Contents

Project Abstract.

Two clustering approaches namely; K-means clustering and hierarchical clustering are to be implemented on a dataset of multivariate valued data instances. For the K-means clustering, the optimal k value is to be reported. And for the hierarchical clustering, the optimal clustering level is to be reported. A comparison of the results of the two clustering methods is to be carried out.

Also, the clustering accuracies are to be compared, when both the methods are implemented on the first four features i.e., 'Mean of the integrated profile', 'Standard deviation of the integrated profile', 'Excess kurtosis of the integrated profile', 'Skewness of the integrated profile' and the last four features i.e., 'Mean of the DM-SNR curve', 'Standard deviation of the DM-SNR curve', 'Excess kurtosis of the DM-SNR curve' and 'Skewness of the DM-SNR curve'.

For computation and clustering and visualization, Python and the Weka software will be used. Google Collab is utilized as the web IDE and Weka software version 3.8.5 will be employed to work with ARFF data file.

Dataset description.

The dataset used in this project is named 'HTRU 2', created in 14 February 2017. The authors and donors of this dataset is Rob Lyon of The University of Manchester, Manchester. This dataset is a sample of the pulsars candidates gathered during the 'High Time Resolution Universe Survey'.

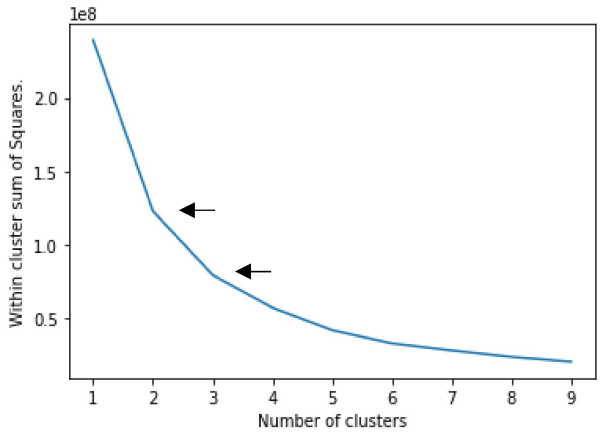
Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter (D. R. Lorimer and M. Kramer, 2005).

The dataset in use contains 16,259 dummy examples caused by noise, and 1,639 real pulsar examples. Hence, there are a total of 1,639 positive instances and 16,259 negative instances; which comes up to a total of 17,898 data instances.

A total of 9 attributes make-up the dataset. There are 8 continuous variables in the dataset by the following names; 'Mean of the integrated profile', 'Standard deviation of the integrated profile', 'Excess kurtosis of the integrated profile', 'Skewness of the integrated profile', 'Mean of the DM-SNR curve', 'Standard deviation of the DM-SNR curve', 'Excess kurtosis of the DM-SNR curve' and 'Skewness of the DM-SNR curve'. And the 9th attribute is reserved for the class labels. The class labels are 0 for negative and 1 for positive.

Elbow Method.

The Elbow Method is an effective approach to finding out the optimal number of clusters. This method provides a line graph which indicates the what would be the ideal number of clusters. Python's matplotlib library is used as a data visualization tool.

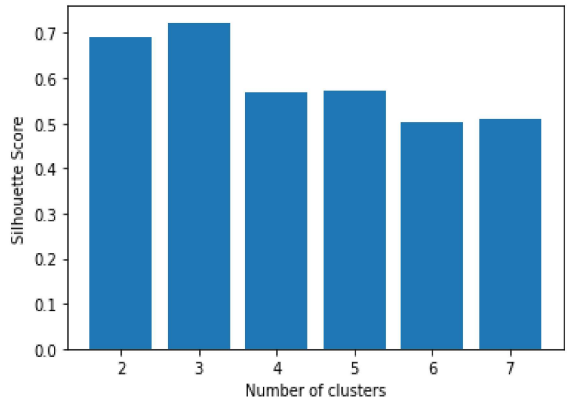


The line graph created by the Elbow Method assists in identifying the optimum number of clusters. At the points on the graph, where the data line abruptly changes direction is focused upon. These points are called the 'elbow points'. As seen in the figure, two elbow points are present at 'no. of clusters = 2' and 'no. of clusters' = 3.

