IDM CHALLENGE 3

Younas Mahmood 22854

**Business Understanding: -**

This challenge is related to the symptoms of diseases that the patients experience, using clusters we are identifying the patterns of symptoms each cluster (disease) manifests. We are pre-processing and using various kinds of models to show that the model predicts results consistent with the actual results, this would be useful for doctors in the medical domain.

**Data Understanding: -**

This dataset consists of 2783 rows and 132 columns, each row represents a patient suffering from a disease, each column represents a symptom the patient has such as ‘Skin rash’, ‘Shivering’, ‘Joint Pain’, etc. Each element in a column contains either a ‘0’ value or a ‘1’ value, 0 showing the patient does not have the symptom, and 1 showing the patient has the symptom.

**Data Preparation: -**

Since the data was in binary values, there was no need to use dummy encoding or one hot encoding, normalization, and removing outliers even though they improve K-Means. There were no missing values in data hence there was no need to impute it with the most common value of the column. I also dropped columns that were multicollinear to see if it improves the score.

**Modelling: -**

I used K-Means Clustering, Agglomerative Clustering, and Hierarchical Clustering with various kinds of parameters. Furthermore, the usage of Dendrogram, Silhouette curve, and Elbow curve gave me better understanding of what number of clusters to pick. Therefore, on most entries number of clusters I used were 6 or 8. Using different random states also improved/decreased the score.

|  |  |  |  |
| --- | --- | --- | --- |
| **Entry** | Data Pre-processing | Model Configurations | **Score %** |
| 1 | None | Used default K-Means | 0.61441 |
| 2 | None | Used K-Means with clusters = 4 | 0.66589 |
| 3 | None | Used K-Means with clusters = 15 | 0.59702 |
| 4 | None | Used K-Means with clusters = 17 | 0.44386 |
| 5 | None | Used K-Means with clusters = 14, 100 max iterations | 0.48598 |
| 6 | None | Used K-Means with clusters = 9, 100 max iterations | 0.6242 |
| 7 | None | Used K-Means with clusters = 8, 130 max iterations | 0.69158 |
| 8 | None | Used K-Means with clusters = 7, 200 max iterations | 0.52381 |
| 9 | None | Used Agglomerative Clustering with clusters = 4 and linkage = Ward | 0.31867 |
| 10 | None | Used Agglomerative Clustering with clusters = 6 and linkage = Ward | 0.58181 |
| 11 | Dropped 5 columns that influenced multicollinearity > 0.90 | Used K-Means with clusters = 12 | 0.57937 |
| **12** | **None** | **Used K-Means with clusters = 6, Random State = 0** | **0.80009** |
| 13 | None | Used K-Means with clusters = 8, Random State = 0 | 0.44585 |
| 14 | None | Used K-Means with clusters = 8, did not note down its random state | 0.6518 |
| 15 | Dropped 7 columns that influenced multicollinearity > 0.80 | Used K-Means with clusters = 12, dropping more columns did not increase the score | 0.51232 |
| 16 | None | Used K-Means with clusters = 5, Random State = 0 | 0.45279 |
| 17 | None | Used K-Means with clusters = 6, Random State = 0, Fitted model 10 times | 0.80009 |
| 18 | None | Used hierarchical clustering in Knime | 0.4551 |
| 19 | None | Accidentally uploaded last entry | 0.4551 |
| 20 | None | Accidentally uploaded last entry | 0.4551 |
| 21 | None | Used Agglomerative Clustering, linkage = single, number of clusters = 6 as elbow graph showed one elbow point at n cluster = 6 | 0.48605 |
| 22 | None | Used Agglomerative Clustering, linkage = single, number of clusters = 13 | 0.35848 |
| 23 | None | Used K-Means with clusters = 6, n\_init = 100 | 0.34242 |
| 24 | None | Used Fuzzy C Means with clusters = 6, max iter = 60 | 0.48602 |
| 25 | None | Used Fuzzy C Means with clusters = 6, max iter = 130 | 0.52847 |
| 26 | None | Used Fuzzy C Means with clusters = 6, max iter = 150 | 0.48512 |
| 27 | None | Used K-Means with clusters = 6, n\_init = 100, Last 5 submissions showed, a good random state was especially important for a good score | 0.73779 |
| 28 | None | Used Agglomerative Clustering, linkage = average, number of clusters = 15 | 0.68636 |
| 29 | Dropped 5 columns with multicollinear > 0.90 | Used Agglomerative Clustering, linkage = ward, clusters = 20 | 0.44199 |
| 30 | None | Used K-Means with clusters = 6, algorithm = “full” | 0.68636 |
| 31 | None | Did not record parameters of Hierarchical clustering | 0.45438 |
| 32 |  | Error in writing in csv file | 0 |
| 33 | None | K-Means with clusters = 5 | 0.65729 |
| 34 | None | Repeated entry 17 without random state to see how much it effects score. | 0.58478 |
| 35 | None | K-Means with clusters = 6, algorithm = “Lloyd” | 0.49609 |
| 36 | None | K-Means with clusters = 6, algorithm = “Elkan” | 0.61444 |
| 37 | None | Used K-Means after identifying clusters = 8 used silhouette plot to set number of clusters | 0.54639 |
| 38 | None | Used K-Means after identifying clusters = 17 using dendrogram to set number of clusters | 0.52663 |
| 39 | None | Used Agglomerative clustering, linkage = single, clusters = 8 using silhouette plot to set number of clusters | 0.45738 |
| 40 | None | Used Agglomerative clustering, | 0.64169 |
| 41 | None | Did not record parameters of K-Mean | 0.49554 |
| 42 | None | Did not record parameters of K-Mean | 0.15725 |
| 43 | None | Repeated previous file | 0.15725 |
| 44 | None | Kmeans clusters = 7, random state = 0 | 0.49483 |
| 45 | None | Kmeans clusters = 8, random state = 0 | 0.51232 |
| 46 | None | Kmeans clusters = 9, random state = 0 | 0.49442 |
| 47 | None | agglomerative, single linkage, clusters = 14 | 0.64879 |
| 48 | None | agglomerative, single linkage, clusters = 12 | 0.55748 |
| 49 | None | agglomerative, single linkage, clusters = 11 | 0.51168 |
| 50 | None | agglomerative, single linkage, clusters = 13 | 0.5614 |

**Evaluation: -**

1. I think K-Means, worked the best for this dataset with the Elbow point being 6 number of clusters. Pre-processing of data did not improve score, which is why in most entries, I did not apply them again. This algorithm was the best because K-Means computed centroids of the dataset and using many iterations it found the best centroid for the clusters.
2. The optimal number of clusters for the dataset in my understanding was 6 and 8. I checked the elbow graph to see that there was a slight elbow point at cluster = 6. I used the silhouette technique for each cluster size which resulted in 8 being the best. The dendrogram did not provide a suitable number of cluster as it did not result in a good score.
3. I could not figure out what random state was the best for the K-Mean model, it was also a problem to understand which cluster was the best for each model. The models were also not consistent, I applied the same parameters except for the random state, and they provided an enormous difference in score. Also using the methods to provide the optimal number of clusters were very distinct.