**LAB REPORT 2**

**DataSet :** The DataSet which I am currently working for this Lab is “**BasketBall College Team**”.

This dataset consists of all the details of USA inter college Basketball competitions held in the year 2018.

This DataSet consists of 350 rows and 16 columns.

The column labels are : ['Games Played', 'Won', 'Offensive Efficiency', 'Defensive Efficiency', 'Power Rating',

'Field Goal Shot', 'Field Goal Allowed', 'Turnover allowed',

'Turn Over Commited', 'Free Throw Shot', 'Free Throw Allowed', 'Two Point Shot', 'Two Point Allowed', 'Three Point Shot',

'Three Point Allowed']

**Task 1: data clustering and decimation**

Implement random sampling and stratified sampling (remove 75% of data)

Stratified sampling includes the need for k-means clustering

A screenshot of a cell phone

Description automatically generated

Above snippet contains code for implementing random sampling.

A screenshot of a social media post

Description automatically generated

Above snippet contains code for implementing Stratified sampling. As we can see, stratified sampling is implemented using k-means clustering. Code is k-means clustering is mentioned below

A screenshot of a cell phone

Description automatically generated

Both random sampling and Stratified Sampling mentioned above removes 75% of the data.

A picture containing photo, kitchen, lot, table

Description automatically generated

Elbow curve used for selecting number of clusters in stratified sampling. In this case, elblow is inbetween 4 – 8 and I’ve chosen 4 as number of clusters for k-means clustering.

A screen shot of a smart phone

Description automatically generated

In the above code snippet, line plot used to draw lines in d3.js is shown.

**Task 2: dimension reduction on both org and 2 types of reduced data**

1. Find the intrinsic dimensionality of the data using PCA

**A close up of text on a screen

Description automatically generated**

As shown in the above code snipper, **PCA()** is used to apply Principal Component Analysis to the entire dataset. After that we used pca.explained\_variance\_ to determine the intrinsic dimensionality of the each pca component.

1. Produce scree plot visualization and mark the intrinsic dimensionality

Above figure consists of code which is used to generate a dict of pca\_variance and no of pca components. Once the dict is completed, it is sent to front end in json format and D3 creates a line plot from the data.

**A close up of a map

Description automatically generated**

Apart from PCA, we can use MDS i.e., Multi Dimensional Sampling and this can be done using python libraries MDS()

**A screenshot of a cell phone

Description automatically generated**

Shown above is the code for determining the intrinsic dimensionality.

**Observation:** If the data set is huge and dimensions are more, then PCA is preferred to MDS the execution time of MDS is too high when compared to PCA. For smaller datasets, the difference between PCA and MDS is smaller.

1. **Show the scree plots before/after sampling to assess the bias introduced**

As shown in the below figures, after sampling(Random as well as Stratified), there is a minute level of bias introduced.

**A picture containing text, photo, table, laptop

Description automatically generated**

**Scree Plot of Original data**

**A picture containing text, map, photo, table

Description automatically generated**

**Scree Plot of Randomly Sampled data**

**A picture containing photo, table, white, laptop

Description automatically generated**

**Scree Plot of Stratified Sampled data**

1. **Obtain the three attributes with highest PCA loadings**

As shown in the below code, “**Turn Over Commited,**

**Turnover allowed, Free Throw Allowed**” are the three highest

PCA attributes. Scatter Matrix is also plotted with these three

attributes.

**A screenshot of a social media post

Description automatically generated**

**Task 3: Visualization of both original and 2 types of reduced data**

1. **Visualize the data projected into the top two PCA vectors via 2D scatterplot**

**A picture containing text, table, photo, large

Description automatically generated**

Based on the loadings we have PC1 and PC2 as top two PCA vectors so we will plot both the components in x axis and y axis respectively.

1. **Visualize the data via MDS (Euclidian & correlation distance) in 2D scatterplots**

**A screenshot of a video game

Description automatically generated**

Shown above is the code for implementing scatter plot in 3D.js

A picture containing large, light, computer, lot

Description automatically generated

**MDS Eucilidean Scatter plot**

**A picture containing large, filled, lot, white

Description automatically generated**

**MDS Corelation Scatter plot**

1. **Visualize the scatterplot matrix of the three highest PCA loaded attributes**

As we already mentioned, “**Turn Over Commited,**

**Turnover allowed, Free Throw Allowed**” are the three highest PCA loaded attributes.

A close up of a map

Description automatically generated

Scatter Plot of highest three loaded Attributes

A screenshot of text

Description automatically generated

**Code for implementing scatter plot matrix**

A screenshot of a cell phone

Description automatically generated

**D3 code for scatter plot matrix**

A screenshot of a social media post

Description automatically generated

**Code for determining highest PCA attributes**

**YouTube Video Link:** <https://youtu.be/BLbLjkZdJQw>