NEW YORK INSTITUTE OF TECHNOLOGY

New York Institute of Technology

Machine Learning Project Assignment 2 - Clustering Report

DTSC 710/ Fall 2021 - M01

Course Instructor: Dr. Kiran Balagani

Title: Classification Project

Hui (Henry) Chen

Table of Contents

Dataset

The data set comes from the CMU data repository (https://archive.ics.uci.edu/ml/datasets/HTRU2). In total, there are 9 features, 17898 observations, and 0% missing data. The data set described the sample of pulsar collected from the *High Time Resolution Universe Survey*. The first four features described the mean, standard deviation, excess kurtosis, skewness of the integrated profile, while the last four features described the DM-SNR curve. The last feature is the class label of the data set.

Data Exploration

Since the goal of this project is to perform clustering on the dataset, it is necessary to explore the class weight of the labelled observations. After some analysis through programming with Python, we discovered that the dataset is imbalanced as the below figure demonstrates.

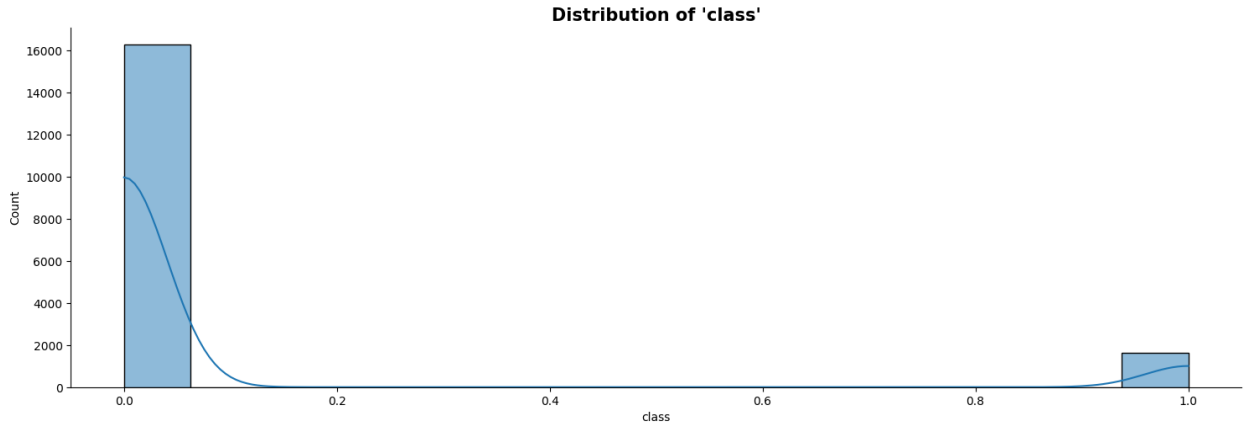

Fig 1: Distribution of the class label

As we can see, there are more observations belonging to classes 0 and around 2k observations belonging to classes 1.

Optimal *K* value and Clustering Level

Since the dataset does not have any missing data, we don't have to deal with it. However, in the clustering project, we do need to find the optimal value of *K* for the K-Means model. In this project, we utilized the Elbow Method to find the optimal number of K-means clusters with respect to the first four, last four and all features. The metric used in the Elbow method is "distortion" where computes the sum of squared distances from each point to its assigned centre. The Elbow Method result is demonstrated as follows:

