Introduction

The Iris dataset is a widely-used dataset in the domain of machine learning and pattern recognition. It is one of the most fundamental datasets. This report encapsulates a meticulous analysis of the balanced Iris dataset, delineating the various methodologies employed, the challenges encountered, and the insights gleaned from the balanced data with a total of **150** data points with each class having **50** data points. Three classes of Iris include *Iris-setosa*, *Iris-versicolor* and *Iris-virginica.*

Approach

In undertaking the analysis of the Iris dataset, a commonly used structured and systematic pipeline was adopted to ensure a comprehensive understanding of the data, optimal preprocessing, and effective modeling, which is similar to any machine learning pipeline. The approach encompassed the following key stages:

1. Exploratory Data Analysis (EDA)

At the outset, an in-depth Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset's characteristics and distributions. This foundational step entailed *Histograms*, *Scatter plots* and *box plots* to visualize the distribution of each feature and also to discern the inter-relationships between different features.

2. Preprocessing and Feature Engineering

A pivotal phase in the data science pipeline i.e. preprocessing, was meticulously undertaken to ensure data integrity and readiness for modeling. Key preprocessing steps encompassed Missing Value Handling, Data Splitting and Label encoding.

3. Data Modeling

With the data suitably preprocessed, the modeling phase ensued, leveraging a diverse suite of algorithms to ascertain the most adept model for the classification task. The Algorithms mainly employed includes:

a) **Logistic Regression:** A quintessential algorithm, renowned for its simplicity and efficacy in binary classification scenarios.
b) **Decision Tree Classifier:** A versatile non-linear classifier, adept at partitioning feature spaces to delineate distinct classes.
c) **Random Forest:** Harnessing the power of ensemble learning, Random Forests amalgamate multiple decision trees to amplify prediction accuracy and resilience to overfitting.

d) **Support Vector Classifier:** Embracing the principles of maximizing the margin between classes, SVC endeavored to discern a hyperplane that optimally separates the 3 distinct Iris species.

Some Relevant Observations

1. From the distributions of the *4* features, *SepalLengthCm* seemed to have a right skewed one similar to the *PetalWidthCm* and these features are positively correlated as well with a correlation coefficient of **0.82**. On the other hand *SepalWidthCm* feature has a normal distribution and the distribution of *PetalLengthCm* is observed to be partly normal (on the right side) and right skewed. The Plots are shown below for reference.

