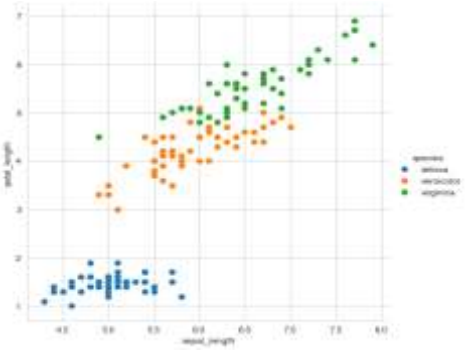
Analysis and Visualization of Iris Dataset

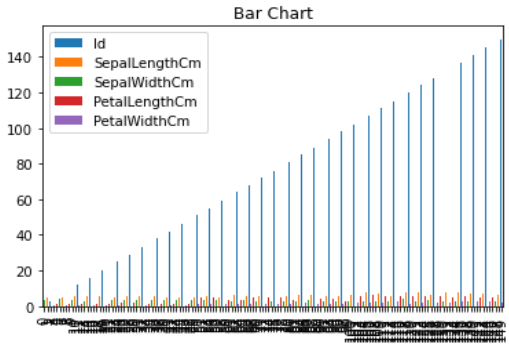
This report discusses Exploratory Data Analysis, a technique for analysing data that employs some visual techniques. We can obtain detailed information about the statistical summary of the data using this technique.

Iris dataset is one of the popular data set mostly used in the field of pattern recognition. The dataset consists of 3 classes of 50 instances each, where each class mostly refers to Iris plant. One of the classes is directly separable from others, the others 2 are not separable from each other.

There are no Null Entries in any of the columns, four columns are numerical in nature, and only one column is categorical in nature.



The above scattering shows the divergence of petal length vs petal width. Iris Setosa species has smaller sepal length but higher width.

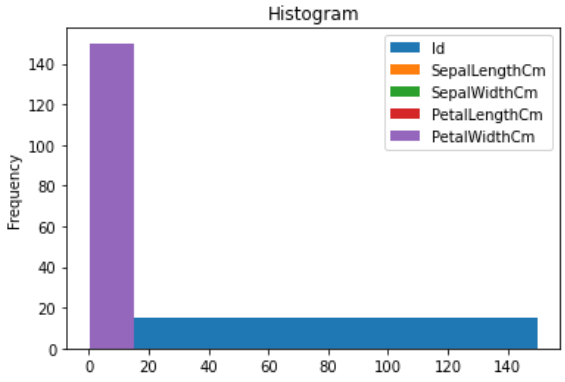


The results show changes in id, sepal length, sepal width, petal length, and petal width, starting from zero to maximum value 140.

DATASET from Kaggle.com

Statistical analysis shows the results however, the class diagnosis for this data set is same that is 33.3% for all 3 classes.

This further visualizes that species are well balanced. Each species (Iris Virginica, Setosa, versicolor) has 50 as its count.



So, we can use Petal Length and Petal Width as the classification feature. According to data analysis, the Iris Setosa species has a shorter sepal length but a wider sepal width. Versicolor is almost in the middle in terms of length and width, whereas Virginica has longer sepal lengths and narrower sepal widths.

DATASET from Kaggle.com