The WCSS i.e., Within Cluster Sum of Square, is the sum of squared distances between each data point from the dataset and its respective cluster's centroid. As a general rule, the point most parallel or most closes to the X axis, indicates the optimal number of clusters.

Silhouette approach.

While working with K-means, another method to find the optimum number of clusters is the Silhouette method. This method identifies how a data point is similar to its own cluster. Python's matplotlib library is used as a data visualization tool.



The Silhouette bar graph represents the same findings as the Elbow approach, where there is a close correlation between n_clusters 2 and 3. However, the Silhouette bar plot clearly emphasizes that the optimal number of clusters should be 3.

K-means.

The Elbow and Silhouette methods indicated that the optimal number of clusters is 3. The Weka software is used for computation and clustering.

Following were the results from Weka's clustering functionality, when all the features were considered.

```
Clustered Instances

0       8067 ( 45%)
1       2131 ( 12%)
2       7700 ( 43%)
```

```
    0    1    2  <-- assigned to cluster
 7740  912 7607 | 0
  327 1219   93 | 1
```

The instances which were incorrectly clustered were 8939 i.e., 49.94%. Hence the correct clustering accuracy when all the attributes are considered is 50.055%.

```
Incorrectly clustered instances :       8939.0   49.9441 %
```

The next clustering task is to consider only the first four features i.e., ‘Mean of the integrated profile’, ‘Standard deviation of the integrated profile’, ‘Excess kurtosis of the integrated profile’, ‘Skewness of the integrated profile’.

```
Instances:    17898
Attributes:   9
              Profile_mean
              Profile_stdev
              Profile_skewness
              Profile_kurtosis
Ignored:
              DM_mean
              DM_stdev
              DM_skewness
              DM_kurtosis
              class
```

```
Clustered Instances                0     1     2  <-- assigned to cluster
                               10670    11  5578 | 0
0       10751 ( 60%)              81   903   655 | 1
1         914 (  5%)
2        6233 ( 35%)
```

Out of the total 17,898 instances, the instances which were incorrectly clustered were 6325 i.e., 35.33%.

```
Incorrectly clustered instances :       6325.0   35.3391 %
```

The instances which were correctly clustered were 11,573. Hence the clustering accuracy when the first four attributes are considered is 64.67%

Further, the next clustering task is to consider only the last four features i.e., 'Mean of the DM-SNR curve', 'Standard deviation of the DM-SNR curve', 'Excess kurtosis of the DM-SNR curve' and 'Skewness of the DM-SNR curve'.

```
Instances:    17898
Attributes:   9
              DM_mean
              DM_stdev
              DM_skewness
              DM_kurtosis
Ignored:
              Profile_mean
              Profile_stdev
              Profile_skewness
              Profile_kurtosis
              class
```

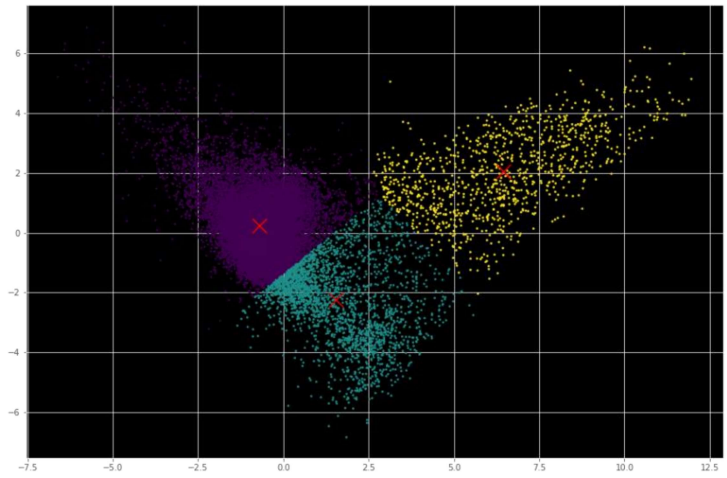
```
Clustered Instances                    0    1    2  <-- assigned to cluster
                                    5520 1176 9563 | 0
0       5570 ( 31%)                   50 1042  547 | 1
1       2218 ( 12%)
2      10110 ( 56%)
```

