Data Science Project

Tanishk Thomas

189302075

Dataset Description:

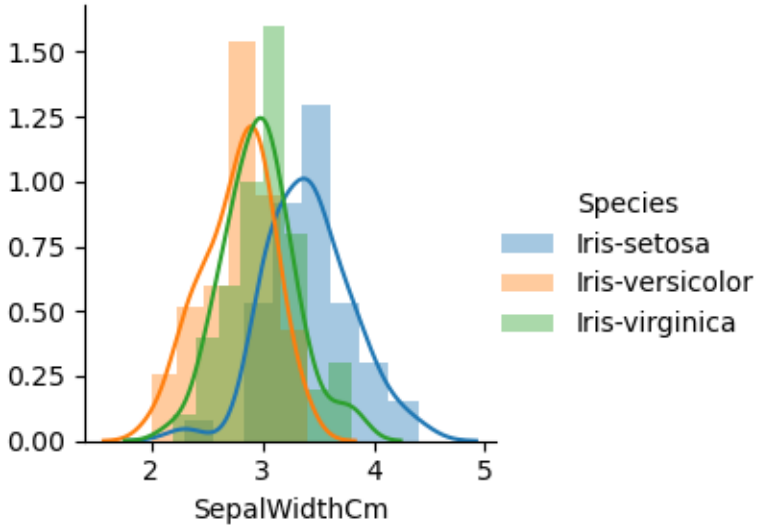
The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris-setosa, Iris-virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification, and clustering examples and to test algorithms.

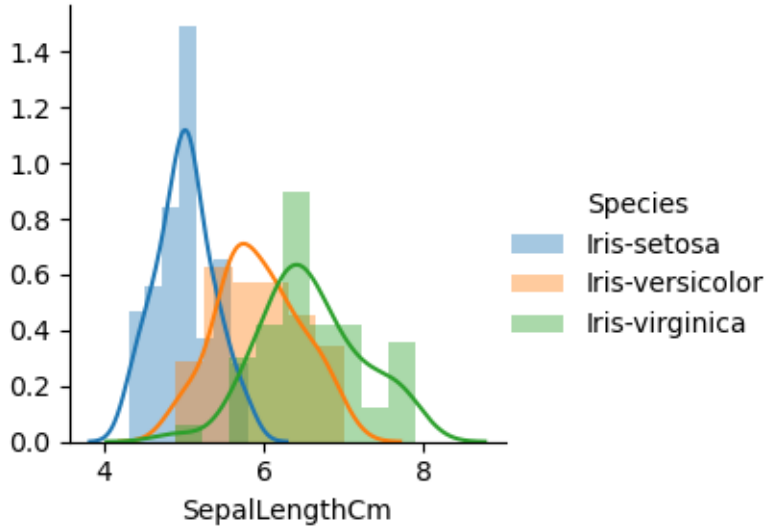
Algorithm Used :

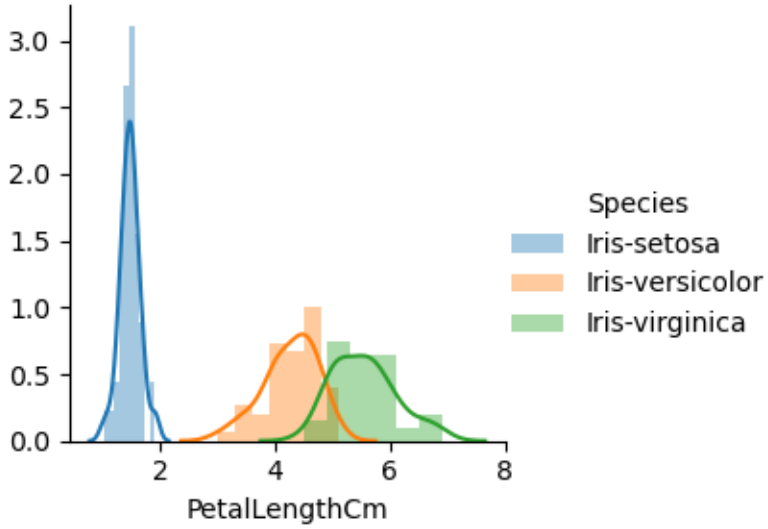
Decision tree algorithm is used to perform classification on the IRIS dataset. The algorithm uses **Information Entropy** to identify the segments in the data and perform the respective classification. Decision Tree (DT) is a powerful machine learning algorithm used for both classification as well as regression. DT can be a tree type of structure where each internal node is a test condition for the vector to move further and the terminal nodes represent the class or the prediction value to be predicted. DT are good for the classification of a few class labels but do not produce proper results if there are many classes and less training observations. And moreover, DTs can be expensive to train computationally.

