Identifying Optimal Subgroups of Traumatic Brain Injury Patients using K-means Clustering

Group Members

Dacosta Yeboah

Ayman Haque

Shashi Khanal

**Abstract**

Traumatic brain injury (TBI), also known as concussion can result in speech, language and other psychological problems. Identifying subgroups of patients allows us to determine which clinical features contribute to good or bad patient outcomes. Clustering is an unsupervised machine learning technique that allows us to categorize a dataset into groups such that samples in the same group have similar characteristics and samples in different groups have dissimilar characteristics. The aim of this project is to identify the optimal groups of patients as well as the key features that contribute most to the resulting clustering configuration. By applying k-means clustering with varying numbers of clusters to the data, we used Internal Validity Metrics (IVMs) to obtain the optimal number (5) groups of the data set. A feature selection process was then used to select the optimal set of features that have a relation with the clustering solution.

**Introduction**

Traumatic Brain Injury (TBI) is any kind of disturbance in the normal functioning of the brain [1]. Being able to cluster patients into groups will help clinicians and medical experts to make informed decisions. Clustering is a fast growing pattern recognition technique that has garnered a lot of attention in the data mining research field. It is a machine learning technique for grouping a set of data samples based on a certain criteria [3]. K-means clustering is an unsupervised learning algorithm that is used to partition a number of data points into a number of clusters [2].

**Background**

In [2] ,the authors used k-means clustering for automated detection of erythemato-squamous diseases. They experimented with five classes (psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis) and the algorithm was used to detect the five erythemato-squamous diseases .

In [9], A method has been proposed where a set of weights corresponding to a set of splines fitted to the time series data is represented by the method and utilizes “ goodness-of-a-fit” as means to assign time-series to the clusters. The disease progression profiles of patients suffering from CKD is used as reference to group the patients with the algorithm and the analysis of learnt clusters showed that the approach was successful in identifying groups with interesting and meaningful characteristics.

In [12], k-means clustering is integrated with decision tree, as well as, investigating different ways of selecting initial centroid, that is, inlier, outlier, range, random row method. for the k-means clustering algorithm to diagnose patients with heart disease. The inlier method with two clusters shows to have yielded the best accuracy.

In [13] , the Self-Organizing-Maps (SOM) method was used with a K-means clustering algorithm to efficiently find the number of centroids in traditional K-means clustering algorithms. The reason for the increased efficiency is credited to unsupervised learning method and topology preserving properties. The algorithm works in two stages. Initially it utilizes the SOM to produce prototypes and in the later stage, use those prototypes to create clusters.

**Methodology**

1. **Data preprocessing**

**1.** **Description of Dataset:**

The dataset we are working on is the ‘Traumatic Brain Injury (TBI)’ dataset acquired from Kaggle [15]. The origin of the dataset are the Center for Disease Control and Prevention (CDC) and Veteran’s Brain Injury Association. The source contains three different datasets but the dataset ‘tbi\_age.csv’ is going to be our prime concern in our project.

The following is a detailed summary of the dataset:

- 231 Entries

- Range Index: 0 to 230

- Data columns = 5

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Column Description** | **Non-Null Count** | **Dtype** |
| age\_group | Age group | 231 non-null | object |
| type | Type of measure | 231 non-null | object |
| injury\_mechanism | injury mechanism | 231 non-null | object |
| number\_est | Estimated observed cases in 2014 | 220 non-null | float64 |
| rate\_est | Rate/100,000 in 2014 | 220 non-null | float64 |

**Table 1: Dataset Summary**

The following table consists of the list of unique values in the columns containing categorical tuples:

|  |  |
| --- | --- |
| **Column** | **Names of Unique values** |
| Age group | '0-17' '0-4' '5-14' '15-24' '25-34' '35-44' '45-54'  '55-64' '65-74' '75+' 'Total' |
| Type of measure | 'Emergency Department Visit'  'Hospitalizations'  'Deaths' |
| Injury Mechanism | 'Motor Vehicle Crashes'  'Unintentional Falls'  'Unintentionally struck by or against an object'  'Other unintentional injury, mechanism unspecified'  'Intentional self-harm'  'Assault'  'Other or no mechanism specified' |

**Table 2: List of unique values in each categorical column**

**2.** **Data Cleaning**

For simplicity of calculation, we drop all the rows with Age Group equal to ‘total’ and the columns ‘number\_est’ and ‘rate\_est’ from the data set. From the dataset summary we observe that the dataset contains 11 null values. Missing values are handled by dropping all the rows of the data set containing null values.

**3.** **Encoding Categorical Features**

Some machine learning algorithms cannot efficiently handle categorical data. For that reason, we perform encoding and one hot encoding on categorical values to turn them into numerical data. Hence, we replace the values in ‘Age Group’ and ‘Type’ with numerical values to represent the categorical values. We perform one hot encoding in the‘injury mechanism’ column as the categorical values are nominal and not ordinal. Using encoding with only integer representation will misrepresent the data as having a natural order among the values and yields unexpected outcomes. So instead of integer representation, we represent such categorical data with binary representation for each unique value thus removing any such misrepresentation. After performing one hot encoding, the injury mechanism column splits into 7 different columns and each sample in the dataset will have an integer value 1 for only one column and the rest of the columns will have a value of 0, which uniquely identifies the injury mechanism of that sample.

1. **Clustering**

Using the python programming language with the scikit-learn library, we used the k-means algorithm to cluster the dataset into meaningful groups. The number of cluster parameters of the algorithm varied from 2 to 5.

1. **Visualizing the clustering results**

In order to visualize the clustering results, we used the T-distributed Stochastic Neighbour Embedding (TSNE) and Principal Component Analysis (PCA) dimensionality reduction methods. TSNE [4] variation of Stochastic Neighbor Embedding that is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map. PCA [5] is used to extract the important information from the table, to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity of the observations and of the variables as points in maps .

