Comparision of F1 Scores and Performance of 3 Clustering Algorithms on Iris Data Set

Sumit Mishra

AITS

7 August, 2019

New Delhi, India

sumit.mishra0432@gmail.com

ABSTRACT

Iris dataset is one of the basic datasets. It contains data of various species of flower of Iris plant. SepalLength, SepalWidth, PetalLength, PetalWitdh and Specis are the data contained in this data set.

It includes three iris species that are Iris-Setosa, Iris-Versicolor, Iris-Virginica with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

It's the Assignment-3 given to me as a Machine Learning Intern.I have done the Exploratory data Analysis, Preprocessing, modelled three clustering algorithms and have compared the f1 scores and performances of these 3 Clustering Algorithms.

1. INTRODUCTION

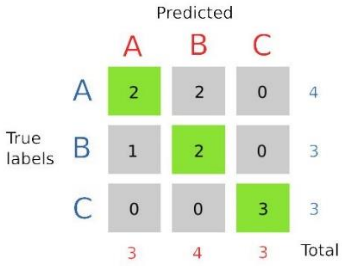
Clustering is a technique to categorize the data into groups. Distance metrics plays a very important role in the clustering process. There are number of algorithms which are available for clustering. In general, K-means is a heuristic algorithm that partitions a data set into K clusters by minimizing the sum of squared distance in each cluster. The algorithm consists of three main steps: a) initialization by setting center points (or initial centroids) with a given K, b) Dividing all data points into K clusters based on K current centroids, and c) updating K centroids based on newly formed clusters. It is clear that the algorithm always converges after several iterations of repeating steps b) and c).

2. METRIC OVERVIEW

- F1 Score is the most common metric for Clustering Algorithms.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

- Confusion Matrix is also used in comparing the Performance of the Clustering Algorithms.

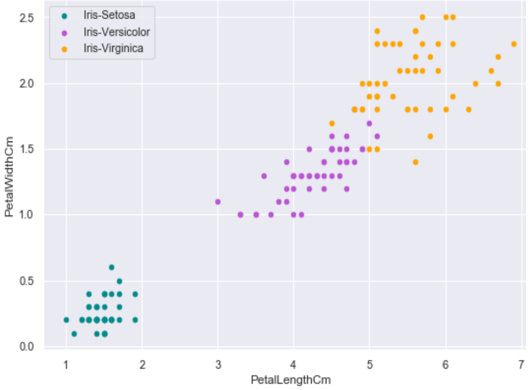


		Predicted			
		A	B	C	
True labels	A	2	2	0	4
	B	1	2	0	3
	C	0	0	3	3
		3	4	3	Total

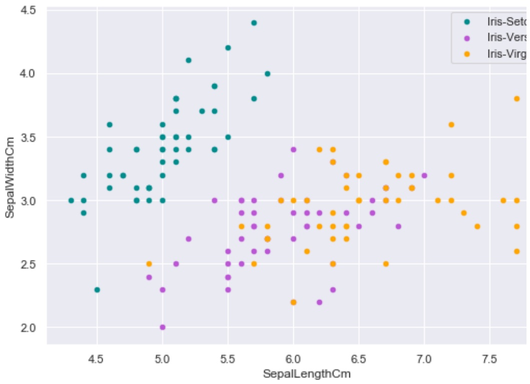
3. EXPLORATORY DATA ANALYSIS

I have done some Exploratory Data Analysis on the Iris Data Set to gain some insight on the data and get as much as information from the data in form of data visualization.

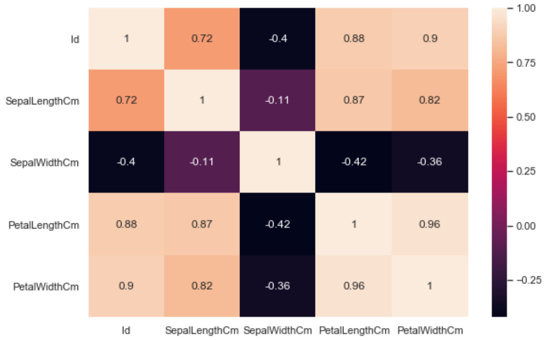
- This scatter plot shows the relationship between the petal width and petal length. By this we can say that the Iris Setosa is linearly separable from other two classes.



