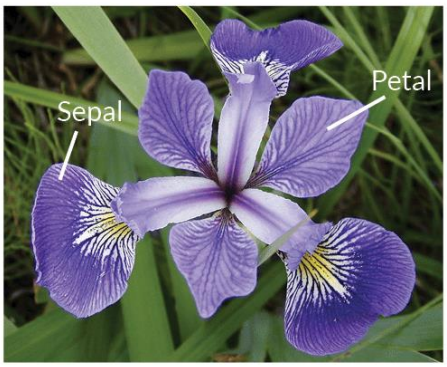

Iris Versicolor

Iris Setosa

Iris Virginica

Iris Dataset report

Ibrahim Tarek Mohamed | intern2grow Internship| 23/11/2023 |Github

DATA PREPRATION

1. After loading the data using pandas package the first step is to check how many samples on each classes using groupby function in the Class column.
2. Checking the columns datatype this will help us to known which columns is a categorical and convert it.
3. Converting the object column into a categorical than use the cat.codes function to label encoding, because there is only 3 classes it good to use the label encoding than one-hot encode as the result would be 0, 1, 2 the distance between them is close not like if we have 100 classes from 0, 100 label encoding would be a problem.
4. There is no feature need to be removed all feature affect the class label.
5. Finally store the class column in variable Y and features in variable X and split it into train and test and normalize the data using MinMaxScaler.

$$MinMaxScaler = \frac{X - X_{min}}{X_{max} - X_{min}}$$

MODEL SELECTION

The model used for this problem is KNN model from sklearn package because it's the simplest model which use the distance measure the classify the current point.

Another model that is good in this problem is linear regression model because the L1 and L2 parameter that can be set for it's loss function that provide regularization but as the problem is very sample KNN can perform fast and accurate.

Model Evaluation

I used accuracy score as my evaluation matrix there is also many metrices that can be use like F1 score R1 score precision and recall.

The model accuracy was 96.67 %