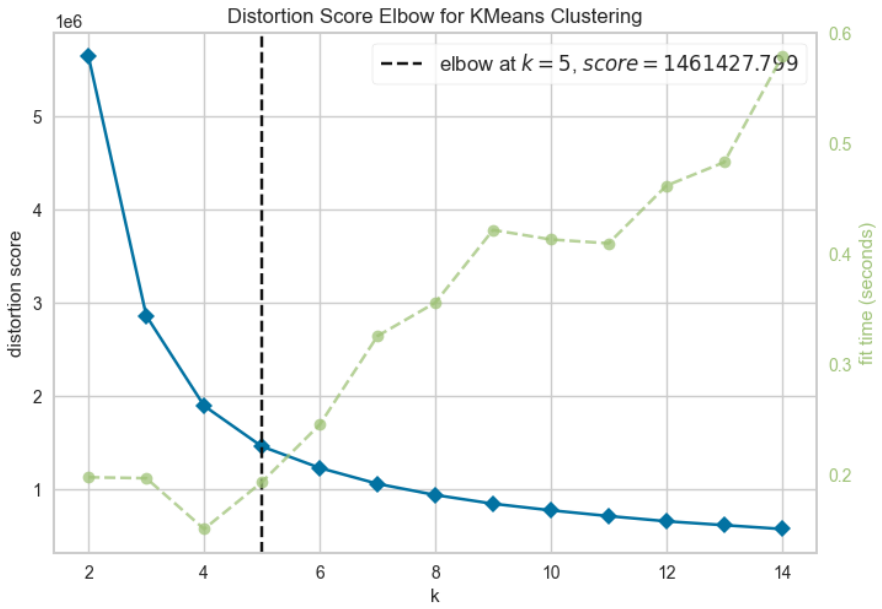

Fig 2: the distortion score of the Elbow Method with the first four features.

The blue continuous line is the distortion score of the first four features on the K-Means Clustering while the green dashed line represents the fit time of the data set to the model in second. As we can see, the optimal value of *K* would be 5 with a distortion score of 1461427.99. With the same approach, we perform this process again on the last four features.

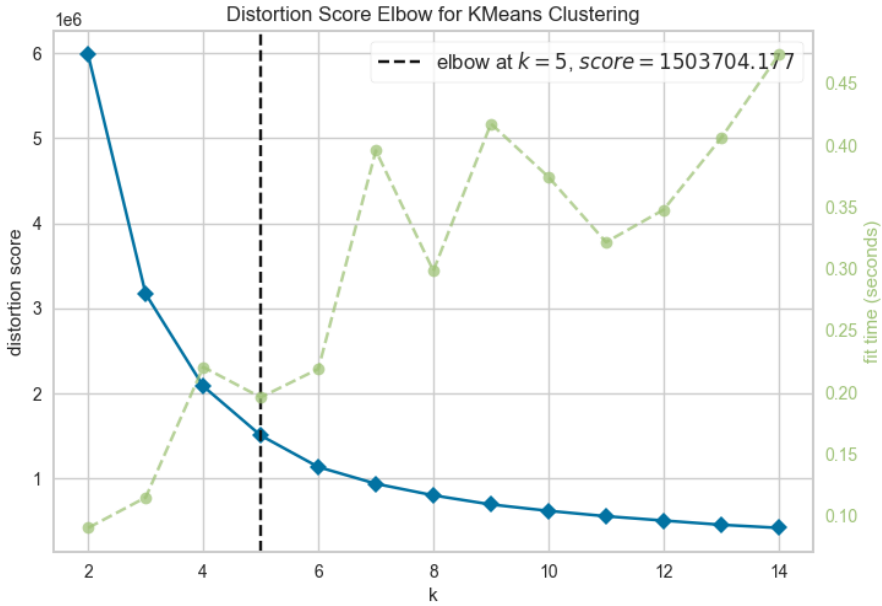

Fig 3: the distortion score of the Elbow Method with the last four features.

After applying the Elbow Method to the last four features, the optimal value for K is 5 with a distortion score of 1503704.177.