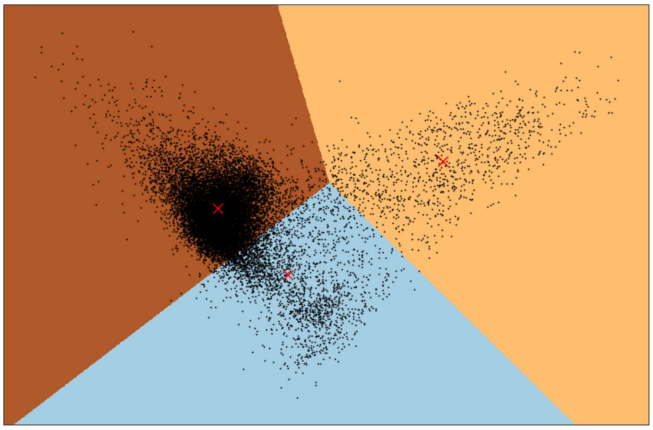
Out of the total 17,898 instances, the instances which were incorrectly clustered were 7293 i.e., 40.74%.

```
Incorrectly clustered instances :       7293.0    40.7476 %
```

The instances which were correctly classified were 10,605. Hence, the clustering accuracy when the last four attributes are considered is 59.25%

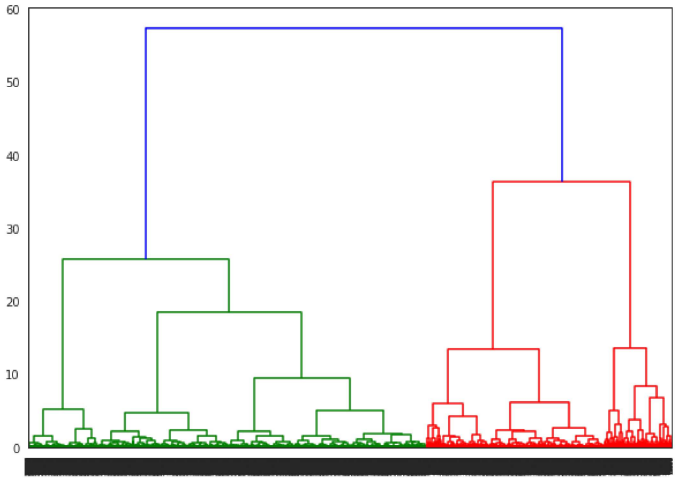
The K-means cluster scatterplot is displayed below. The cluster centroids of the respective clusters are marked with a red cross.



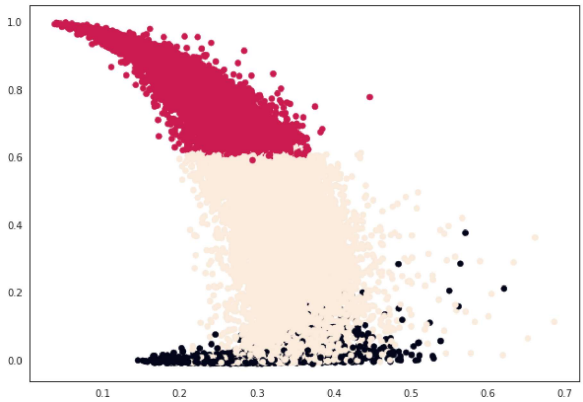


Hierarchical Clustering.