1. **Cluster Evaluation using Internal Validation Metrics**

The purpose of clustering is to determine the intrinsic grouping in a set of unlabeled data, where the objects in each group are indistinguishable under some criterion of similarity [6]. Internal validity metrics [8] is to evaluate the goodness of a data partition using quantities and features inherited from the datasets. Davies-Bouldin (DB) index [6] aims to identify sets of clusters that are compact and well separated. Smaller DB values indicate a better clustering solution. Silhouette (Si) index [11] computes clustering performance based on the pairwise difference of between-and within-cluster distances. Calinski-Harabasz (CH) index [10] measures between-cluster isolation and within-cluster coherence.

1. **Selecting the key Features that separate the patients into different groups**

Feature selection is proposed to select a small subset of relevant features to reduce the dimensionality of the data and maintain or increase the classification performance [14]. In our work, we used the evolutionary feature selection method for our feature selection method [14].

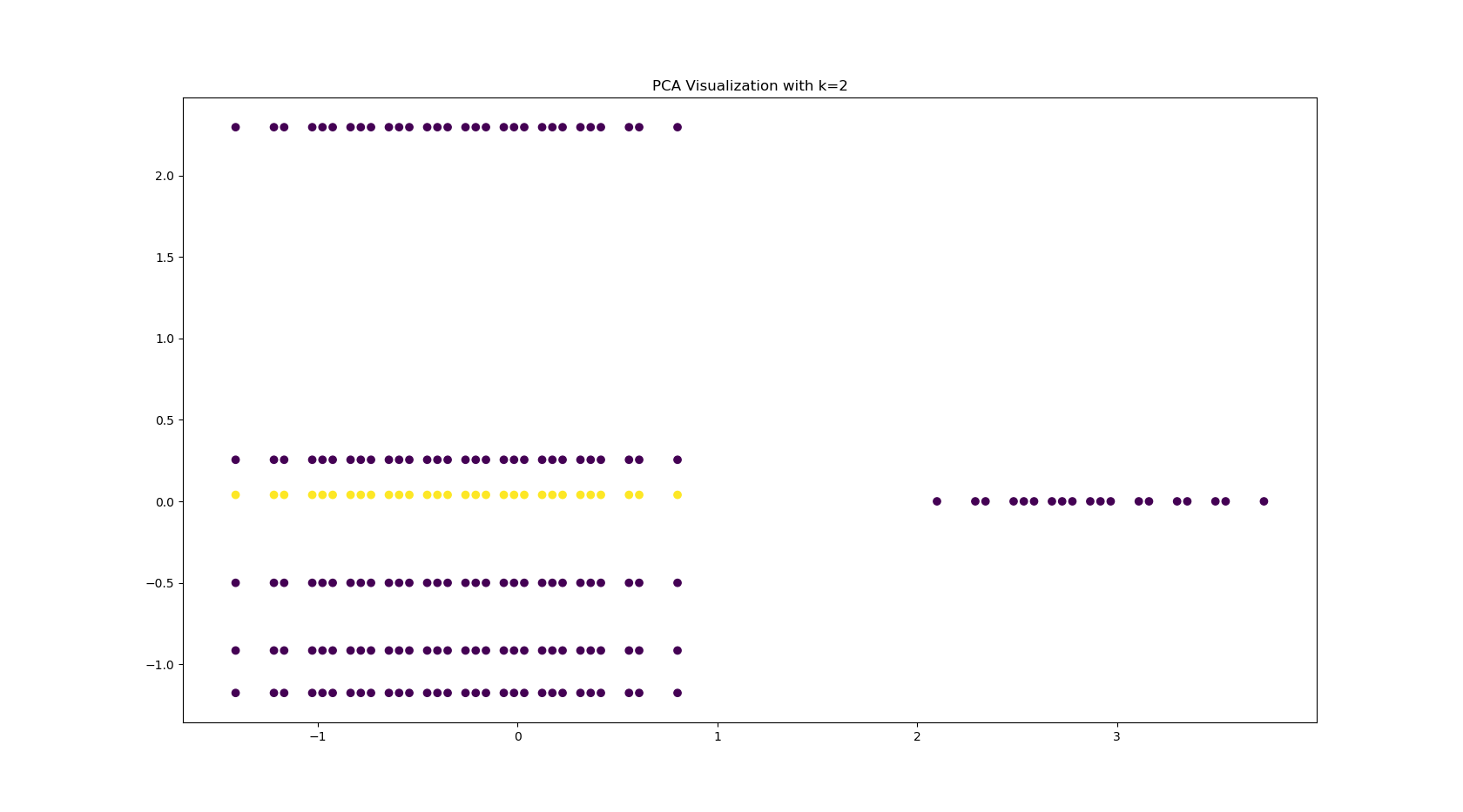
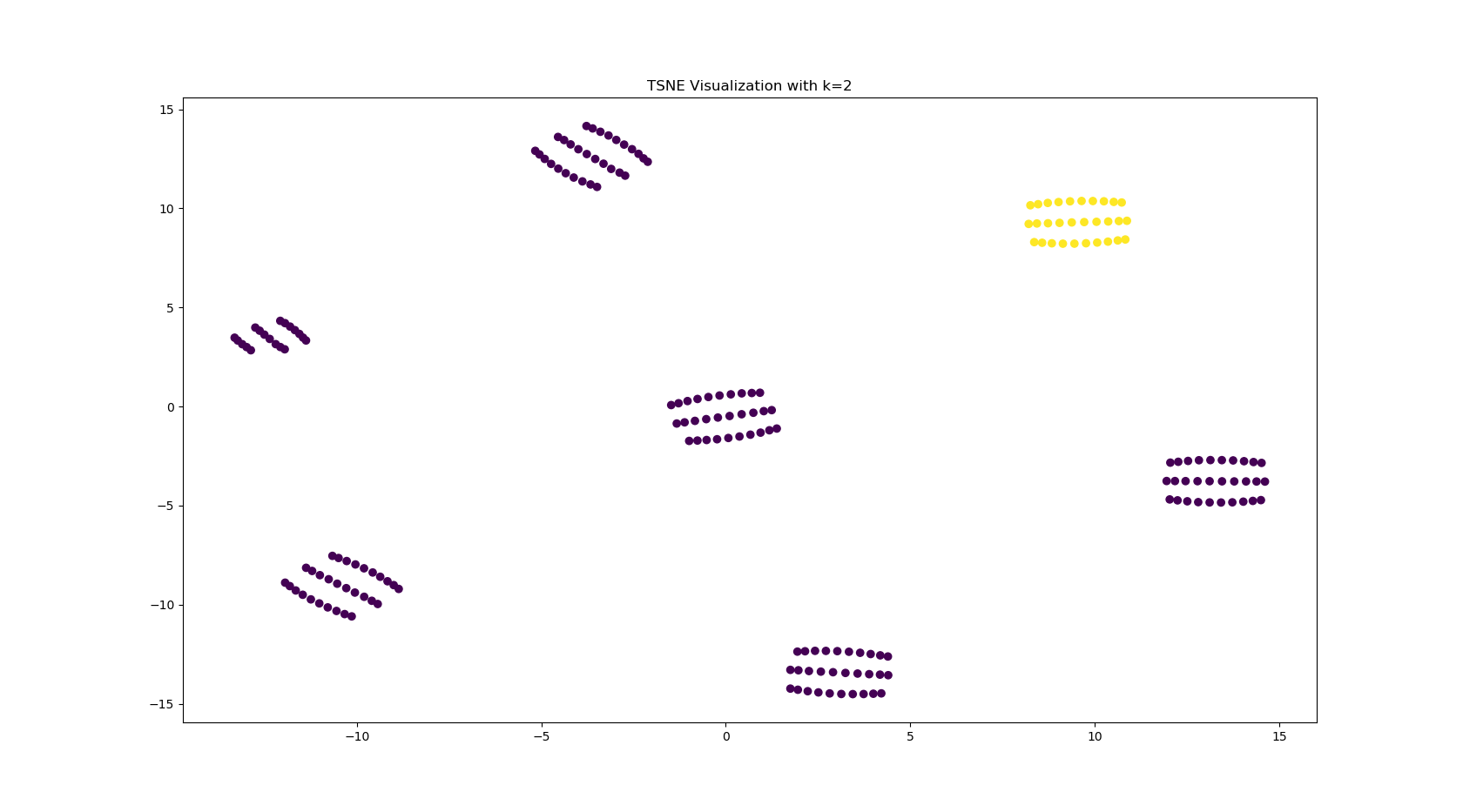
1. **Prediction model to Calculate Accuracy**