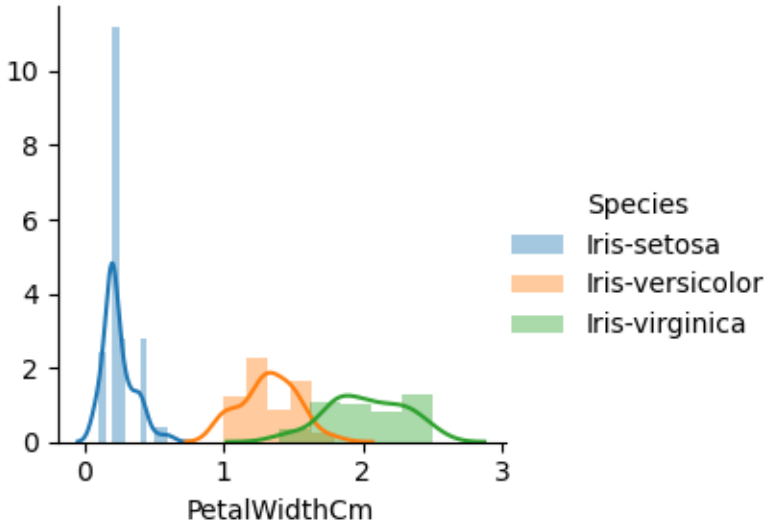
Data Visualisation :

Probability Density Function Graphs -



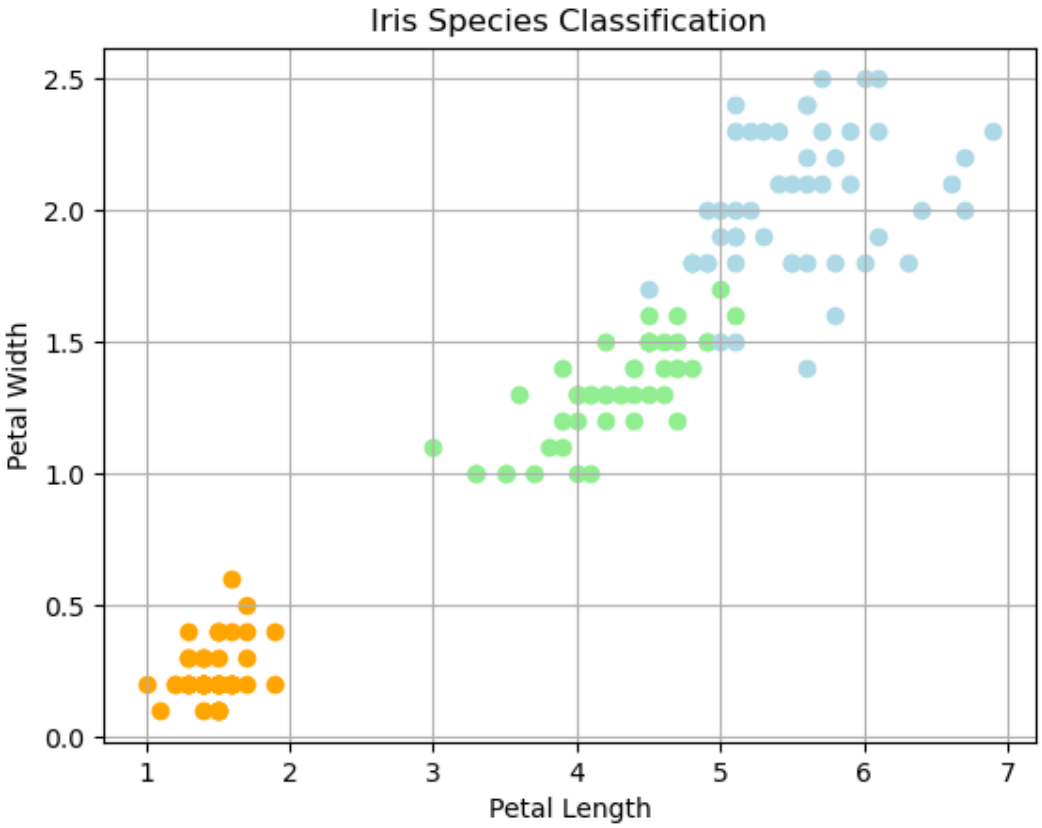






Inference: *Using Distribution plots we cannot infer much information but can only separate the species Iris-Setosa. So, now we use a scatter plot to visually classify the species based on petal length and petal width, since both are a distinguishing factor.*

Scatter Plot of PetalLength Vs PetalWidth :



Results :

Confusion Matrix:

60:40 Train-Test Split :

[23, 0, 0]
[0, 19, 0]
[0, 1, 17]

70:30 Train-Test Split :

