ML Lab Report 12

Name: Raman Kumar
Reg NO: BL.EN.P2DSC22009
Course: Mtech Data Science

1. Describe a scenario where PCA would not be able to help reduce the dimensionality of the available training data.

In a scenario where Netflix is trying to cluster movies and TV shows based on their features (e.g. genre, director, actors, etc.), PCA might not be the best option if the structure of the data is highly non-linear.
Imagine that we have a dataset of movies and TV shows that we want to cluster based on their features (e.g. genre, director, actors, etc.). The relationship between these genres is not necessarily linear, and it might be difficult to capture this structure by a linear projection.
Additionally, movies and TV shows may share some genres, but appeal to different audiences, this aspect is not captured by PCA. For example, a movie like "Jurassic Park" is of the genre "action" and "sci-fi" but it may appeal more to kids than an action movie like John Wick.

In this case, other dimensionality reduction techniques such as t-SNE, which are able to capture non-linear structures could be used instead. Moreover, a more advanced technologies such as deep learning-based autoencoder could also be applied to learn the underlying structure of the data and cluster them. It's more important that PCA or any other dimensionality reduction technique alone may not be sufficient to give good clustering results and other techniques such as k-means, hierarchical clustering, or clustering using deep learning models should be considered along with dimensionality reduction. Also, The success of clustering depends on the quality of the data, feature representation, and the number of clusters.

Q.2 Compare and summarize the effects of various data transformation techniques on your project. Which is the most suitable technique for your project data. Justify your answer.

Data transformation techniques are used to preprocess the data in a way that makes it more suitable for analysis and modeling. Some common data transformation techniques include

- Normalization,
- Standardization, and
- Scaling.

Normalization is a technique that is used to scale the data so that it falls within a specific range, usually between 0 and 1. This is useful for data that has a large range of values, as it can help to prevent certain variables from having too much influence on the analysis. One example of normalizing data is Min-Max normalization which scales the data to the range of [0,1].

Scaling is a technique that is used to change the range of the data while maintaining the relative proportions of the values. This can be useful for data that has outliers or extreme values, as it can help to prevent them from having too much influence on the analysis.

Standardization is a technique that is used to center the data around a mean of 0 and a standard deviation of 1. The Z-score method is most commonly used.

In this Netflix Movies and TV shows clustering project I want to group similar movies and shows together. To do this, it's important to make all the data similar in size so that one variable does not influence the analysis more than the others. That's why the best technique to use for this project is standardization which makes all the data in the same range.