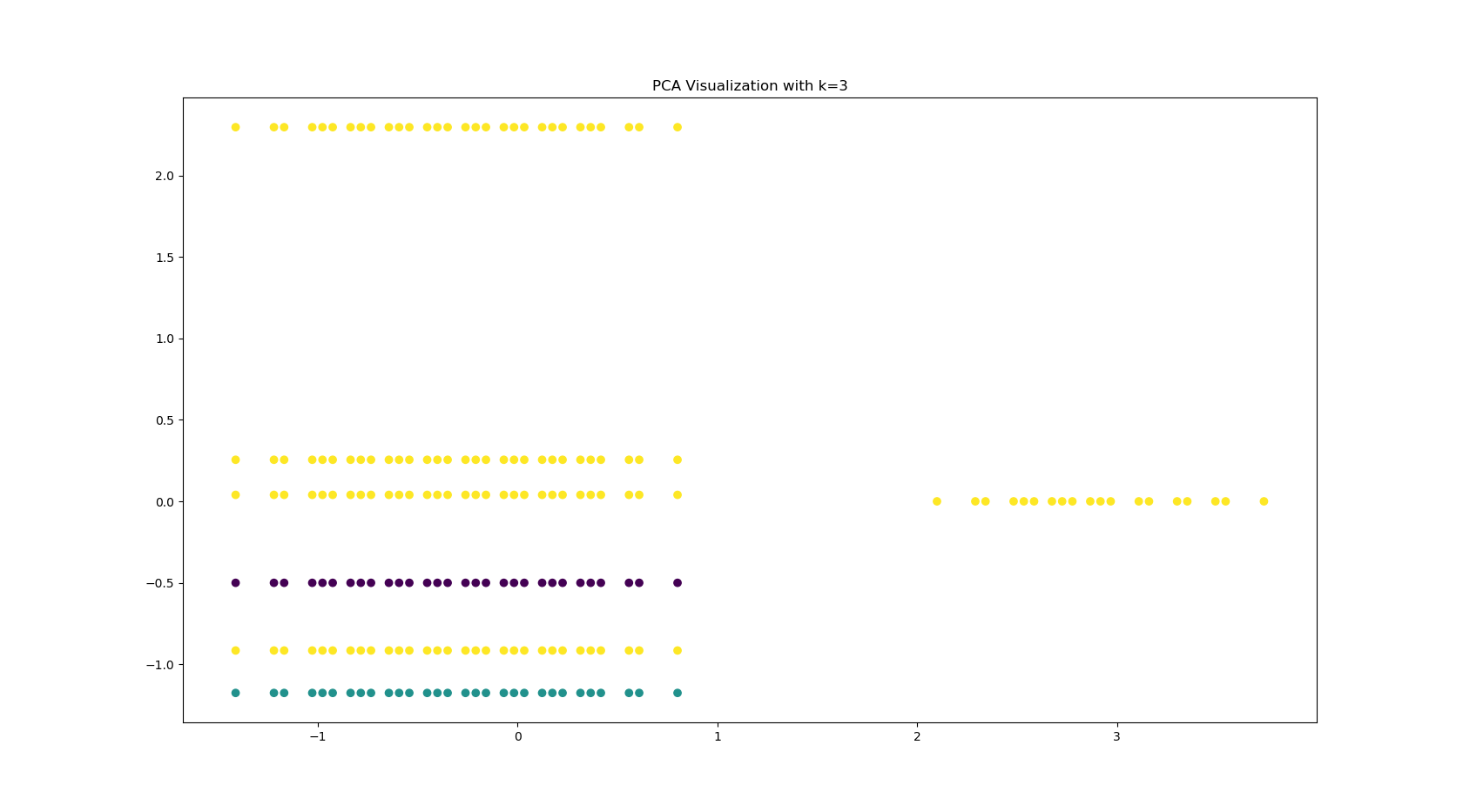
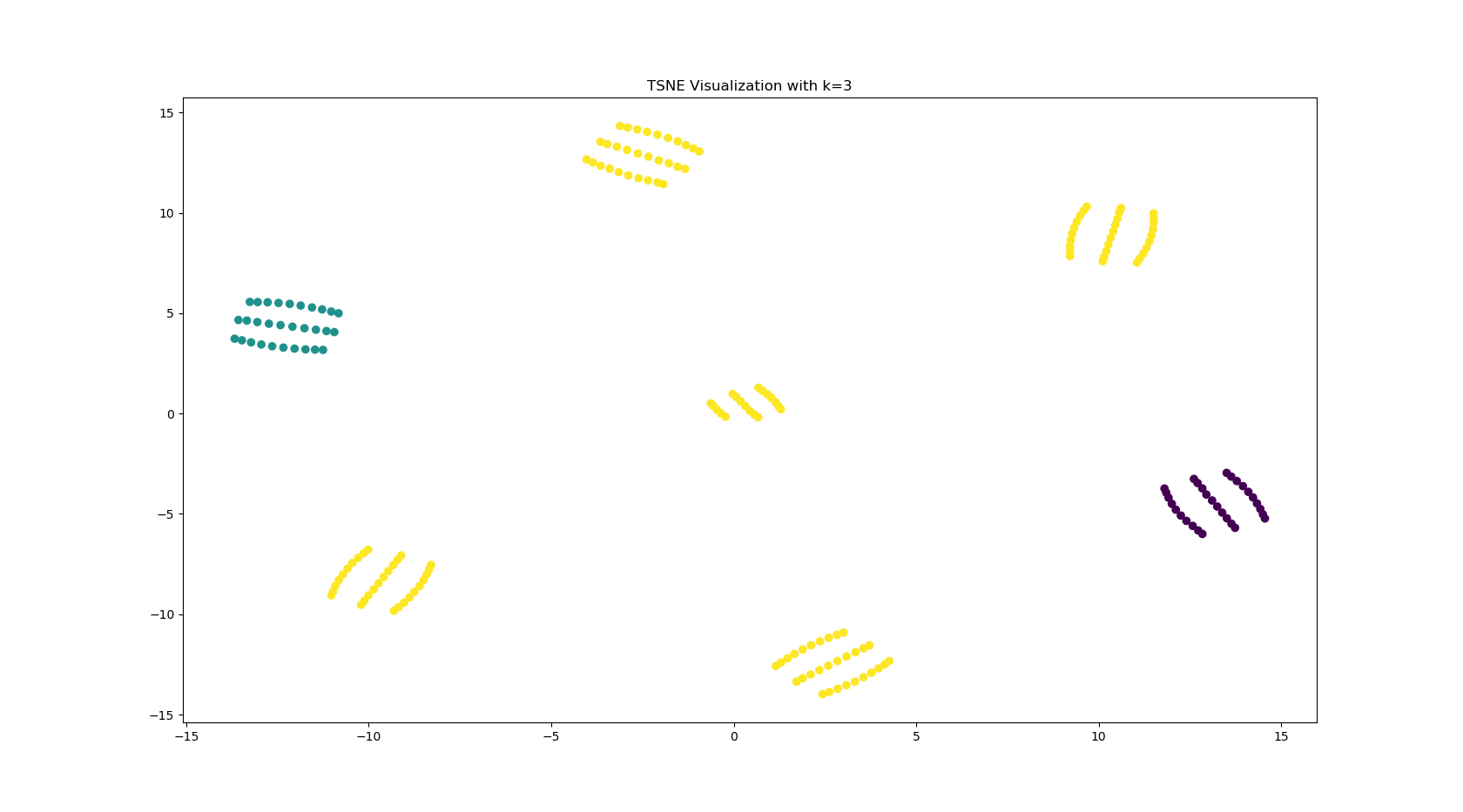
Classification models are mostly used to find out how well a dataset can be used to predict the classification classes. In this work, we treated our clustering labels as classification classes and then used the Multi-Layer Perceptron classification algorithm to classify the dataset samples into the respective clusters (classes). The process was done to measure the performance of the algorithm.