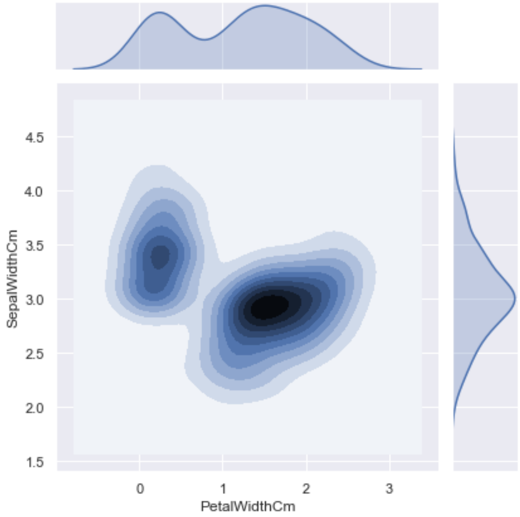
- This scatter plot shows the relationship between the sepal width and sepal length.

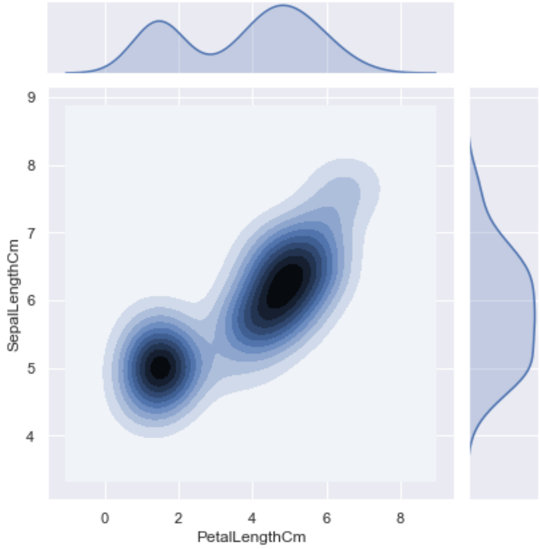


- The Heatmap of the correlation matrix has also been plot to know the feature dependency. By this we know if there's a positive or negative dependency of the features.



- There are also the joint plots that shows the cross relation between the features i.e. Sepal length and petal length, Sepal width and petal width.





- Box plots are also plotted against the 3 different Species to get to know about the outliers of the different Features as per the interquartile range.

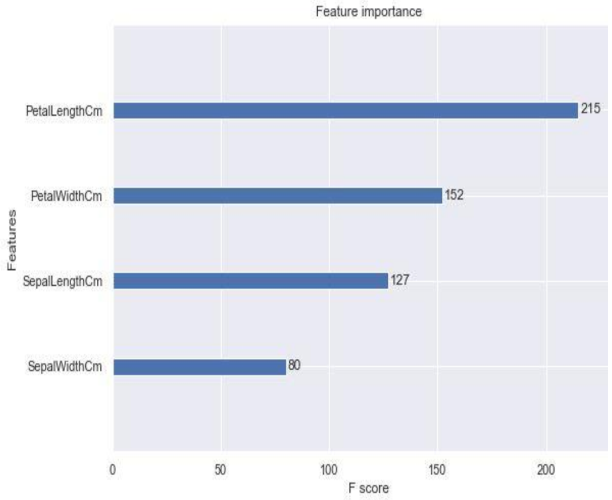
4. Preprocessing

The difference of Mean and Median between the different features are checked get the idea about the outliers of the features. I have replaced the categorical values of the Species i.e. the name of the Species is replaced by the numerical values.

I have used the SepalWidthCm, SepalLengthCm, PetalWidthCm, PetalLengthCm as the features for the model and the Species as the Target Variable for the Clustering.

I have used the train_test_split from the sklearn model selection to split the Data set into Train and test set. The 70% of the Data is used as the train data and 30% as the test data.

I have used the Boosted Gradient Descent to get insight of the feature importance and plotted the feature importance bar graph. From the graph is can be clearly seen that the Petal Length is the most important feature in the dataset.



