Student ID : 221183244

SIT720 Assignment 2 Report

Question 1 –

The SCADI dataset includes 6 classes namely : class1 = Caring for body parts problem; class2 = Toileting problem; class3 = Dressing problem; class4 = Washing oneself and Caring for body parts and Dressing problem; class5 = Washing oneself, caring for body parts, Toileting, and Dressing problem; class6 = Eating, Drinking, Washing oneself, Caring for body parts, toileting, Dressing, Looking after oneself health and Looking after oneself safety problem; class7 = No Problem.

From the chart below, we implement the elbow method via k-means clustering to choose the right number of clusters. The sum of the squared distance (Euclidean distance) within-the-cluster variances against the number of clusters is plotted. The first turning point of the curve is when k = 4, hence there are 4 number of clusters present. However, in the original dataset there are 6 classes (class 1 – 6) presented by attribute 206. Therefore, the subgroups and classes are different.

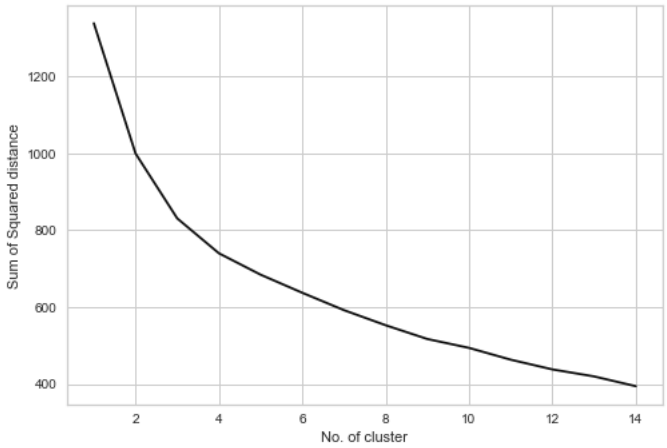

Figure 1 Cluster diagram

Question 2 –

i) The minimum dimension that captures 89% variance is 20 Principal components from the chart below.

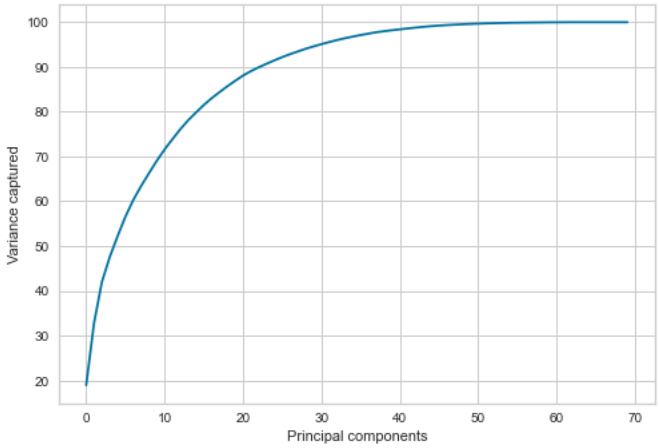

Figure 2 PCA graph

ii) Whereas 45 Principal components capture 99% variance.

We use PCA here in order to decompose a multivariate dataset for a set of successive orthogonal components that carry highest variance.

Question 3 –

The evaluation metrics for the k-means clustering models here is taken to be the purity score computed as the number of correctly matched class and the cluster labels divided by the number of total data points in the dataset.

The purity score for 89% variance captured is calculated as –

➔ Sum of accurately assigned clusters / Total = 2/70 = 0.0285

For 99% variance captured,

➔ the purity score is 1/70 = 0.0142

The figure below shows the purity scores plotted against the captured variance, we can see that the line graph is a straight line and indicates

