Clustering
K-means

Applying two clustering methods in our dataset

Team SheCodes

Monica Dommaraju
Asha Balshiram Aher
Swati Ganesh Narkhede
Sri Sruthi Chilukuri

1. Our main objective of performing clustering on the dataset, is to group employees with similar professional and personal characteristics. Then for each cluster, the percentage of employee attrition can be calculated from the class label. This helps the HR department to focus on employee groups with the highest attrition rate and take necessary actions (may be by adjusting salaries/ providing additional benefits/perks etc) to retain the employees.

2. We were able to deduce the following results from our clustering results.

   ● Our data was divided into 3 clusters with sizes of (645, 622 and 203) employees respectively.

   ● Employees belonging to the 3rd group earns high monthly salary and most of them are older than employees in other groups

   ● From correlation matrix, we can deduce that higher the performance rating, higher the percentage of salary hike.

   ● Higher the age, higher the monthly income

   ● Higher the number of working years, higher the monthly income.

   ● We have also found that the attrition rate is only 7% for the 3rd cluster, whereas it is close to 18% for the other two groups. So, the HR department can most focus on employees of cluster 1 and 2 to reduce the attrition rate.

   ● Also, Job Role and Manager Position had higher correlation which means employees which managerial roles are more likely to stay back in the company compared to others.

   ● Also, YearsSinceLastPromotion and TotalWorkingYears are highly correlated which mean that higher the employee's tenure is in the company; more likely he is to retain in the same; given the rest of the factors like salary, work-life balance, etc don't change.

3. If we have to say whether our data has features that support our hypothesis ; Yes, we have features such as Monthly Income, Performance Rating, Total Working years, Years at Company, Gender, etc which were used for clustering. The clustering results were then used along with the feature named Attrition (which is our class label) to find the attrition percentage per cluster.

4. Comparisons:

   We have performed K-Means clustering, Agglomerative, DBSCAN, Ward and Spectral Clustering.

● **K-Means and DBSCAN Comparison:** K-Means performed better in clustering into groups. With DBSCAN, the main advantage is that we don't need to provide the number of clusters before hand and it can also be used to detect noise. But for our dataset it did not go well and we got all the data points into a single cluster.

● **K-Means and Ward clustering comparison:** K-Means can be a beneficial method when we are sure of how many clusters we would want from our dataset. However; in Ward Clustering, which is a hierarchical technique, focuses on computing the least squared distance between clusters or data points. Hence; it strives to lessen the cost of merging which can be an advantage computationally.

Although Kmeans performed better for our dataset; the hierarchical clustering is a more prefered way of cluster analysis because of the underlying fact that there are less data assumptions(like value of 'k') and when the data is comparable.

Note: According to our observation; among single link, complete link, average link and ward clustering techniques, complete link could perform better on our dataset.

● **K-Means and Spectral Clustering Comparison:** K-Means and Spectral Clustering requires number of clusters as parameters. For performing K-Means and Spectral Clustering, we considered total 13 features and applied the PCA algorithm for dimensionality reduction. We performed Spectral clustering and K-Means clustering on the same PCA Components generated by the PCA algorithm. Out of 3 clusters, data points are changed in cluster 1 and cluster 2. In spectral clustering, cluster 1 has less data points than cluster 2. Whereas,In K-Means, cluster 2 has less data points as compared to cluster 1. Cluster 3 remains same in both clustering algorithms.

● **K-Means and Agglomerative Clustering comparison :** Agglomerative clustering finds points that are close to each other and then group them together. This clustering is hierarchical because it performs operations sequentially. This algorithm comes to rescue where you want to make a decision about how roughly or finely you want to group your data.

5. Yes, we have inspected the outliers after the KMean clusters were formed. We have done the following steps to detect outliers

a. Calculated the distance from cluster centroids to each of the records that belongs to the cluster b. Found the mean distance for each of the clusters. c. If the distance between the cluster centroid and the record is greater than 2 times the mean, then that record is considered as the outlier for that cluster. d. For our implementation, we found that there are 6 outliers in Cluster 1 (6 / 645),

15 outliers in cluster 2 (15 / 622) and 3 outliers in cluster 3 (3 /203)

5.a) Individual data points that are nearer to the centroid are the ones which have the highest similarity among each other and are hence clustered around the centroid; similarly the data points that are farther from a cluster are the ones that have high-dissimilarity with the features belonging to the same cluster and these are the ones that are often considered as outliers.

Also, from the cluster analysis, we have not found any outliers but we assume that with a deeper exploratory data analysis studying each individual parameter and applying other evaluation metrics; we would encounter outliers in our dataset.

We have mentioned the algorithms each of us from our team have implemented along with the Google Colab Links:

- Monica Dommaraju- KMeans and DBSCAN
https://colab.research.google.com/github/monicasjsu/ML_Clustering_Assignment/blob/master/clustering_Monica.ipynb

- Asha Aher- Means and Agglomerative Clustering

  https://colab.research.google.com/drive/128LNGwL16eTV1NUPN7igYJZL6xPb6 CE5#scrollTo=uEp7gX8sAAQF

- Swati Ganesh Narkhede- KMeans and Spatial Clustering
https://colab.research.google.com/drive/17TWbCoGfOJcu1UxJ9-vXbA_FcqVR0WIK

- Sri Sruthi Chilukuri - Kmeans and Ward Clustering
https://colab.research.google.com/drive/1qPBeCiEQ-JzS_mSs76cR1MgwVAHLmd8E