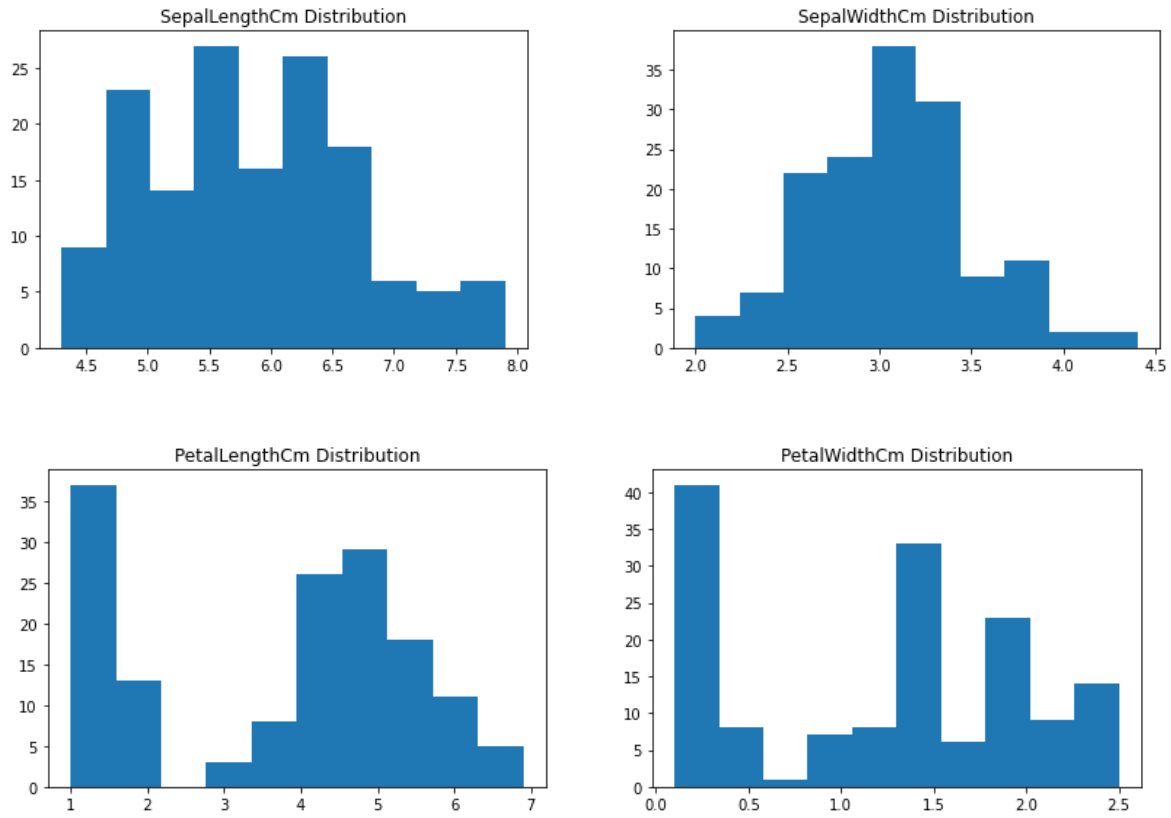

Fig 1. Histograms

2. For the correlation among features, we observed that *PetaWidthCm* and *PetalLengthCm* features are highly correlated with a coefficient of **0.96**, so the *PetalLengthCm* and *SepalLengthCm* with a coefficient of **0.87** (indicated by their scatter and box plots as well). In the opposite case, *SepalLengthCm* and *SepalWidthCm* features are not closely

correlated with a very low coefficient. Some plots are given below along with the heatmap.