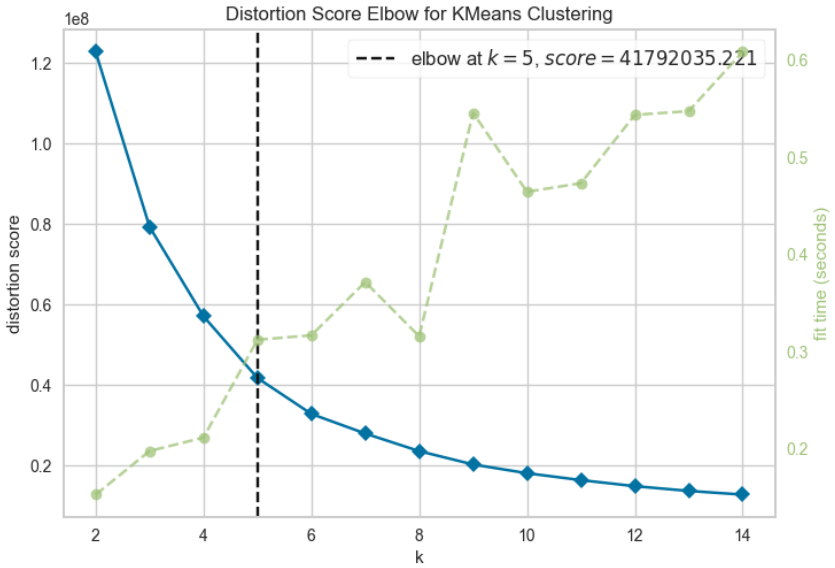

Fig 4: the distortion score of the Elbow Method for all features.

The same process applies to all features and the result shows the optimal value for *K* is 5 with a distortion score of 41792035.221. The distortion score here is significantly higher compared to the first and last four features, which demonstrates the K-Means model would not work well when we include all features. Note, for all Elbow Method analyses, we disregard the "class" label. The only time we use the class label is when comparing the clustering result since the "class" feature is the ground truth of the observation.

Another clustering method is Hierarchical Clustering and we need to find the optimal clustering level for such clusters. In this project, Dendrograms is utilized to find the optimal clustering.

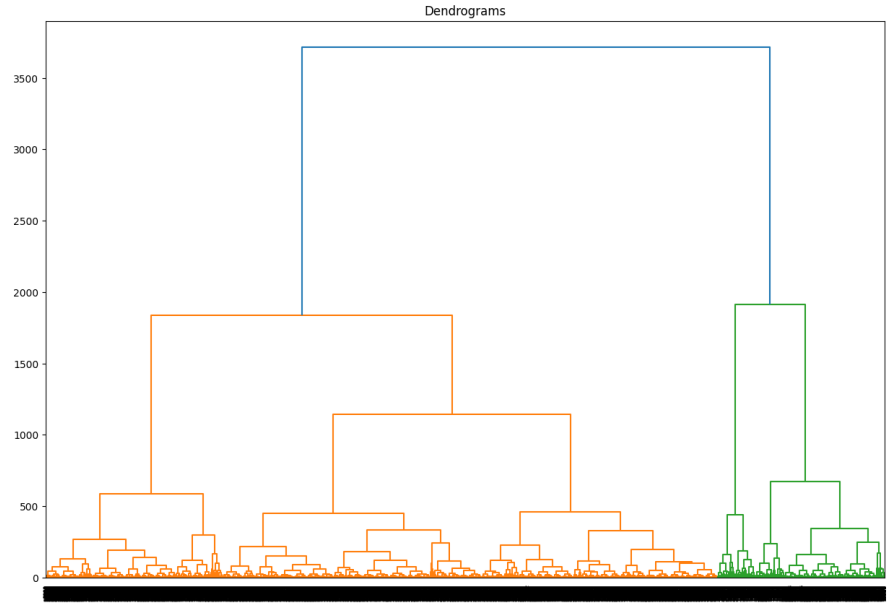

Fig 5: Dendrograms for the first four features

According to the dendrogram, the optimal clustering level for the first four features is 4 as the longest vertical distance/line did not intersect with other splits/horizontal lines. The same method applies to the last four features and the result is shown below:

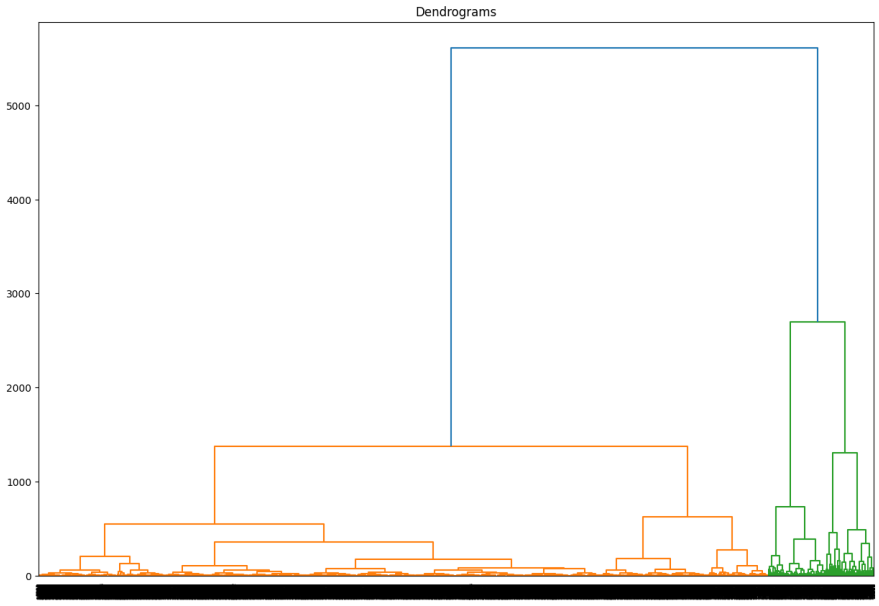

Fig 6: Dendrograms for the last four features

As we can see, the optimal clustering level for the last four features is 3 since the blue vertical line did not intersect with the other splits/ horizontal lines. Then the same method was applied to all features and the result is shown below:

