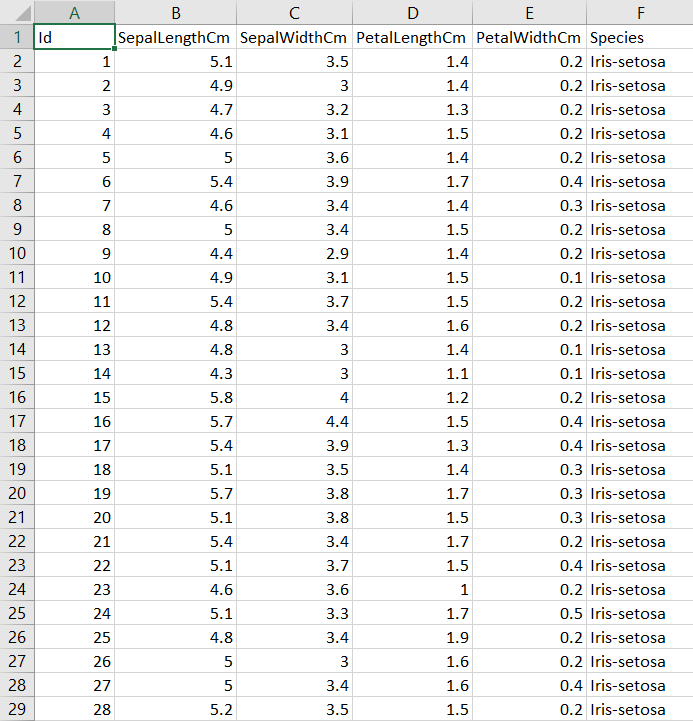
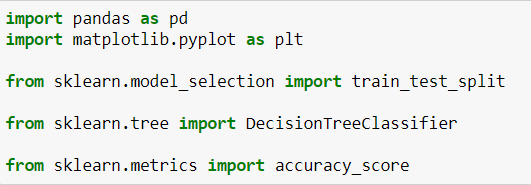
**Code Documentation: Assignment#4**

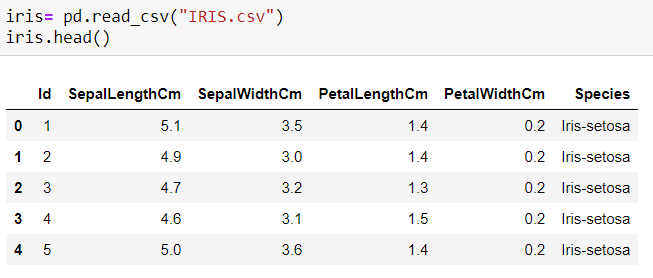
Downloaded the dataset from the Kaggle link provided [here](https://www.kaggle.com/uciml/iris) to solve the problem given. The Iris dataset consists of six columns namely: **Id, SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm** and **Species**. IRIS dataset is a very standard dataset in the field of Data Science. The dataset includes three iris species with 50 samples as well as some properties about each flower.



Importing the libraries (pandas, train\_test\_split, matplotlib.pyplt, DecisionTreeClassifier, accuracy\_score) which will be used in the code as the first step.



The next step is to create a DataFrame **iris** by exporting the csv file to Jupyter notebook using the **read\_csv()** method present in the pandas library. Later, we print the first five rows of the DataFrame using the **head()** function.

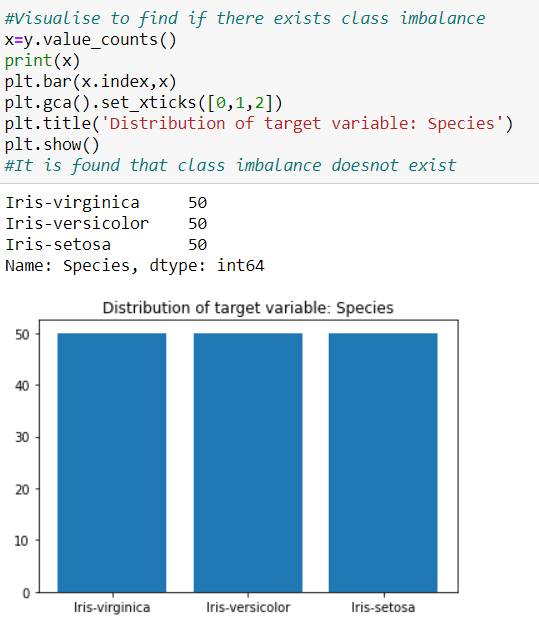


According to the problem statement, we are supposed to understand the Decision Tree Algorithm. Then split the IRIS dataset into 2 subsets: **Training Dataset (70% rows)** and **Testing Dataset (30% rows)**. Later implement the decision tree classifier algorithm on the IRIS Training dataset. We are instructed to test the accuracy with Testing Dataset. We are allowed to use any ready-made library for decision tree if needed.

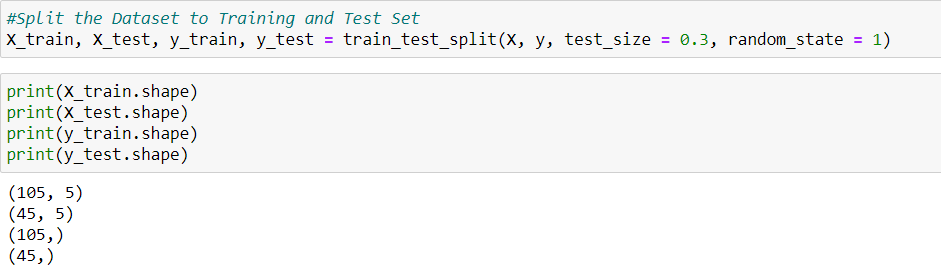
As the first step of the Data Analysis, we split the given IRIS Dataset into ***X (Dataset containing all independent variables)*** and ***y (Dataset containing the Target variable)***.



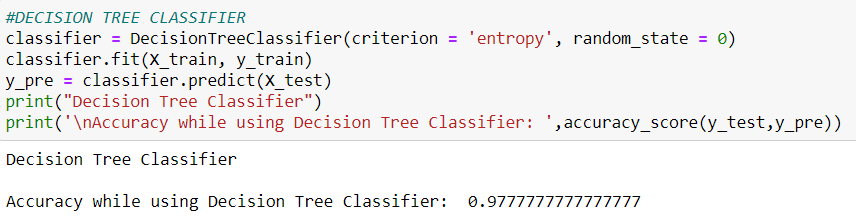
When applying any machine learning model to a Dataset, if there exists class imbalance in the dataset, the accuracy in the results reduce substantially. Therefore, we check if there exists any class imbalance in the given dataset IRIS and the below given results show that there ***doesn’t exist any class imbalance*** in the given IRIS dataset as the distribution (**Species: Iris-versicolor, Iris-setosa and iris-virginica**) present in the Target class is equal (**50-50-50**).



The next step is to split the given dataset into Training and Test Dataset as given in the problem statement (**70:30**) using the **train\_test\_split** function present in the **sklearn.model\_selection**. Later, we check if the split made has been done properly by using the **shape** parameter.



As the last step in our Data Analysis, we prepare a **Decision Tree Classifier** named **classifier** and set the criterion to “**entropy**” and fit the **X\_train** and **y\_train** to train the model. Using the classifier, we predict the y values (**y\_pre**) given the data to be **X\_test**. Comparing the original **y\_test** values with the y\_pre values gives the overall **accuracy** of the model.



The results show that the model which we trained has an accuracy of **97.77%** using the Decision Tree Classifier.