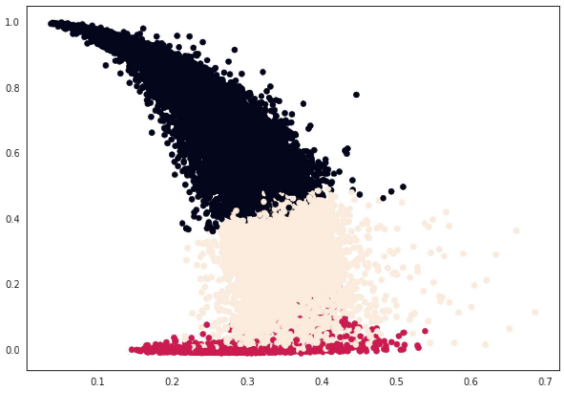
In the beginning, all the features are considered for the clustering task. The dendrogram and all the plots with all different linkages are depicted below. Also displayed is the clustering task while considering only the first four task and while considering the last four task.



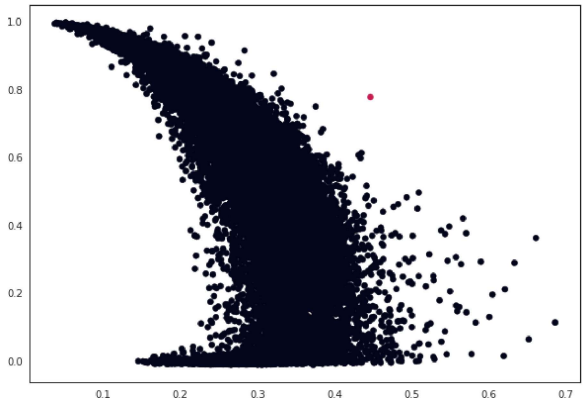
Complete Linkage.



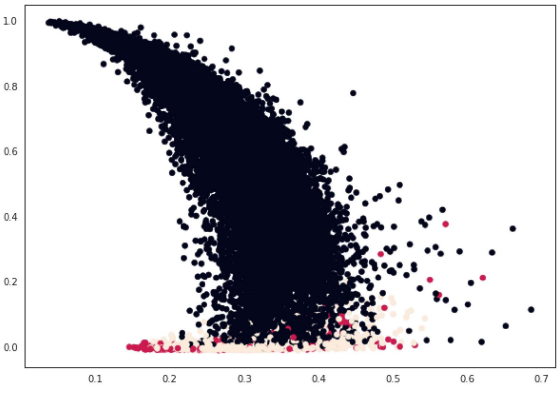
Ward Linkage.



Single Linkage.



Average Linkage.



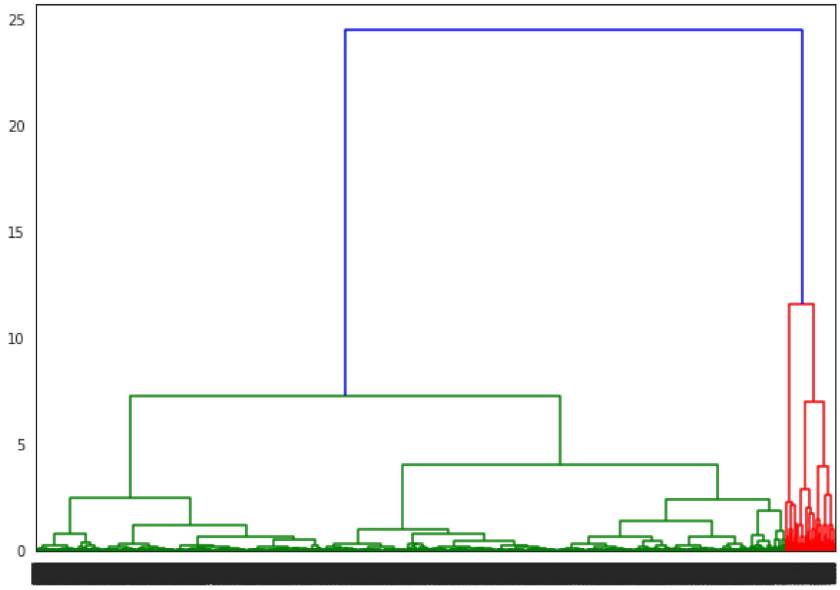
First four features

The next hierarchical clustering task is to consider only the first four features i.e., ‘Mean of the integrated profile’, ‘Standard deviation of the integrated profile’, ‘Excess kurtosis of the integrated profile’, ‘Skewness of the integrated profile’.