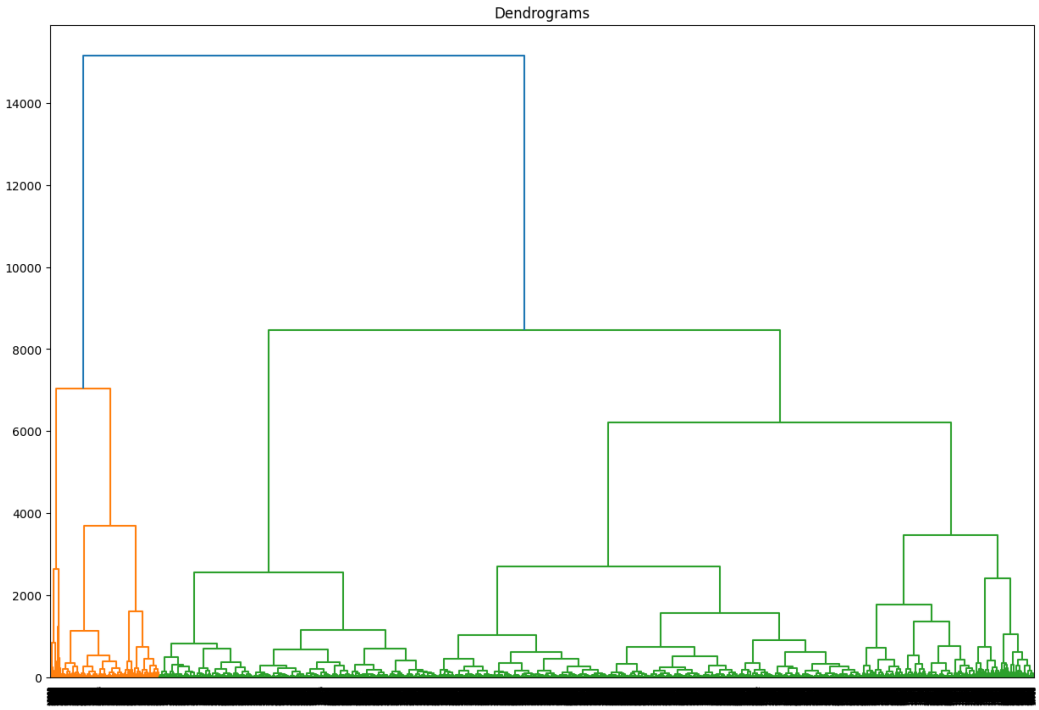

Fig 7: Dendrograms for all features

Based on Fig 7, the optimal clustering level is 3 as the longest distance (the blue line) did not intersect splits/ vertical lines. Note, the same set of data we used in the Elbow Method is also applied here.

Model Analysis 1: K-Means Clustering

After training the K-Means model with the respect *K* value that we found, we got the following results:

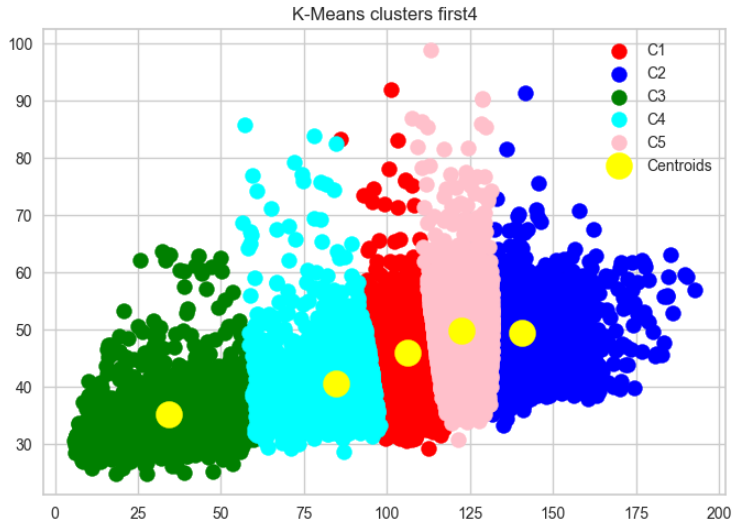

Fig 8: Visualization of the K-Means model for the first four features

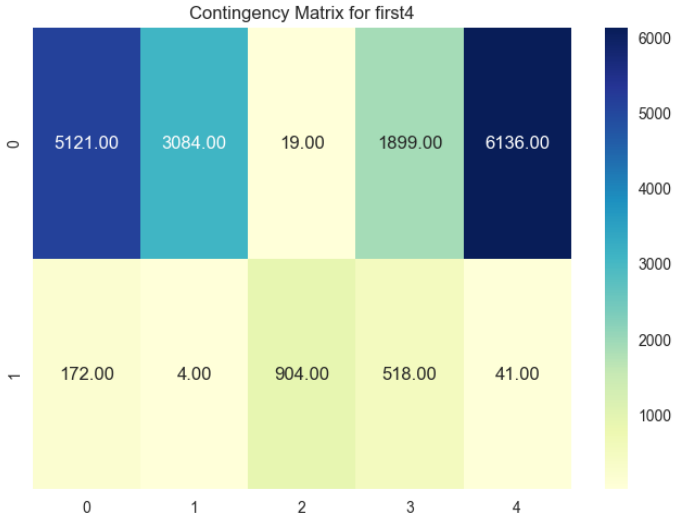

Fig 9: Contingency Matrix of the K-Means model for the first four features

As we can see from Fig 8 and 9, with $K = 5$, the model was able to classify the first four features into 5 clusters well. According to Fig 9, most of the observations were clustered on 5 with class 0 while the least observations were clustered on 3 with class 1.