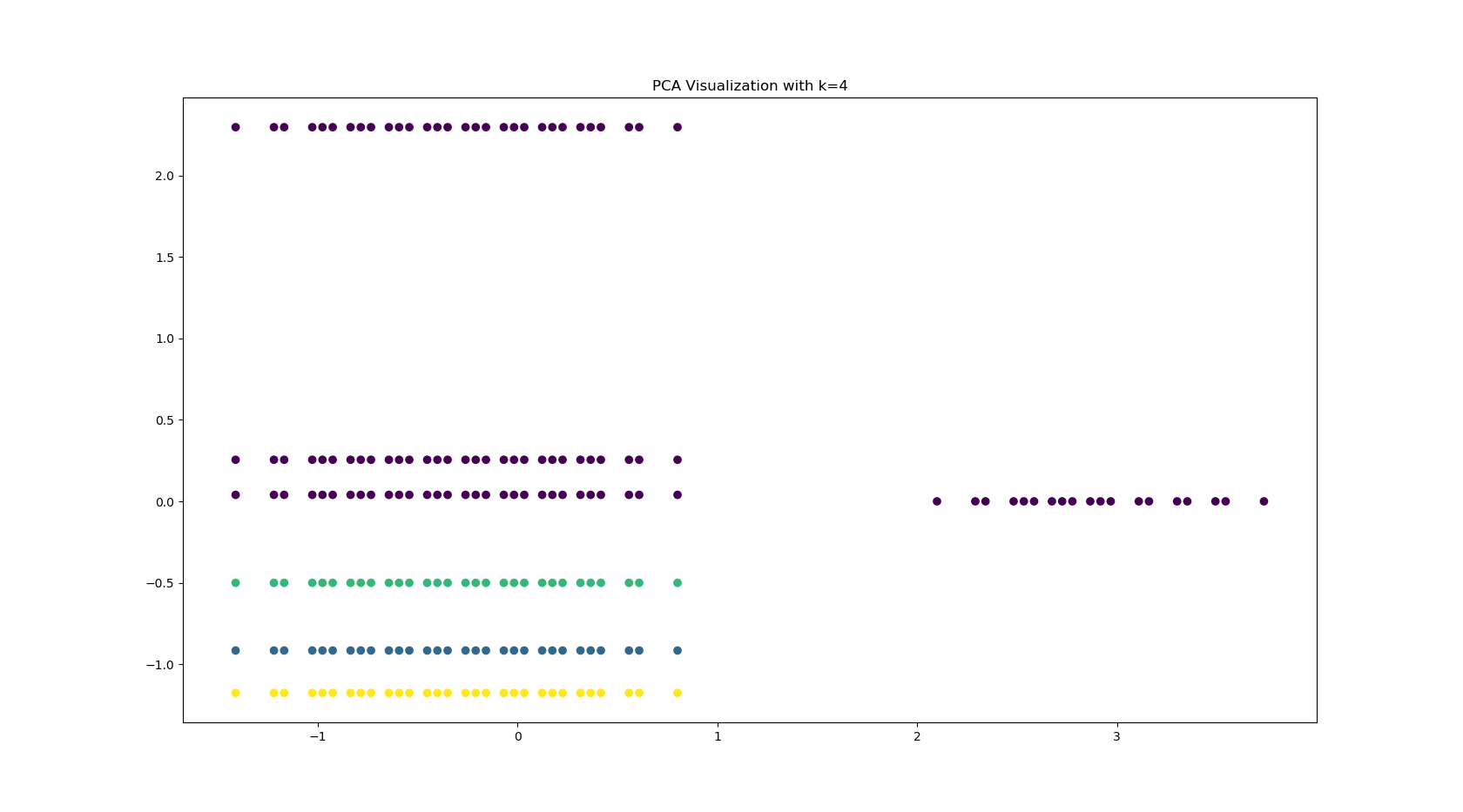
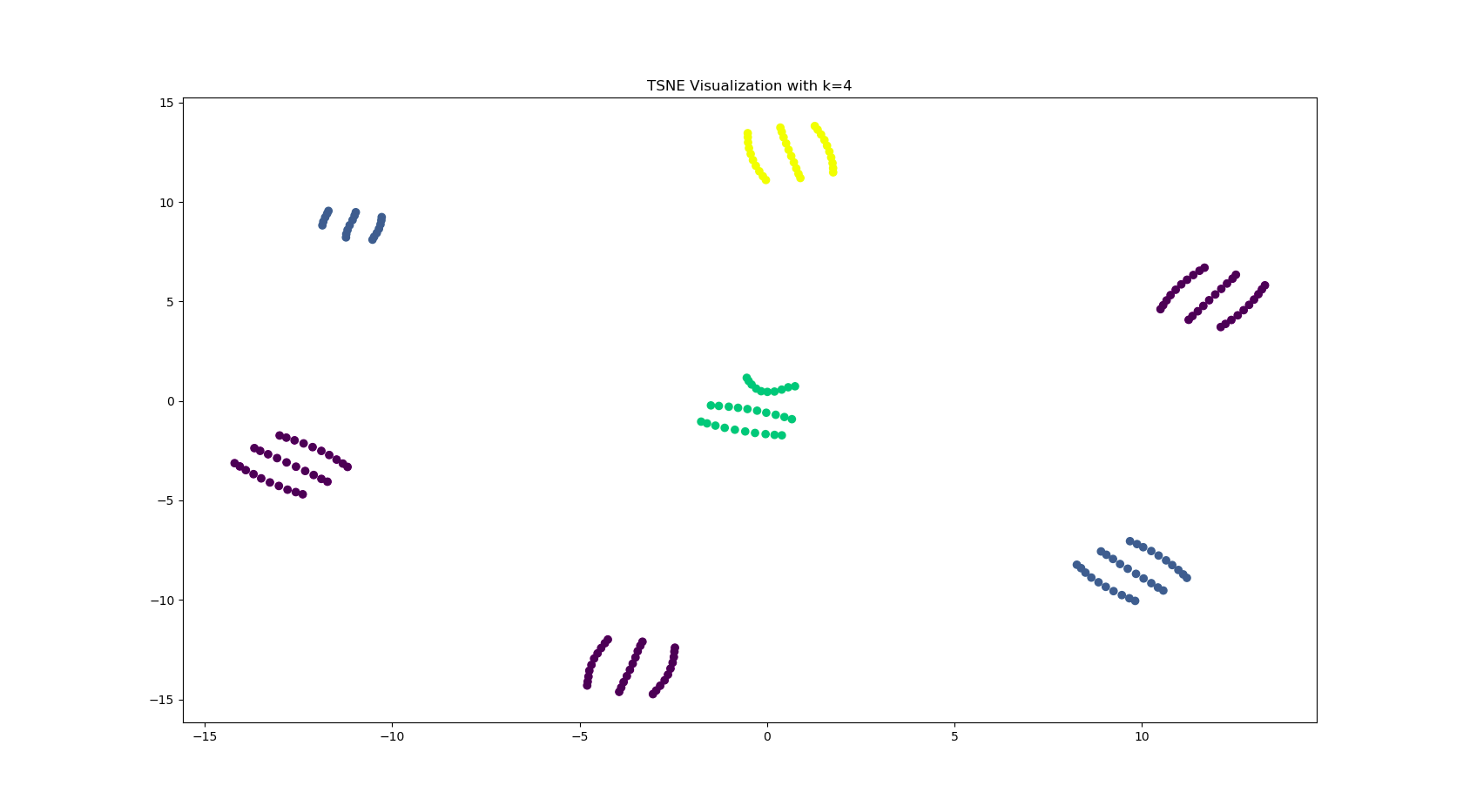
**Results**

