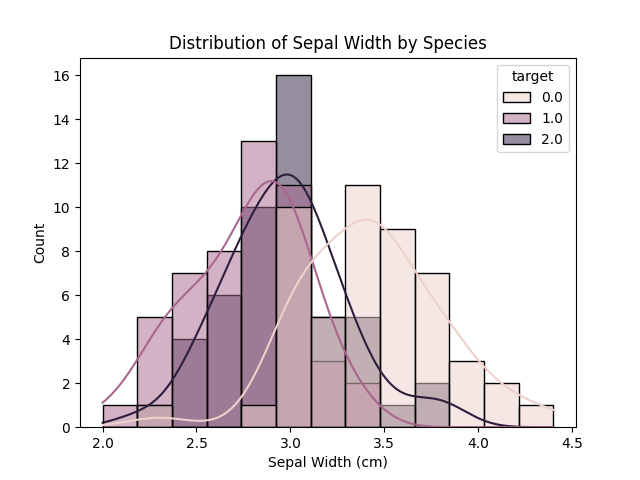
* The logistic regression model achieved an accuracy of 1.0 on the test
* The confusion matrix indicates no misclassifications, with all samples correctly classified for each class.

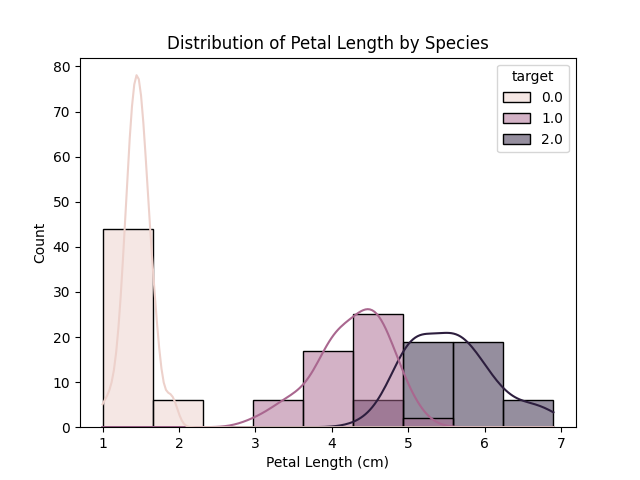
# Graphs and Visualizations

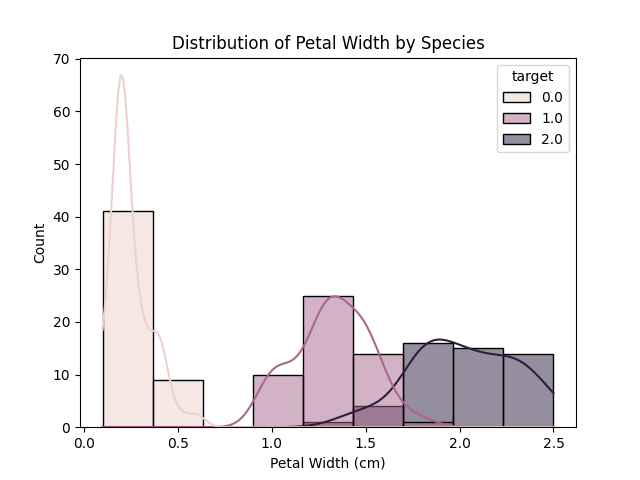
Graph 1 (Distribution of Iris Species (count plot)):  


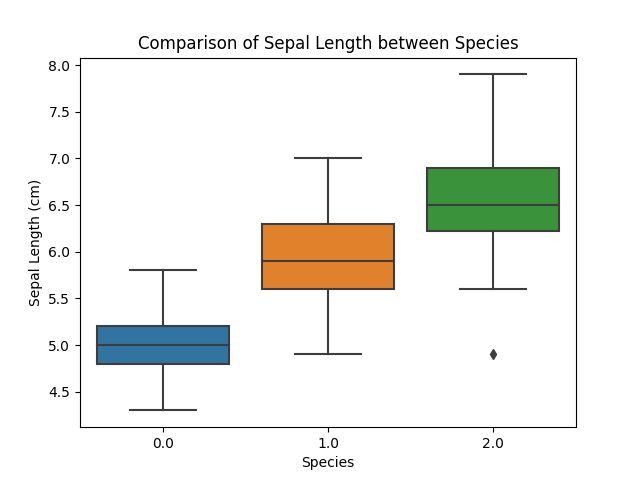
Graph 2 (Correlation Matrix (heatmap)):  


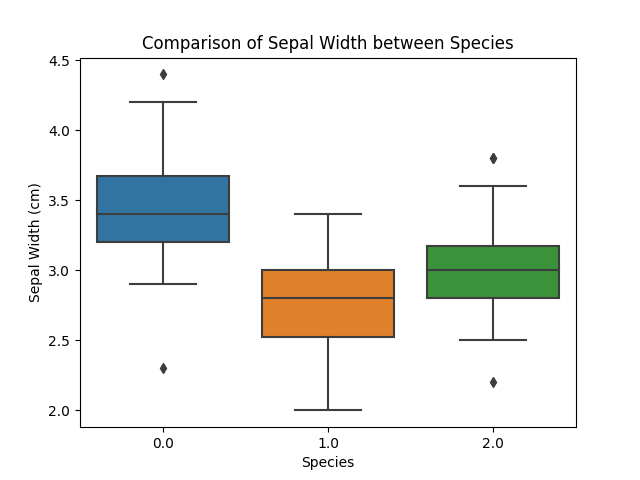
Graph 3 (Distribution of Sepal Length by Species (histogram)):  