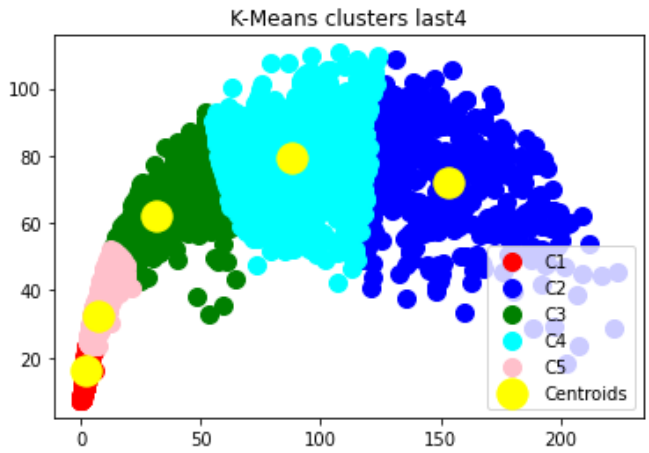

Fig 10: Visualization of the K-Means model for the last four features

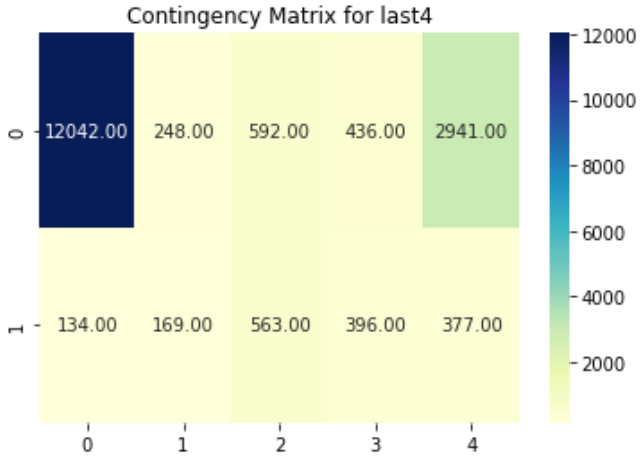

Fig 11: Contingency Matrix of the K-Means model for the last four features

Based on Fig 10 and 11, with $K = 5$, the estimator was able to cluster the last four features well. However, the Contingency Matrix demonstrated that most of the observations were in cluster 0 with class label 0.