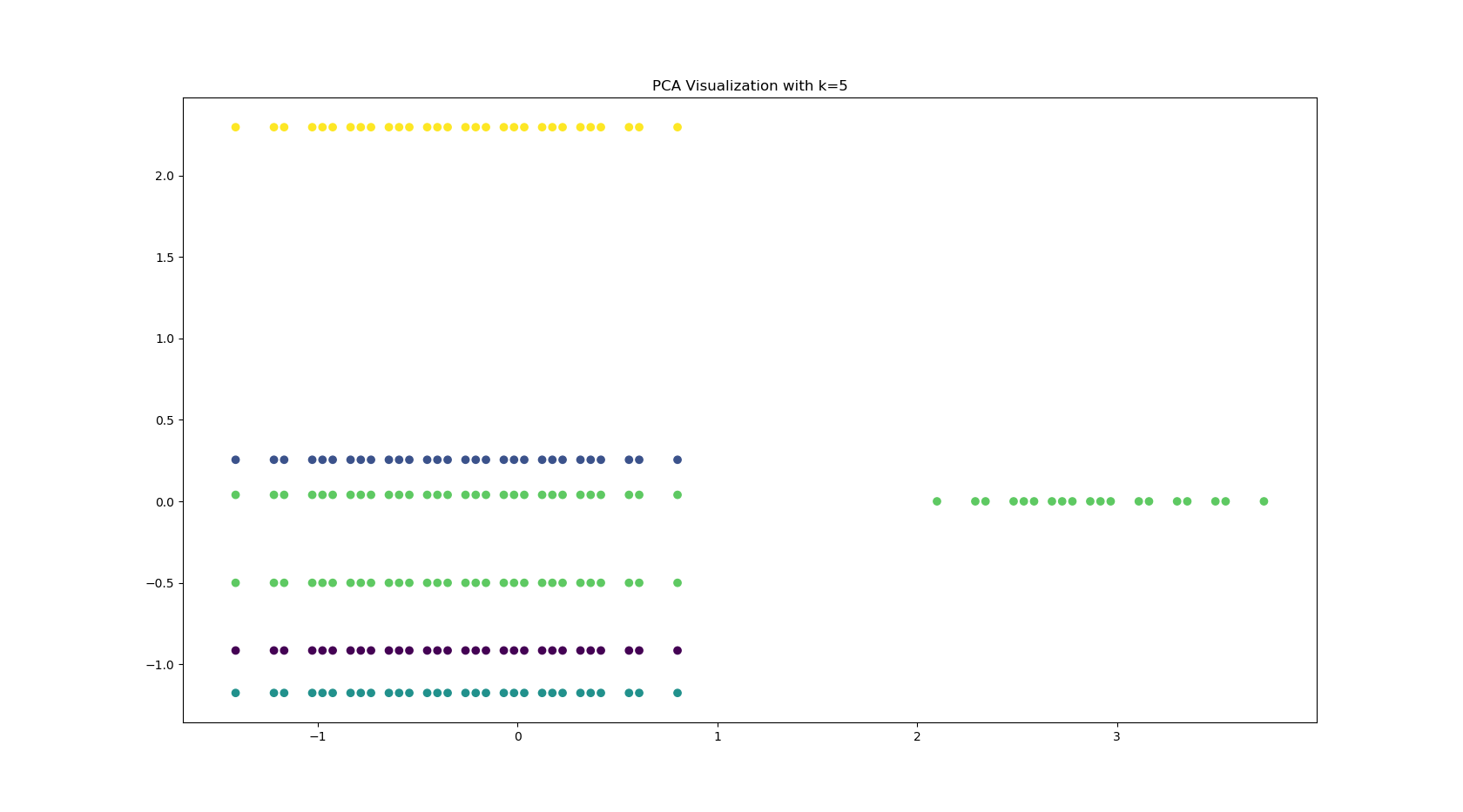
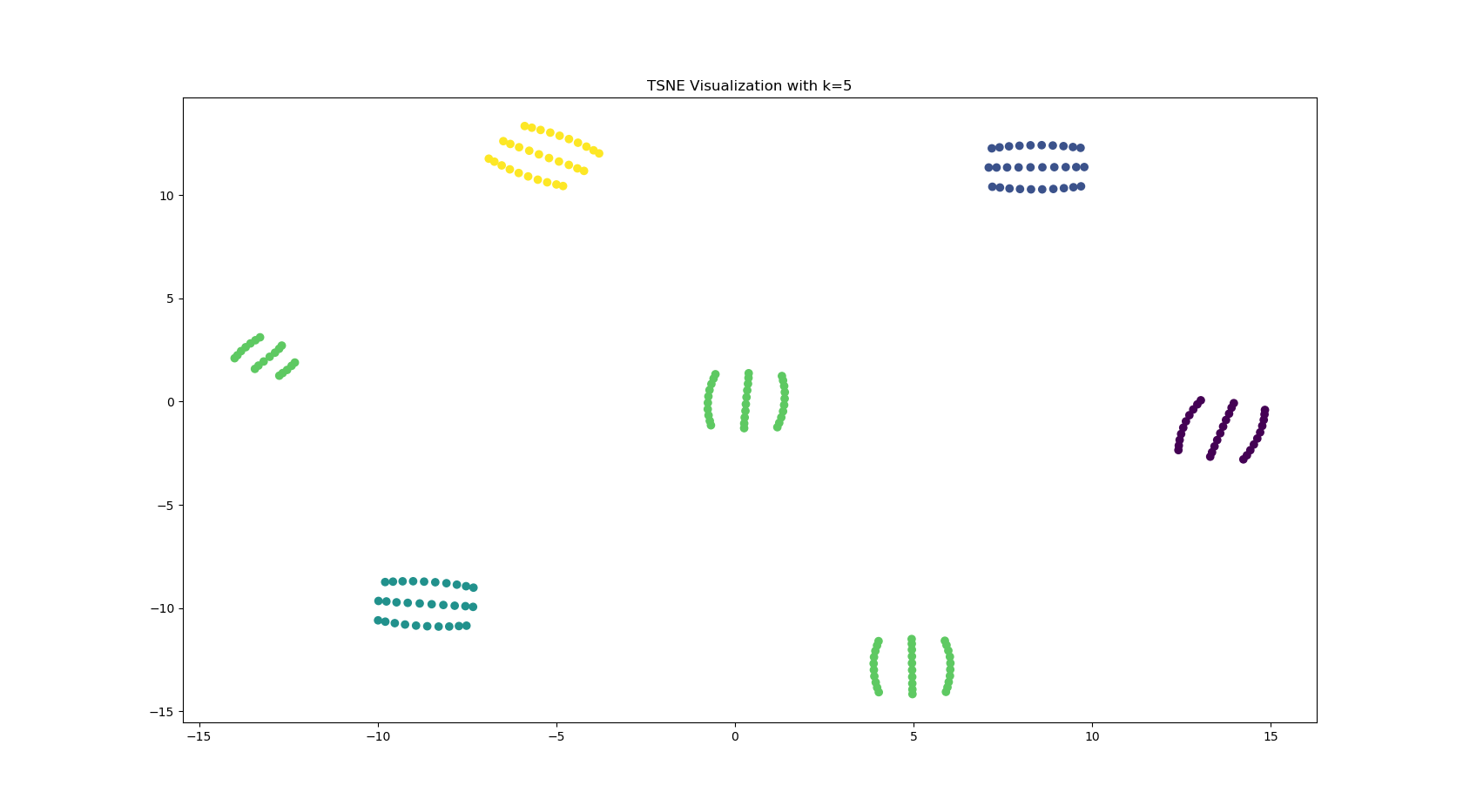
1. **Clustering Outcome**