5. Clustering Models

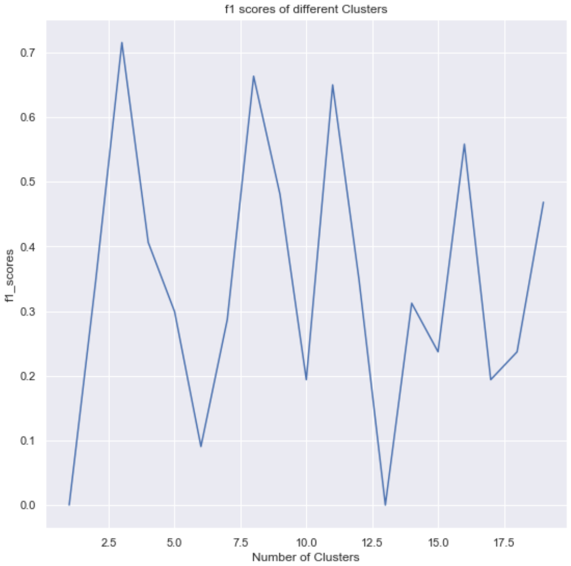
I have used 3 Clustering models which are K-Means Clustering, Mean Shift Clustering and Agglomerative Clustering.

5.1 K-Means Clustering

***k*-means clustering** is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the clusters.
This algorithm aims at minimizing an objective function know as squared error function given by.

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

I have used a loop to iterate over the number of clusters and hence taken the best model for the final prediction.



In the Clustering of the Iris Data set K-Means Algorithms gives good f1 score of 0.7156 and an accuracy of 75.56%.

5.2 Mean Shift Clustering

Mean shift clustering aims to discover “blobs” in a smooth density of samples. It is a centroid-based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region.

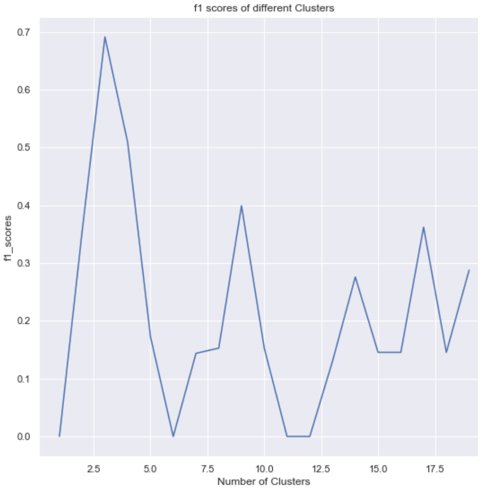
$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x)x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

The mean shifting Clustering Algorithms has an accuracy of 62.22% and a f1 Score of 0.6756.

5.3 Agglomerative Clustering

Agglomerative Clustering algorithms is a hierarchal clustering algorithm. It recursively merges the pair of clusters that minimally increases a given linkage distance.

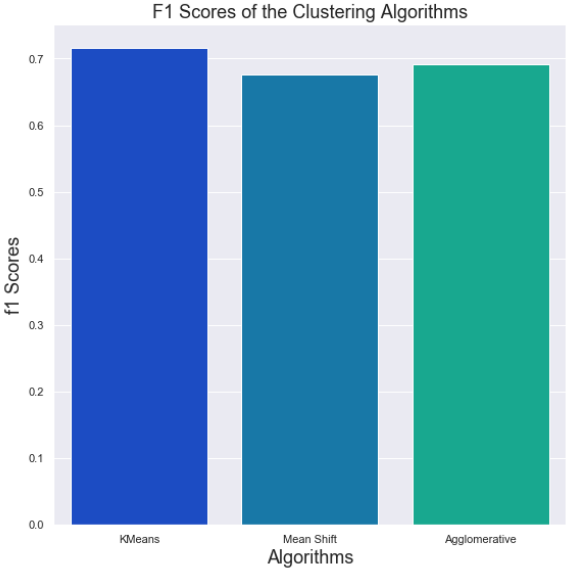
I have also used a loop to iterate over the number of clusters and used the best model for the final prediction.



Agglomerative Clustering had an accuracy of 64.44% and a f1 Score of 0.6910.

6. Result

I have plotted a bar graph that compares the f1 scores of the three clustering algorithms.



The K-Means Clustering Algorithm is the best performing model with a f1 score of 0.7156.

7. Conclusion

I have performed the Exploratory Data Analysis on the Iris Data set and Preprocessed it for the modelling and successfully Modelled 3 Clustering Algorithms and compared their f1 Scores and Found out that the K-means is the best performing clustering algorithm on the iris data set with a f1 Score of 0.7156.