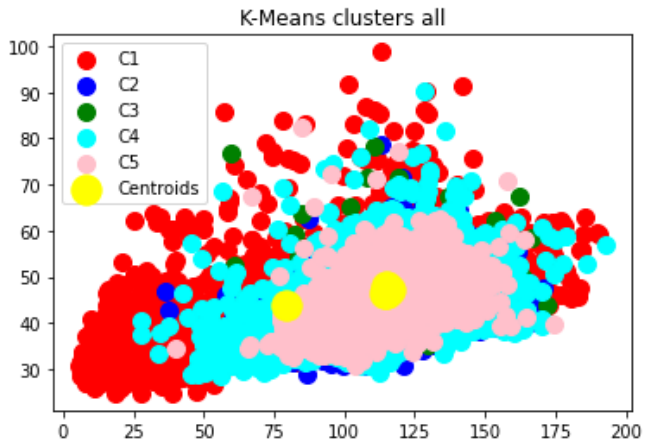

Fig 12: Visualization of the K-Means model for all features

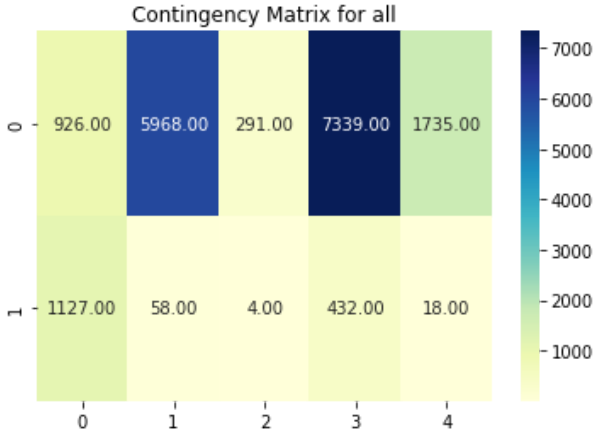

Fig 13: Visualization of the K-Means model for all features

Based on Fig 12, the value of $K = 5$ did not fit the data. The two centroids suggest the data set would fit the data better if the K is 2. Therefore, this suggests that during the Elbow Method, the analysis is incorrect. Also, from the Contingency Matrix, most of the observations were clustering into clusters 4 and 2 with class label 0. The least cluster is 3 with class label 1. Compare to the first and last features, all features tend to cluster the observations into clusters 2 and 4. Overall, class label 1 is less clustered compared to label 0 due to the imbalance dataset.

Model Analysis 2: Hierarchical Clustering

After training the data through Hierarchical Clustering, we got the following results:

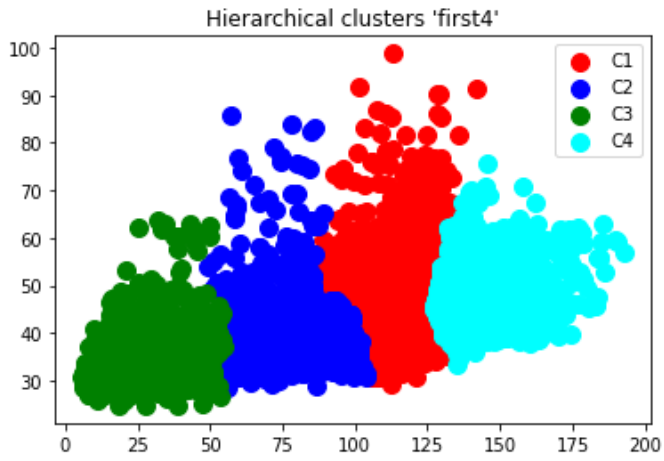

Fig 14: Visualization of the Hierarchical Clustering model for the first four features

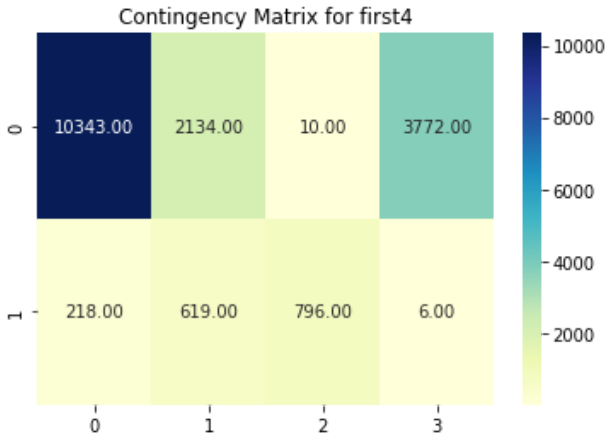

Fig 15: Contingency Matrix of the Hierarchical Clustering model for the first four features

Based on the Contingency Matrix, most of the observations with class label 0 were clustered in cluster 1, whereas cluster 4 with class label 1 has the least number of observations in the cluster.

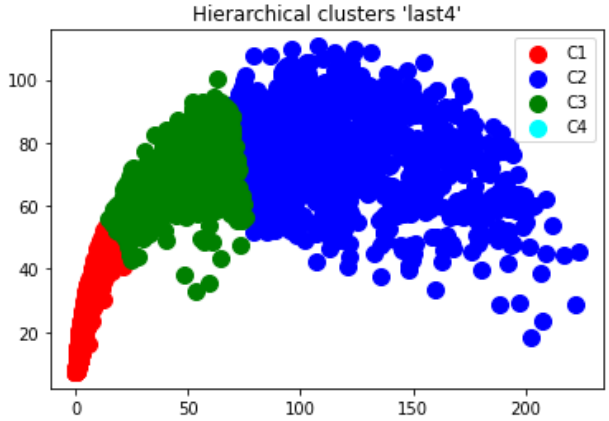

Fig 16: Visualization of the Hierarchical Clustering model for the last four features