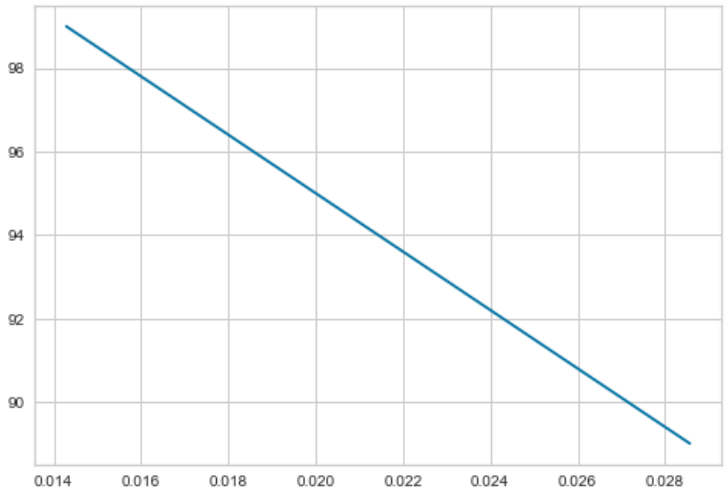

Figure 3 Purity scores plot

Therefore, as the number of clusters increases, purity score increases.

Question 4 –

Principal Component Analysis (Using Principal Component Analysis (PCA) for Machine Learning, 2022) is a method for retaining the most important features (components) from the dataset. Mostly, it removes low dimensional set of features by projecting irrelevant dimensions from a high dimension dataset to capture as much information as possible. PCA is also more useful when dealing with 3 or higher dimensional data. PCA is also always applied on a symmetric correlation or covariance matrix having numeric and standardized data. There is no evidence that shows that PCA methods can be implemented on curved structural data, hence, PCA can only be applied on linear dataset.

Question 6 –

There are two clusters (0 and 1) showcased here which is equal to the number of classes in the dataset shown indicating the diagnosis of the heart disease in the patient :

Value 0 - <50% diameter narrowing (CAD, no)
Value 1 - >50% diameter narrowing (CAD, yes)

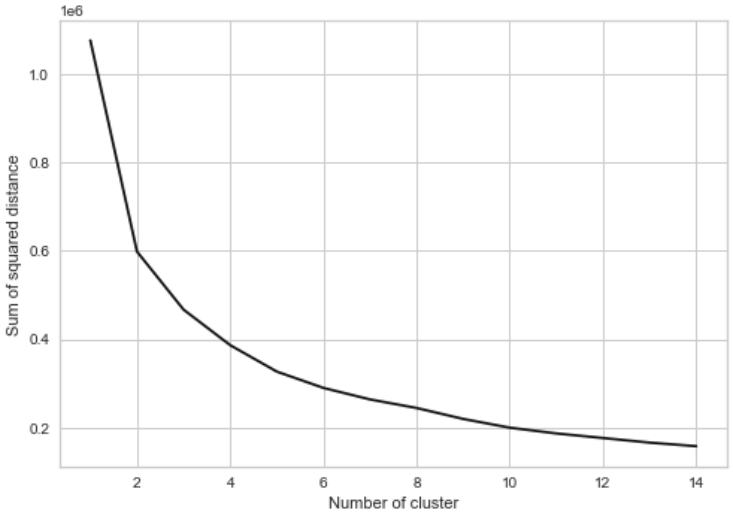

Figure 4 Number of clusters, k

Question 7 –

The purity score is calculated as the number of correctly assigned classes divided by total datapoints :

= 126/297 = 0.424

Question 9 –

We determine the three best features using the SelectKBest function

Using the chi2 selection criteria, the three best features selected are :

Thal : 3: normal, 6:fixed defect, 7: reversable defect having 6.522 e+01 as the significant F-stat.
Thalach : maximum heart rate achieved having 1.870 e+02 F-statistic
Oldpeak : ST depression induced by exercise relative to rest having 5.418e+01 F-statistic.

This displays whether there is a statistically significant difference between the expect and the observed metrics output, and higher the number, better the significance.

Using the ANOVA test, the three best features selected are :

Thal : 3: normal, 6:fixed defect, 7: reversable defect having 1.132e+01
Ca : Number of major vessels (0-3) coloured by fluoroscopy having 8.0577e+01
Thalach : maximum heart rate achieved having 6.458e+01

Purity Score	**Chi2**	**ANOVA**
	0.562	0.562

The purity scores for both chi2 and ANOVA are similar, I would recommend we can use any set having features thal, thalach and olpeak or ca, since they have the highest significance level with the target variable 'num'.