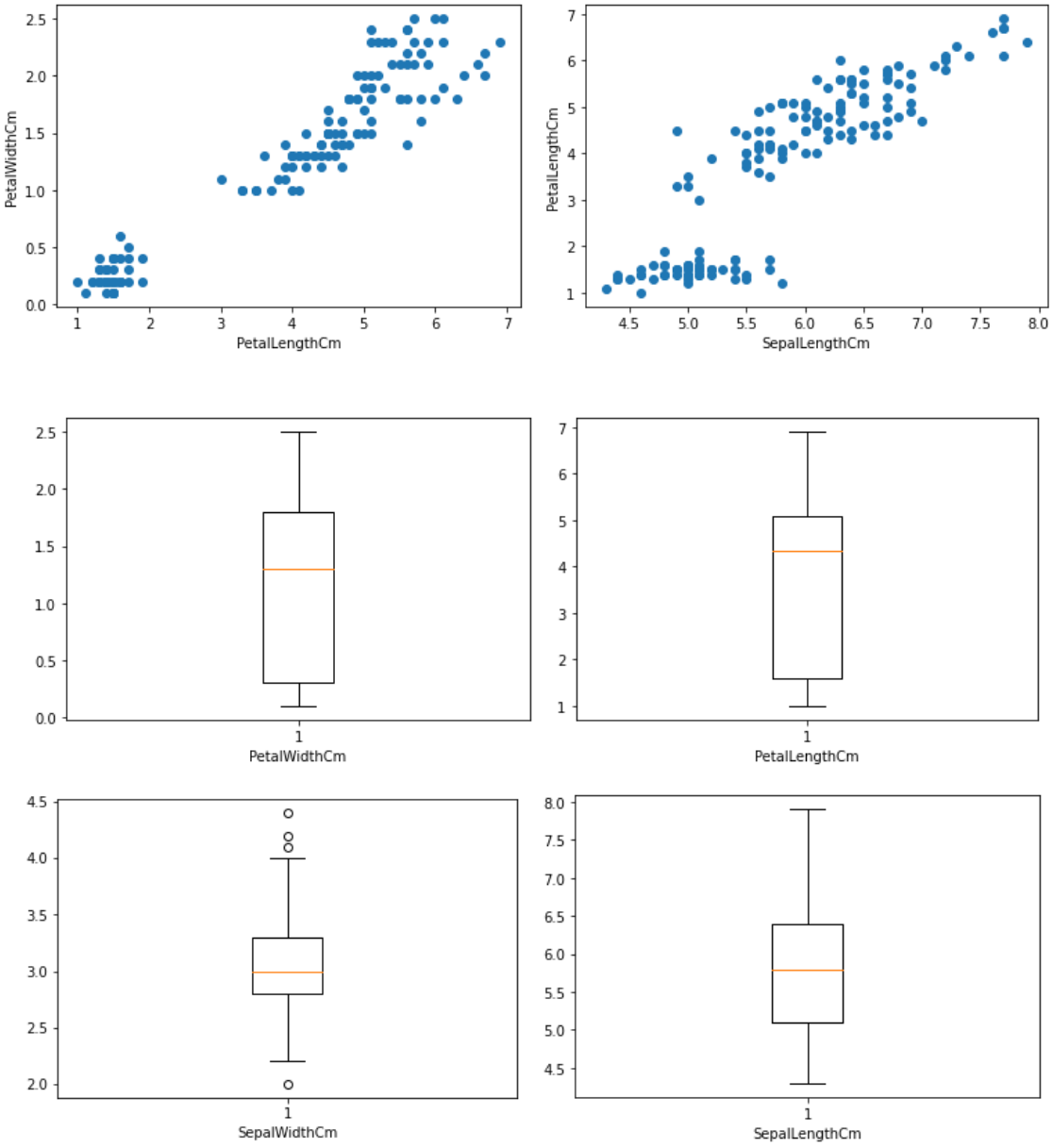

Fig 2. Box and Scatter Plots

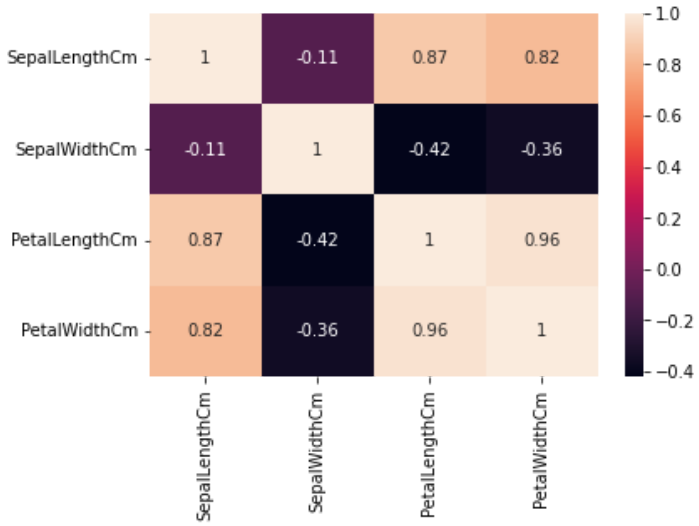

Fig 3. Heatmap

3. 3 Labels were in textual format, so we have to convert them into integers, otherwise it will not be possible to train the model, thus we have simply encoded the 3 labels as 0,1 and 2.There is also no *NaN* values in the dataset and even the dataset is *balanced*, hence no sampling techniques is involved over here.
4. We have splitted the models, with 3 different test sizes of 0.1, 0.2 and 0.3. The results obtained came best in case of *0.1* for *Logistic Regression* and *SVC*, while accuracy remains the same for ensemble methods like *Decision Tree* and *Random Forest*, making them the most out performing models in this case with a **100%** accuracy. Some statistics on performance of other models is shown below.

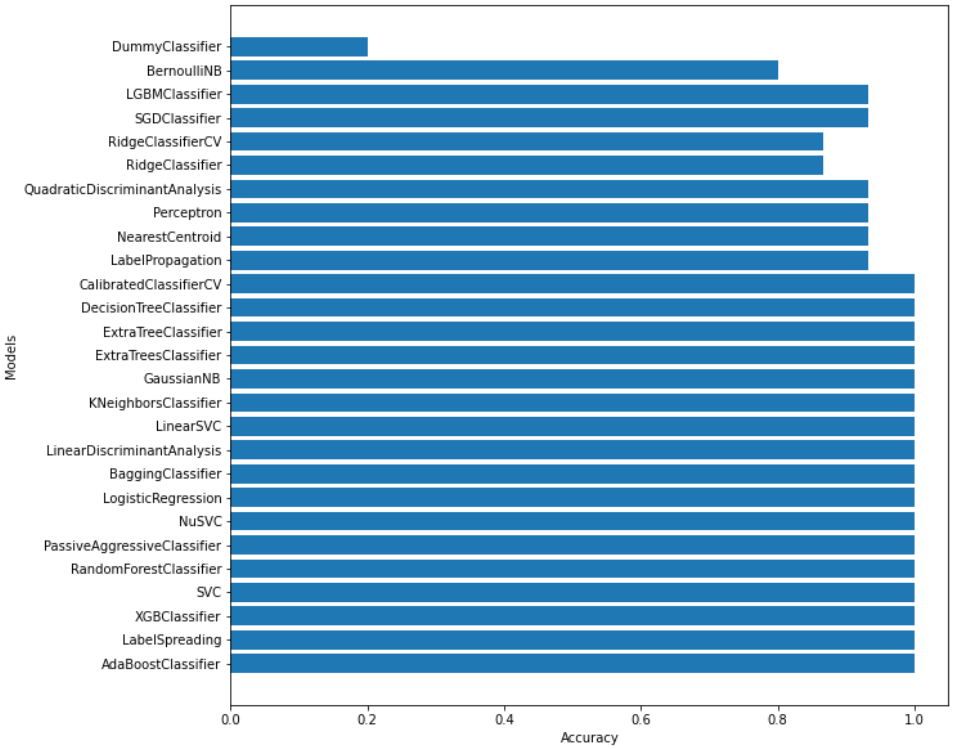

Fig 4. Performance of various models

Challenges Faced

1. **Model Selection:** Choosing the appropriate algorithm was a major issue as more than 1 model was out performing. Thus we have to test multiple models with lots of different hyper-parameters and it consumes a lot of time.
2. **Hyperparameter Tuning:** Finding optimal hyperparameters for each model to enhance performance is a necessity.
3. **Overfitting:** Most important issue we have is that the dataset was extremely small and most likely for any model to overfit. Thus, we have to work on multiple test splits and random forest deals with over-fitting very diligently.

Results and Conclusion

In the analysis of the Iris dataset, the modeling phase revealed compelling results with the *Random Forest* and *Decision Tree classifiers* both achieving a remarkable **100%** accuracy on the test dataset, underscoring their unparalleled adeptness in discerning intricate patterns and facilitating flawless classification across the 3 distinct Iris species. Random Forest chooses a random split one at a time and tests itself against that split, by training on other splits with a majority voting of N (we chose *n_estimators* as standard as **100**) decision trees. Thus **Random Forest** or **Decision Tree** has obtained the maximum score in this case.