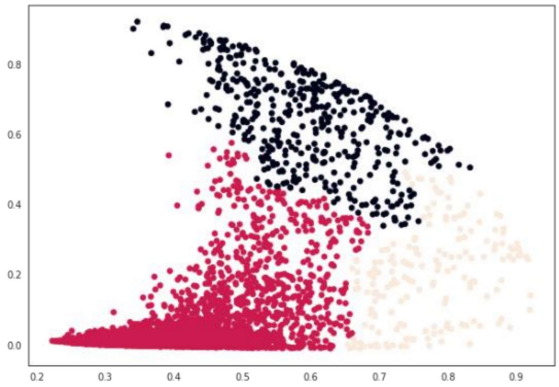
	Mean of the integrated profile	Standard Deviation of the integrated profile	Excess kurtosis of the integrated profile	Skewness of the integrated profile	Class
0	140.562500	55.683782	-0.234571	-0.699648	0
1	102.507812	58.882430	0.465318	-0.515088	0
2	103.015625	39.341649	0.323328	1.051164	0
3	136.750000	57.178449	-0.068415	-0.636238	0
4	88.726562	40.672225	0.600866	1.123492	0
...	...	...	...	...	...
17893	136.429688	59.847421	-0.187846	-0.738123	0
17894	122.554688	49.485605	0.127978	0.323061	0
17895	119.335938	59.935939	0.159363	-0.743025	0
17896	114.507812	53.902400	0.201161	-0.024789	0
17897	57.062500	85.797340	1.406391	0.089520	0

17898 rows × 5 columns

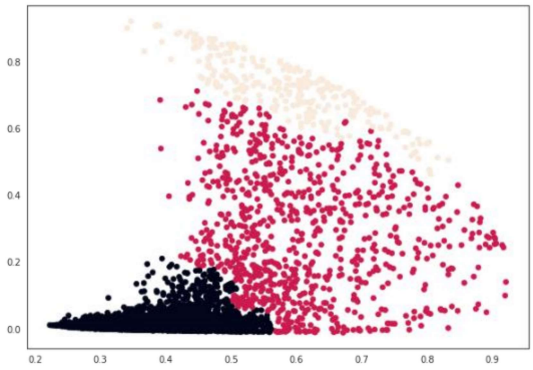
Dendrogram for first four features.



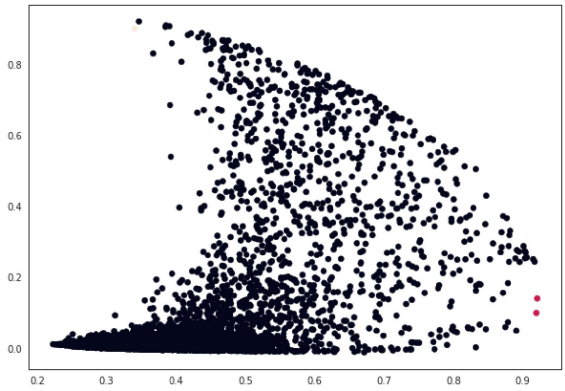
Complete Linkage.



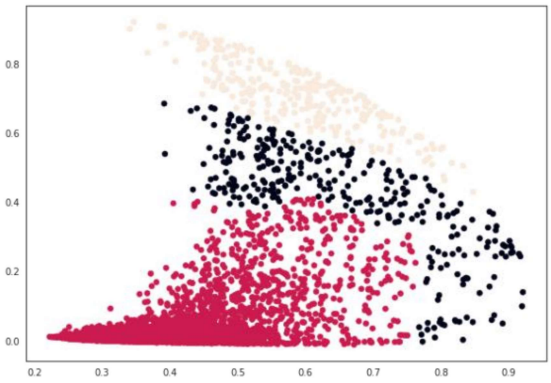
Ward Linkage



Single Linkage.



Average Linkage.