Q10 –

a) The classification models used in this scenario are :

Random Forest Classifier	Logistic Regression	KNeighbors Classifier	Linear SVC
0.866	0.933	0.633	0.9

From the below chart we can see that without any hyperparameter tuning, the best performing models in order are :

1) Logistic Regression
2) Linear SVC
3) Random Forest
4) KNeighbors

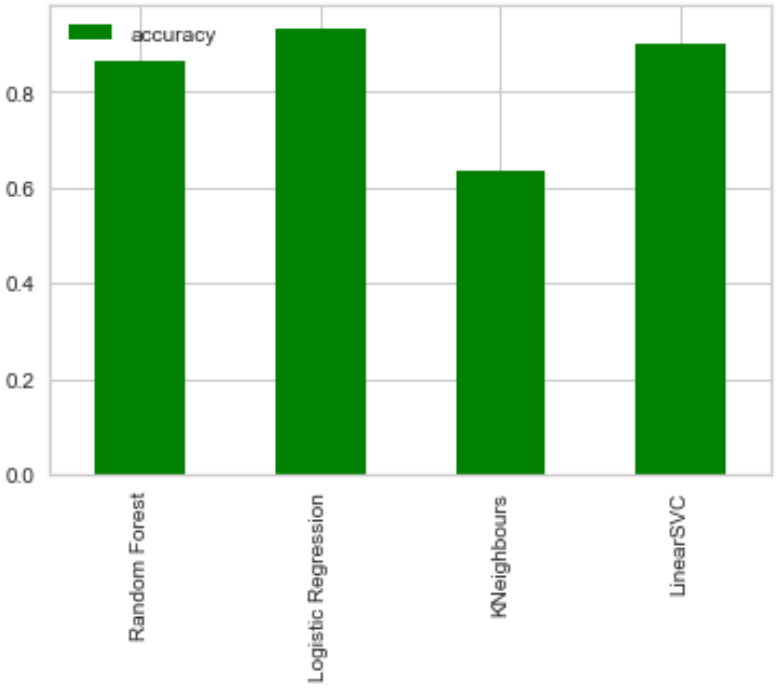

Figure 5 Accuracy of the models

b) The generalisability of the above models would involve hyperparameter tuning, implying that in order to improve the accuracy the model, we would need to change certain features of the model. For instance, we can apply Cross Validation method with GridSearchCV and tune the model with the best parameters. Hence, we can

conclude that Logistic Regression and Linear SVC have the best accuracy scores, which could be further used in analysis.
In this scenario, we will use the Logistic Regression model which works well with identified (structured dataset) rather than Linear SVM model – which works well on unstructured data.

c) I have used several other models for model comparison, such as Random Forest Classifier to report the model scores. Out of all the above models, Logistic Regression performs the best.

Q-11

a) I have created a synthetic dataset from scratch using the dictionaries and pandas libraries.

b) The variables used to create the dataset are :

- Gymnasium – If the resort has a gym, then it is 1 otherwise 0.
- Dine in restaurants – The number of restaurants in the resort , for example – 1,2 or 3 restaurants.
- Swimming pool – 0: No swimming pool, 1: Kid's pool , 2: Indoor swimming pool, 3: Both swimming pools.
- Conference rooms – Number of conference rooms – 0: none, 1: 1 conference room
- Resort Location – 1: Ski resort, 2: Beach resort , 3: Cultural and heritage resort.

c) Here, the target variable is the 'Season' where, 1: Summer, 2: Winter, 3: Autumn

d) We require an ML model for this problem to classify which resorts are appropriate based on the season of travel. For instance, a Random Forest Classifier model would better classify the number of resorts which tells us that it is appropriate during winter season or summer or autumn season. Or a Support Vector Machine (SVM) would better classify the number of resorts – hence, a machine learning model would be the best to use in this scenario.

REFERENCES :

1) Analytics Vidhya. 2022. *PCA: Practical Guide to Principal Component Analysis in R & Python*. [online] Available at: <https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/> [Accessed 1 May 2022].
2) En.wikipedia.org. 2022. *Chi-squared test - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Chi-squared_test> [Accessed 1 May 2022].
3) Medium. 2022. *Using Principal Component Analysis (PCA) for Machine Learning*. [online] Available at: <https://towardsdatascience.com/using-principal-component-analysis-pca-for-machine-learning-b6e803f5bf1e> [Accessed 30 April 2022].