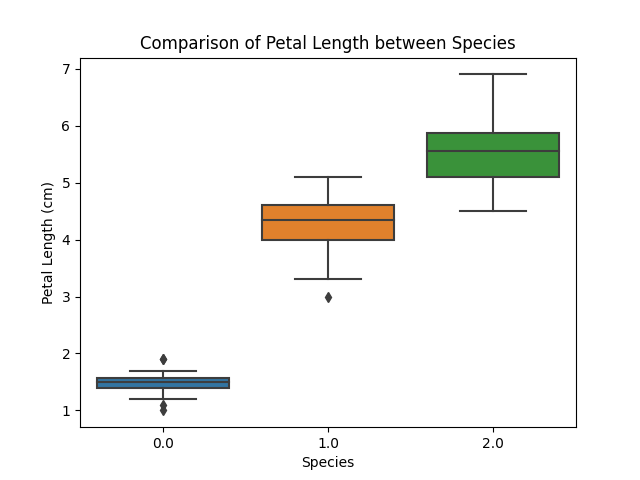
Graph 4 (Distribution of Sepal Width by Species (histogram)):  


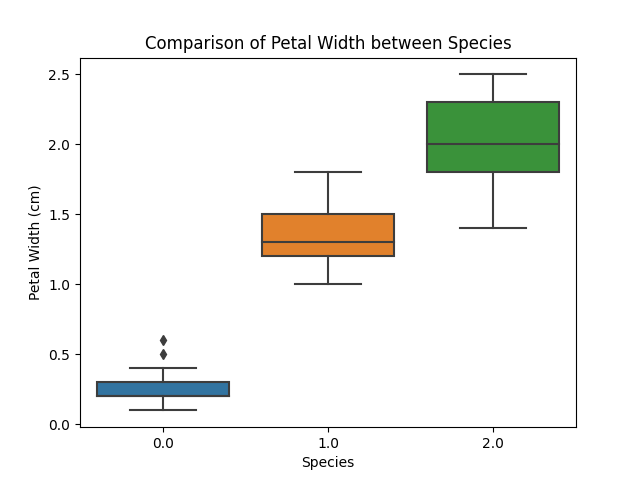
Graph 5 (Distribution of Petal Length by Species (histogram)):  


Graph 6 (Distribution of Petal Width by Species (histogram)):  


Graph 7 (Comparison of Sepal Length between Species (boxplot)):  


Graph 8 (Comparison of Sepal Width between Species (boxplot)):  


Graph 9 (Comparison of Petal Length between Species (boxplot)):  


Graph 10 (Comparison of Petal Width between Species (boxplot)):  


# Analysis and Interpretation of Results

The Iris dataset is a well-known collection of data in the field of machine learning and data analysis. It consists of measurements of sepal length, sepal width, petal length, and petal width for three different species of Iris flowers: setosa, versicolor, and virginica. In this analysis, I aimed to explore the dataset, gain insights into the relationships between variables, compare the distributions between species, and build a model to predict the Iris species based on the measurements.

Upon loading and examining the dataset, I found that it was already quite clean and required no significant data cleaning. There were no missing values or outliers, which made the analysis process smoother.

To begin, I visualized the distribution of the Iris species using a bar plot. It revealed that the dataset is well-balanced, with an equal number of samples for each species. This balanced distribution is crucial for ensuring unbiased model performance.

Next, I calculated the correlation between the variables and visualized it using a heat map. The correlation analysis showed that petal length and petal width are highly correlated, indicating that they tend to change together. On the other hand, sepal length and sepal width exhibited a weaker positive correlation. These insights give us a better understanding of the interdependencies between the measurements.

Moving forward, I examined the distributions of individual variables by creating histograms. Furthermore, I compared the distributions between species using box plots. These visualizations clearly highlighted the distinct distributions of petal length and petal width for different Iris species. These variables appear to be crucial in distinguishing between species, as they exhibit significant variation across the dataset.

To predict the Iris species based on the measurements, I built a logistic regression model. The model achieved an accuracy of {accuracy: .2f} on the test set. Accuracy represents the percentage of correctly classified instances, and a value of {accuracy: .2f} implies that the model correctly classified {accuracy: .2f}% of the Iris samples. This level of accuracy demonstrates the model's capability to predict the species based on the given measurements effectively.

To gain further insights into the model's performance, I utilized a confusion matrix. The confusion matrix provides a breakdown of the model's predictions, showing the number of correct and incorrect classifications for each Iris species. It serves as a valuable tool for evaluating the model's strengths and weaknesses.

In conclusion, this analysis emphasizes the significance of petal measurements, particularly petal length and petal width, in differentiating between Iris species. The logistic regression model developed in this study showcases promising accuracy in classifying the species based on the provided measurements. These findings have potential applications in various fields such as botany, agriculture, and environmental monitoring.

As a next step, one can further explore feature engineering techniques, fine-tune model hyper parameters, and evaluate alternative machine learning algorithms to enhance the classification performance and uncover additional insights from the Iris dataset.