We used the k-means clustering algorithm to obtain 5 different clustering configurations. The number of clusters varied from 2 to 5.

****

****

****

****

**Figure 1: TSNE and PCA visualizations of clustering results**.

1. **Cluster Evaluation Outcome**

We have obtained the Si score, CH score and DB score for each Cluster and ranked them ordinally from best to worst as 1 representing the rank of the best score and the 4 representing the worst. The final rank of a cluster is the summation of the Si, CH and DB rank for that particular cluster. The final rank with the least value is the best choice among the clusters. We find that the k5 Cluster performs the best on all the scores. After calculating the final rank, we discover that the k5 has the best score among the final clusters and the score depreciates with k4, k3, k2 and k1 respectively.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Si score | Si rank | CH score | CH rank | DB score | DB rank | Final rank |
| k2 | 0.15 | 4 | 30.51 | 4 | 1.42 | 4 | 12 |
| k3 | 0.24 | 3 | 28.38 | 3 | 1.39 | 3 | 9 |
| k4 | 0.32 | 2 | 40.72 | 2 | 1.34 | 2 | 6 |
| k5 | 0.41 | 1 | 53.39 | 1 | 1.26 | 1 | 3 |

**Table 3: Internal Validation Metric results.**

1. **Feature Selection and Accuracy Outcome**

The feature selection process resulted in the following 5 features from our dataset:

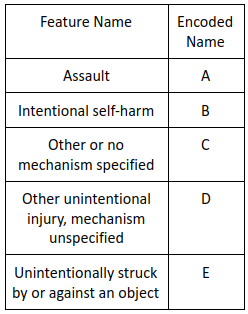
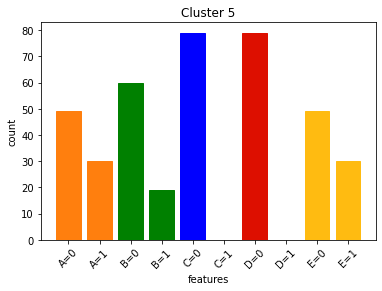
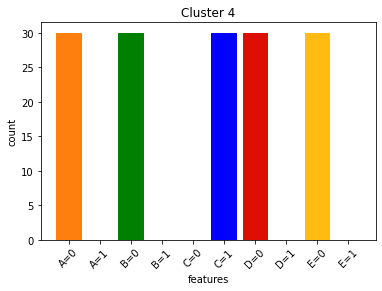
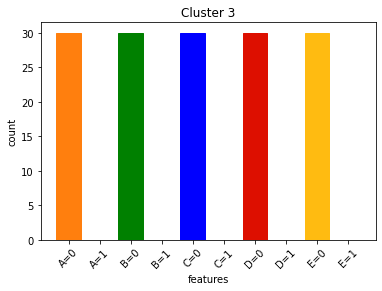
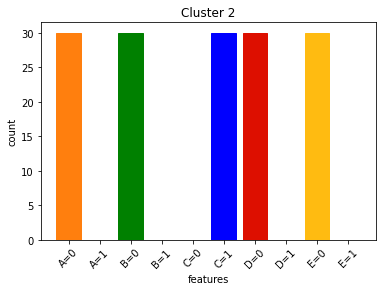
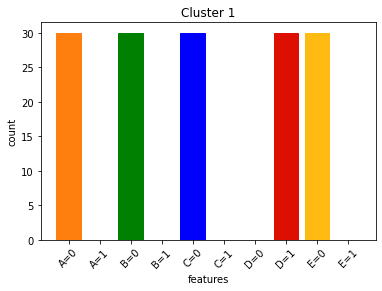
“assault”, “intentional self-harm”, “other or no mechanism specified”, “other unintentional injury”, “mechanism unspecified and unintentional struck by or against an object”. Hence these are the set of features which contribute most to the classification of the patients into the selected clustering configuration. The feature selection process dropped “Motor vehicle crashes”, “age group”, ”type” and “unintentional falls”. Using only the selected features from the feature selection process, a multi-layer perceptron (MLP) was used to classify the dataset into the 5 clustering groups. A k-Fold cross validation with a 10 number of splits was used to train the multi-layer perceptron. The MLP model accuracy was calculated for varying values of learning rates and momentum. From the above table, it can be seen that the model achieves good results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MLP Algorithm | LearningRate=0.2 Momentum=0.1 | LearningRate=0.2 Momentum=0.2 | LearningRate=0.3 Momentum=0.1 | LearningRate=0.3 Momentum=0.2 |
| mean accuracy | 84.94 | 84.94 | 84.94 | 84.44 |

**Table 4: MLP accuracy results.**

**Discussion**

From our results, we can deduce that the age group of a patient does not have any impact on the clustering result (from the feature selection process). The bar chart in Fig. 2 indicates the distribution count of the features across the various clusters. Each histogram refers to the distribution of features for a single clustering group.For simplicity, the names of the features are encoded into letters (A, B, C, D, E). For each feature, a bar is used to represent the number of patients affected by each unique value of the feature .The length of each bar represents the number of patients with a particular value for each feature. It is interesting to know that the various causes of injury did not have any impact on the cluster 3 group.. This could be that these patients could have some pre-existing genetic conditions that could be explored further. Also, only the members in cluster 5 were affected by assault and intentional harm.

**Figure 2: Distribution of features across clusters.**

**Conclusion**

In this work, we applied data-preprocessing steps to the TBI dataset and applied the k-means clustering algorithm with varying k. The results obtained were visualized using the TSNE and PCA dimensionality reduction methods. The clustering results were evaluated using well-known Internal Validation Metrics (IVM). Based on the scores assigned to the different clustering results, we selected the 5-cluster result as the optimum clustering configuration. We investigated further to determine which set of features mostly contribute to the clustering solution using feature selection technique.

**References**

[1] Ghajar, Jamshid. "Traumatic brain injury." The Lancet 356.9233 (2000): 923-929.

[2] Übeyli, Elif Derya, and Erdoğan Doğdu. "Automatic detection of erythemato-squamous diseases using k-means clustering." Journal of medical systems 34.2 (2010): 179-184.

[3] Silitonga, Parasian. "Clustering of Patient Disease Data by Using K-Means Clustering." International Journal of Computer Science and Information Security (IJCSIS) 15.7 (2017): 219-221.

[4] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.Nov (2008): 2579-2605.

[5] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

[6] Rendón, Eréndira, et al. "Internal versus external cluster validation indexes." International Journal of computers and communications 5.1 (2011): 27-34.

[7] Holmes, Geoffrey, Andrew Donkin, and Ian H. Witten. "Weka: A machine learning workbench." Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference. IEEE, 1994.

[8] Fahad, Adil, et al. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis." IEEE transactions on emerging topics in computing 2.3 (2014): 267-279.

[9] D. T. Anh Luong and V. Chandola, "A K-Means Approach to Clustering Disease Progressions," 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, 2017, pp. 268-274, doi: 10.1109/ICHI.2017.18.

[10] Bolshakova, Nadia, and Francisco Azuaje. "Cluster validation techniques for genome expression data." Signal processing 83.4 (2003): 825-833.

[11] Liu, Yanchi, et al. "Understanding of internal clustering validation measures." 2010 IEEE International Conference on Data Mining. IEEE, 2010.

[12] Shouman, Mai, Tim Turner, and Rob Stocker. "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients." *Proceedings of the International Conference on Data Science (ICDATA)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[13] Alsayat, Ahmed, and Hoda El-Sayed. "Efficient genetic K-means clustering for health care knowledge discovery." 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA). IEEE, 2016.

[14] Xue, Bing, Mengjie Zhang, and Will N. Browne. "A comprehensive comparison on evolutionary feature selection approaches to classification." International Journal of Computational Intelligence and Applications 14.02 (2015): 1550008.

[15] Mostipak, Jesse. Traumatic Brain Injury (TBI), Kaggle.com, 2020, www.kaggle.com/jessemostipak/traumatic-brain-injury-tbi?select=tbi\_age.csv.

[16] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." The Journal of machine Learning research 12 (2011): 2825-2830.