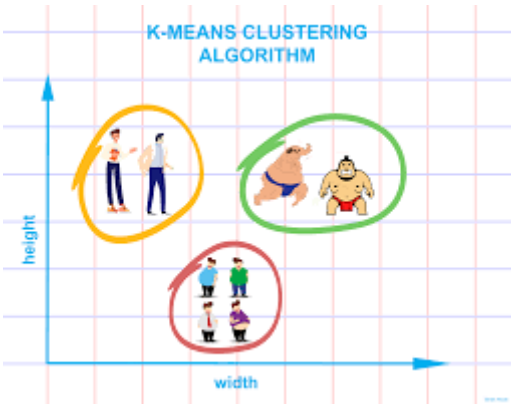
Assignment 1 Report

Introduction

Unsupervised learning, a subset of machine learning algorithms used when the given data is unlabelled i.e., when there is no output column. All Unsupervised learning algorithms fall under one of the two categories, either Clustering or Association Rule Learning. In Clustering, the given data is grouped into desired number of clusters based on their inputs. The grouping or clustering is done by the machine, i.e., there is no need to specify any logic behind the grouping process, only the input data and number of clusters is enough. For more accuracy some hyperparameter tuning is done. Most commonly used Clustering algorithm is KMeans algorithm. Association Rule Learning is used in learning the relation between 2 or more events, similar to the Clustering algorithm, all logic is taken care by the machine i.e., the algorithm. Most commonly used Association Rule Learning algorithm is Apriori algorithm.

Algorithm

KMeans algorithm is based on the K Nearest Neighbours algorithm. For this algorithm we need to specify the number of clusters or groups to be made and give in the input data, after 300(default) epochs we get our clusters.



Inner workings:

Number of clusters is selected and the centres of the clusters are randomly selected initially. Then the nearest object or element to the centre is put in the cluster and the centre of the cluster is recalculated, i.e., the object is taken as the new centre. From there, smallest possible diameter is drawn such that all objects fall into one of the n clusters.

Usage:

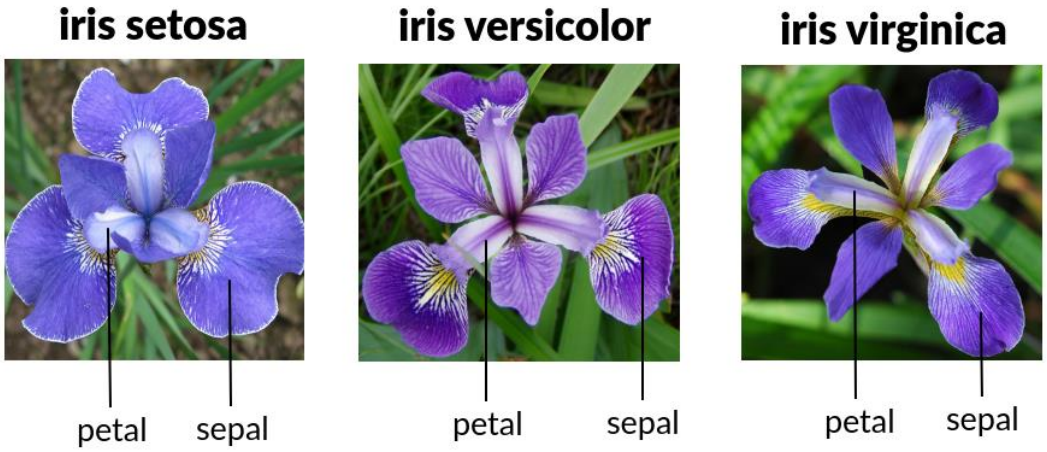
```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3)
kmeans.fit(data)

preds = kmeans.predict(data)
```

Dataset

The given dataset, *Iris.csv* contains 6 columns and 151 rows including header, the shape of the given data set is (150, 6) excluding header. The columns include, Id, Sepal Length in cm, Sepal Width in cm, Petal Length in cm, Petal Width in cm and the Species of the flower. The data was uniformly distributed along Species, i.e., there are 3 different species and each has 50 different collections thus making 150 rows without header.

In this given data, Id column is unnecessary and can be dropped. As we are using Unsupervised learning, Species column also can be dropped. We can store Species column in separate variable to check the accuracy of our algorithm at the end of the prediction, if needed.

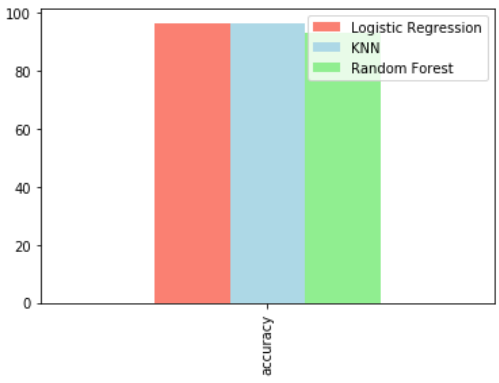
The given data was full, without any null or NaN values and did not need any data pre-processing.