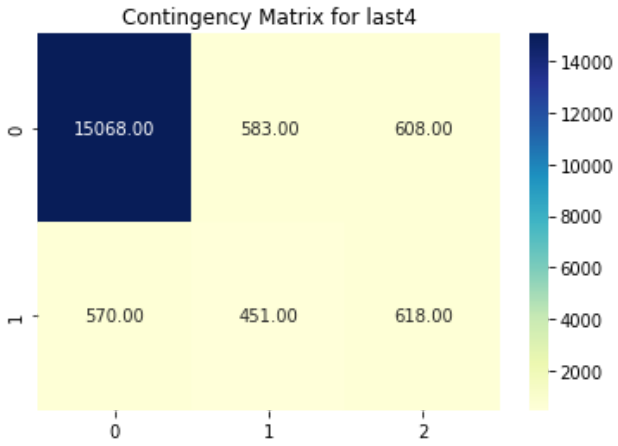

Fig 17: Contingency Matrix of the Hierarchical Clustering model for the last four features

Based on Fig 16 and 17, the cluster level of 3 works well. However, the Contingency Matrix demonstrates many observations with class label 0 were clustered in cluster 1.

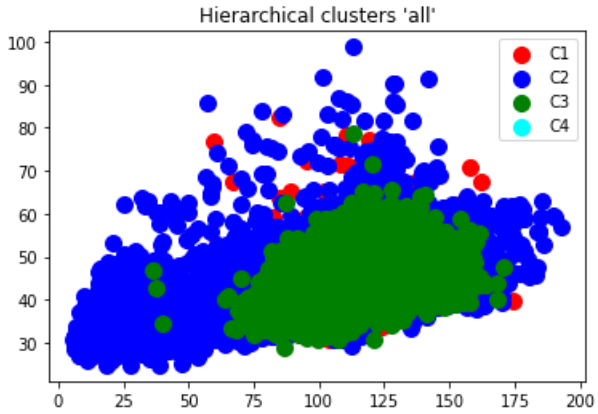

Fig 18: Visualization of the Hierarchical Clustering model for all features

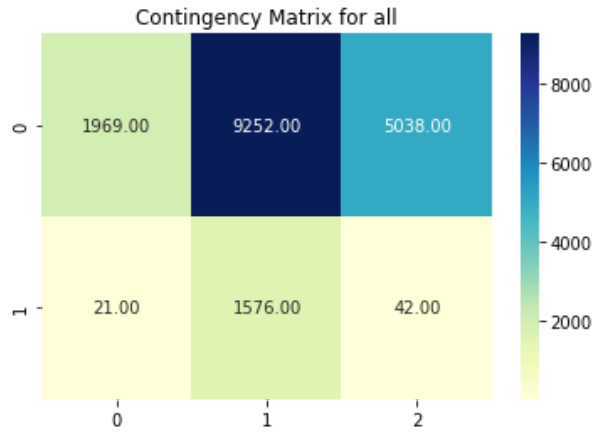

Fig 19: Contingency Matrix of the Hierarchical Clustering model for all features

With a clustering level of 3, the data visualization demonstrated the estimator did not describe the data well as many clusters are overlapped. However, compared to the first and last feature data set clustering, using all features tends to cluster into cluster 2 with class label 0.

Closing Remarks

Clustering is an unsupervised learning technique in Machine Learning. Throughout the project, we utilized K-Means and Hierarchical Clustering and then evaluated the two estimators with the metrics from the Contingency Matrix. However, there are many other metrics that could be evaluated. In the future, this is something that I would like to explore further for this project.

Source Code: https://github.com/hchen98/DTSC710-ML/tree/main/Project%20Assignment%202