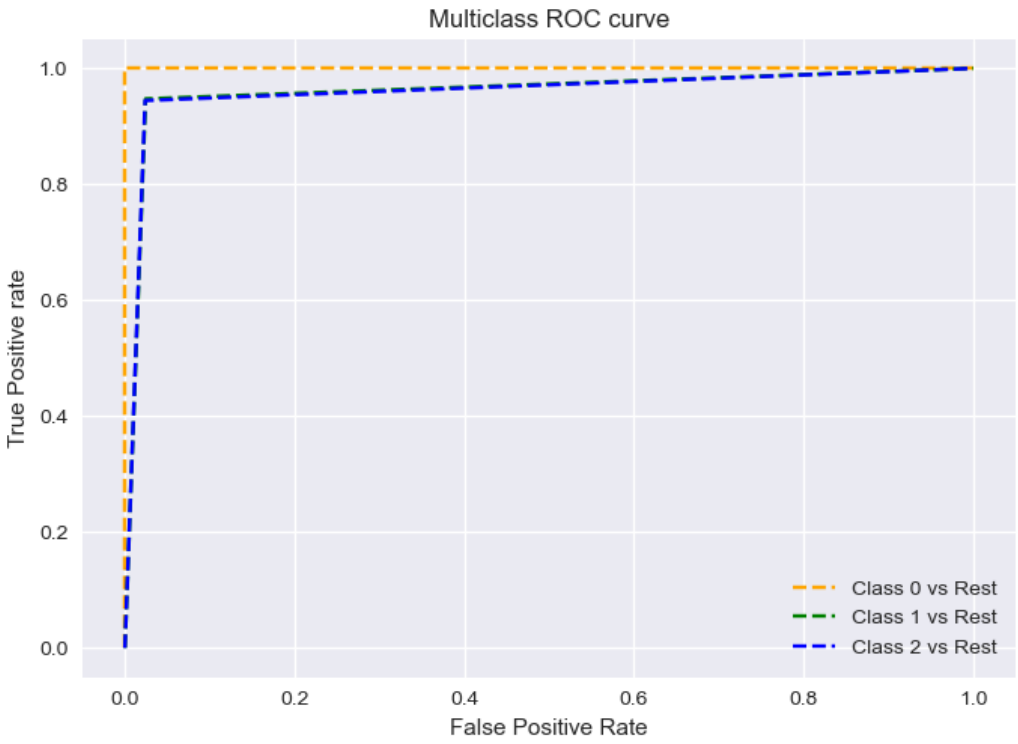
[19, 0, 0]
[0, 11, 2]
[0, 0, 13]

80:20 Train-Test Split :

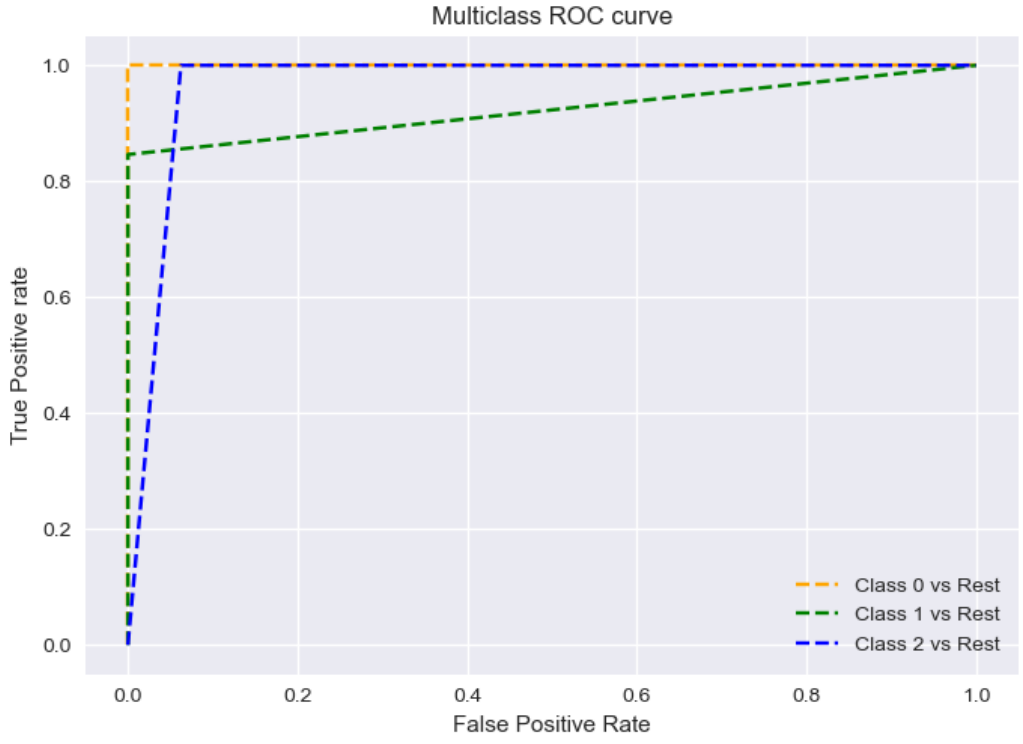
[10, 0, 0]
[0, 9, 0]
[0, 0, 11]