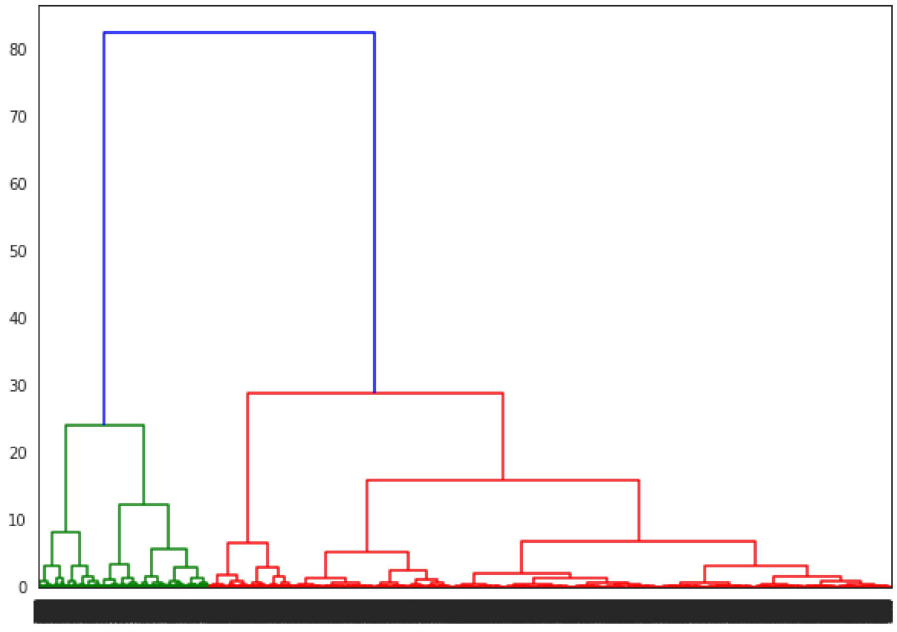


Further, the next clustering task is to consider only the last four features i.e., ‘Mean of the DM-SNR curve’, ‘Standard deviation of the DM-SNR curve’, ‘Excess kurtosis of the DM-SNR curve’ and ‘Skewness of the DM-SNR curve’.

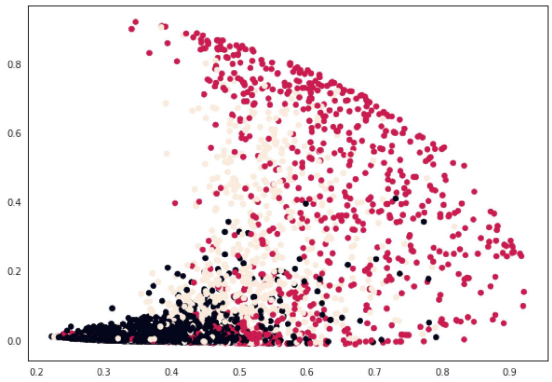
	Mean of the DM-SNR curve	Standard Deviation of the DM-SNR curve	Excess kurtosis of the DM-SNR curve	Skewness of the DM-SNR curve	Class
0	3.199833	19.110426	7.975532	74.242225	0
1	1.677258	14.860146	10.576487	127.393580	0
2	3.121237	21.744669	7.735822	63.171909	0
3	3.642977	20.959280	6.896499	53.593661	0
4	1.178930	11.468720	14.269573	252.567306	0
...	...	...	...	...	...
17893	1.296823	12.166062	15.450260	285.931022	0
17894	16.409699	44.626893	2.945244	8.297092	0
17895	21.430602	58.872000	2.499517	4.595173	0
17896	1.946488	13.381731	10.007967	134.238910	0
17897	188.306020	64.712562	-1.597527	1.429475	0

17898 rows × 5 columns

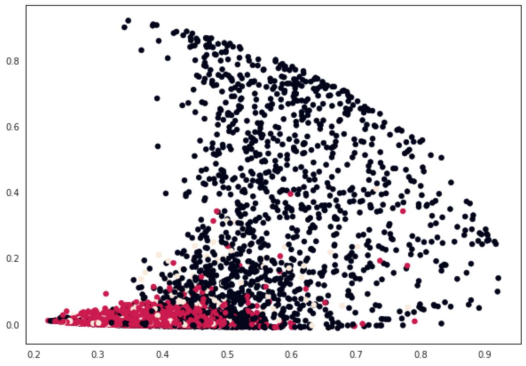
Dendrogram of the last four features.