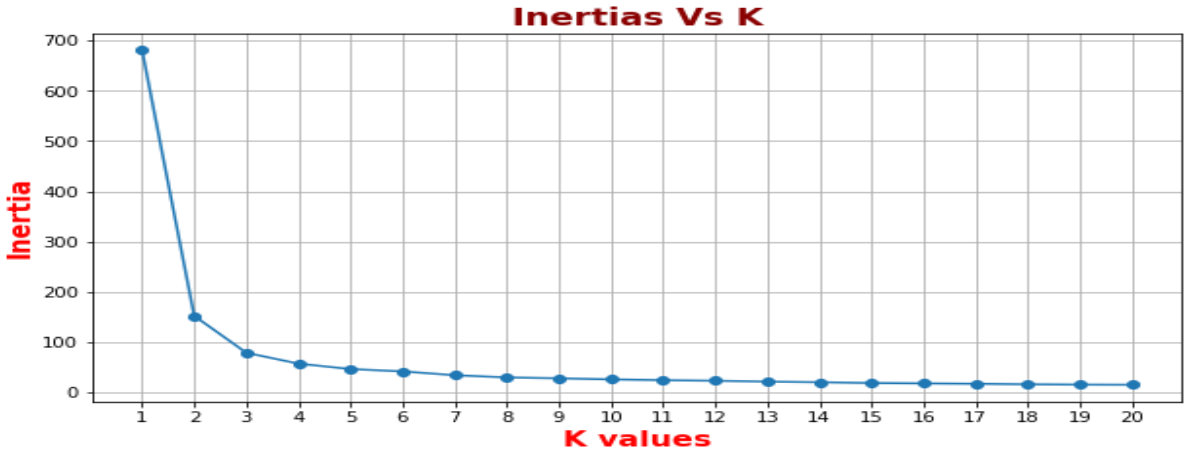
This is one of the most common and famous datasets for machine learning.

Graphs and Plots

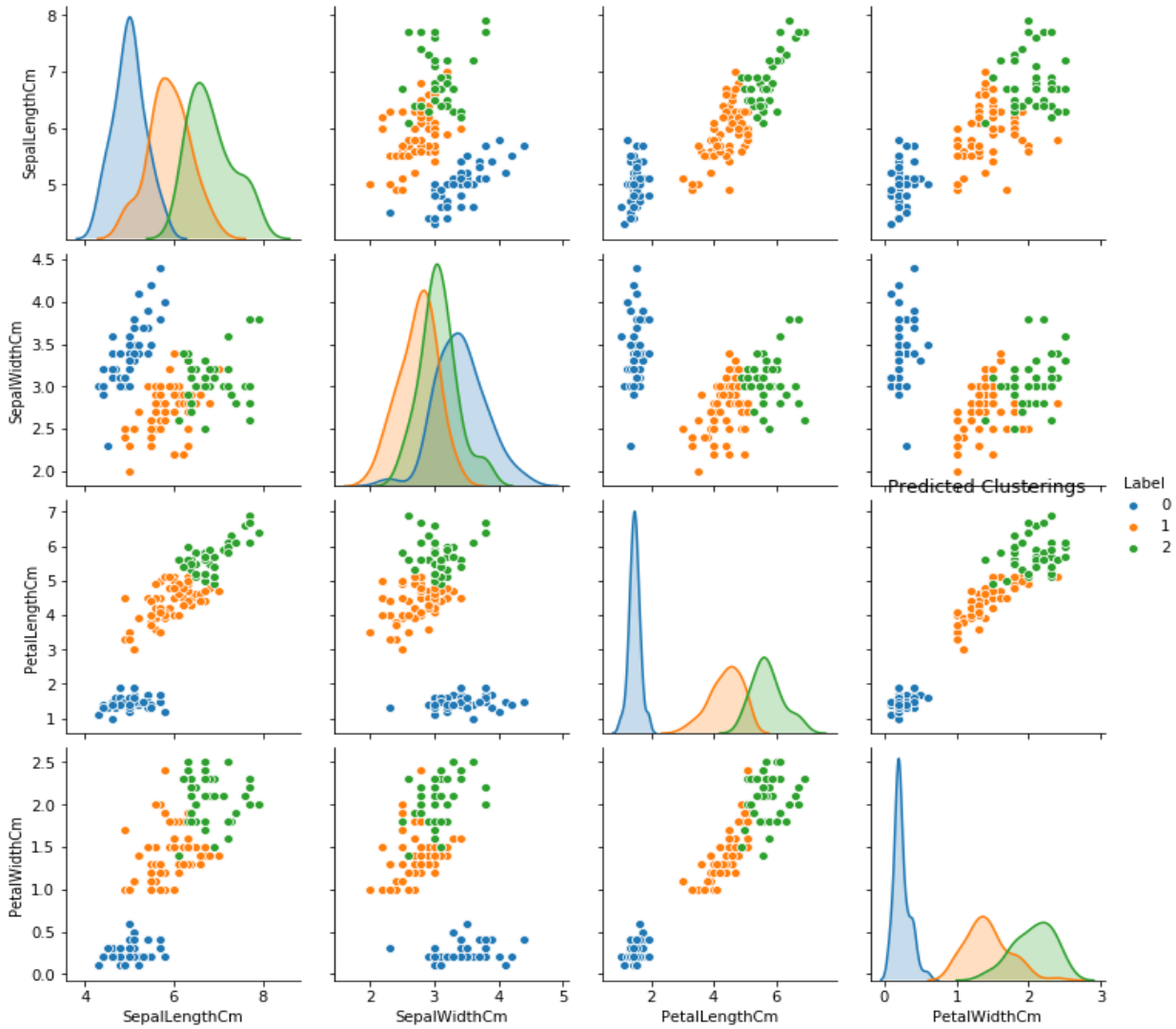
1. The accuracy achieved by 3 different supervised algorithms is plotted, in which Logistic Regression and KNN was good.



2. The curve between kmeans inertia and k value is plotted, for lower k values the inertia is higher.



3. The pair plot has been plotted, to visualize distribution of 3 species based on each given parameter.



Inference

From the given data and respective plots, we can infer,

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	small	large	small	small
1	medium	small	medium	medium
2	large	medium	large	large

Applications:

1. KMeans Algorithm can be used in Recommendation systems where similar people can be grouped and recommended based on the grouping.
2. KMeans Algorithm can also be used for virus classification based on their particular properties.