Receiver Operating Characteristic Curve:

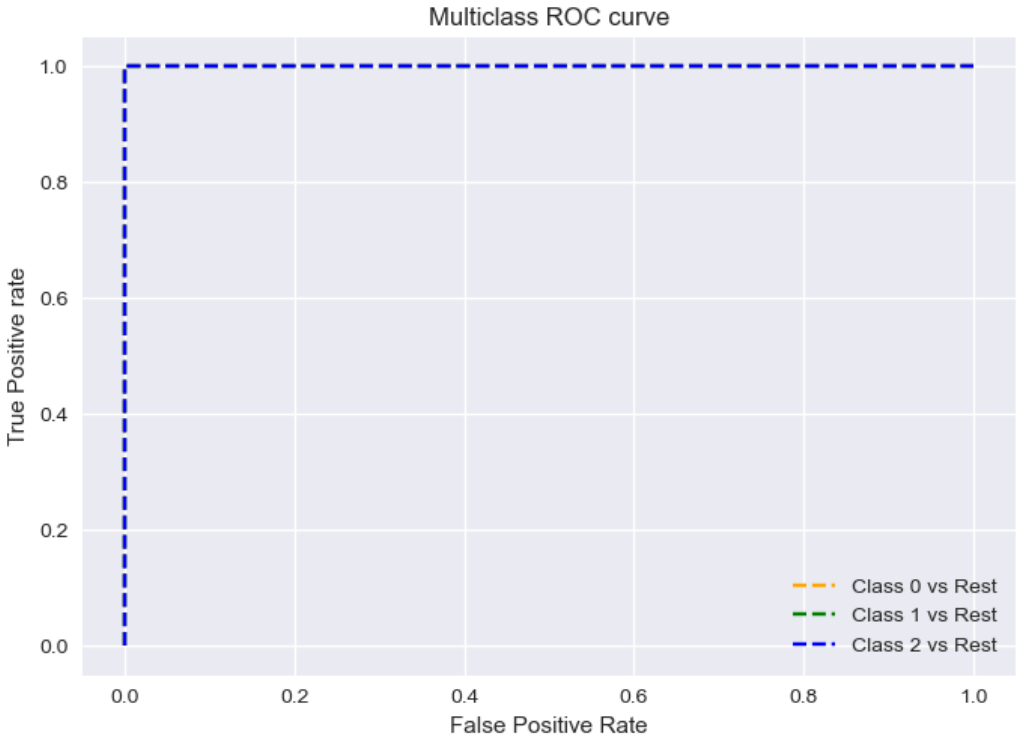
60:40 Train-Test Split:



70:30 Train-Test Split:



80:20 Train-Test Split:



Evaluation Metrics : (S1: Specificity; S2: Sensitivity)

60:40 Train-Test Split:

	Precision	Recall	F1-Score	FPR	FNR	NPV	FDR	MCC	S1	S2
Setosa	1.00	1.00	1.00							
Versicolor	0.95	1.00	0.97							
Virginica	1.00	0.94	0.97							
				0.008	0.016	0.991	0.016	0.975	0.992	0.983
Accuracy			0.98							
Macro Avg	0.98	0.98	0.98							
Weighted Avg	0.98	0.98	0.98							

70:30 Train-Test Split:

	Precision	Recall	F1-Score	FPR	FNR	NPV	FDR	MCC	S1	S2
Setosa	1.00	1.00	1.00							
Versicolor	0.93	1.00	0.96							
Virginica	1.00	0.92	0.96							
				0.022	0.044	0.977	0.044	0.933	0.977	0.956
Accuracy			0.98							
Macro Avg	0.98	0.97	0.97							
Weighted Avg	0.98	0.98	0.98							

80:20 Train-Test Split:

	Precision	Recall	F1-Score	FPR	FNR	NPV	FDR	MCC	S1	S2
Setosa	1.00	1.00	1.00							
Versicolor	1.00	1.00	1.00							
Virginica	1.00	1.00	1.00							
				0.00	0.00	1.00	0.00	1.000	1.000	1.000
Accuracy			1.00							
Macro Avg	1.00	1.00	1.00							
Weighted Avg	1.00	1.00	1.00							

Conclusion :

We get the best logical results when the Train-Test split ratio is 60:40. When the ratio is 80:20, the model overfits the data and hence we get cent percent evaluation values. For the 70:30 split, the results are average.