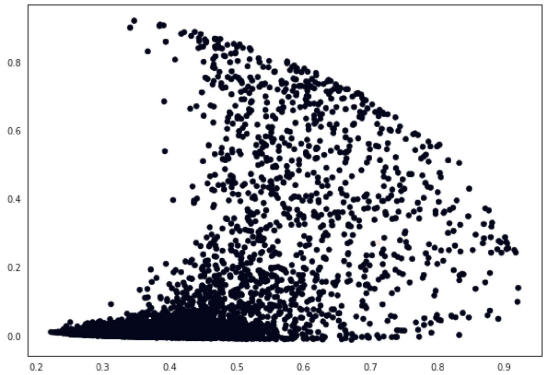
Complete Linkage.



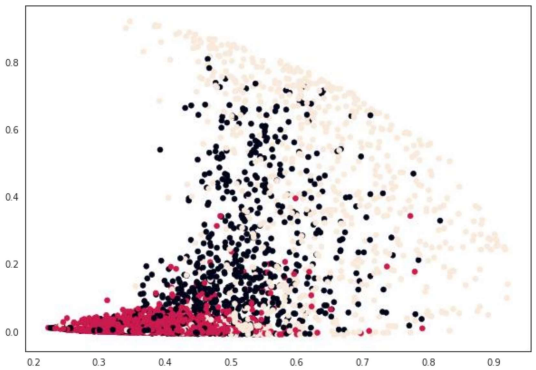
Ward Linkage.



Single Linkage



Average Linkage.



Project Analysis.

To find the optimal number of clusters, two approaches namely; The Elbow Method and the Silhouette Method was used. Both the approaches indicated that the optimal number of clusters was 3. The K-Means clustering was carried out in three ways. The first iteration was the clustering task when all the features were considered. The instances which were correctly clustered were 8,959, hence the clustering accuracy was 50.0558%. The second iteration was the clustering task when only the first features were considered i.e., ‘Mean of the integrated profile’, ‘Standard deviation of the integrated profile’, ‘Excess kurtosis of the integrated profile’,

‘Skewness of the integrated profile’. The incorrectly clustered instances were 6,325. The instances which were correctly clustered were 11,573. Hence the clustering accuracy when the first four attributes are considered is 64.67%. In the third iteration, the last four features were considered i.e., ‘Mean of the DM-SNR curve’, ‘Standard deviation of the DM-SNR curve’, ‘Excess kurtosis of the DM-SNR curve’ and ‘Skewness of the DM-SNR curve’.

Out of the total 17,898 instances, the instances which were correctly clustered were 10,605. Hence, the clustering accuracy when the last four attributes are considered is 59.25%. The highest clustering accuracy was achieved when only the first four features of the HTRU 2 dataset were considered. And the lowest clustering accuracy was observed when all the features from the HTRU 2 dataset were considered.

In hierarchical clustering, the linkages namely; ward, single, complete and average, were implemented. The optimal number of clusters were identified as 3. In the first iteration, where all the features were considered, the ‘ward’ and ‘complete’ linkages were the best suited, as they depicted the 3 clusters more accurately. In the second iteration, where the first four features were to be considered, again the ‘’ and ‘’ linkages were the best suited. In the third iteration, where the last four features were to be considered, the ‘ward’ and ‘average’ linkages, although no proper clustering was observed, were the perfectly suitable ones.

References.

[1] Shrimali, Shubham & Pandey, Amritanshu & Chowdhary, Chiranji. (2020). K-means Clustering-based Radio Neutron Star Pulsar Emission Mechanism. Recent Advances in Computer Science and Communications. 13. 10.2174/2213275912666200129115401.

[2] D. R. Lorimer and M. Kramer, "Handbook of Pulsar Astronomy", Cambridge University Press, 2005.

[3] Jain, S., Aalam, M.A., & Doja, M.N. (2010). K-MEANS CLUSTERING USING WEKA INTERFACE 1.