EXECUTIVE SUMMARY FOR KMEAN

My task for the assignment was on implementing an unsupervised learning algorithm well known for its classic ability to automatically group data into clusters based on inherent features defined within the data itself. Kmeans immense performance is evidence in its efficient clustering of data with no prior labels nor category. In order to garner more insight in the task I employed the usage of the wine dataset which is very unique in machine learning due to its diverseness of features. The aim of the using the KMeans clustering was to separate the dataset into distinct groups based on similar inherent attributes. The wineqality-red.csv which I used is composed of different chemical combinations making up diverse varieties. Each row of the dataset identifies a wine sample with its unique characteristics containing acidity or alcohol content. The task description are as follows.

**Import libraries and data collection**

To ensure proper objective of generating clusters from the winequality -red.csv dataset vital libraries that are necessary where installed and import for use such as pandas for data manipulation, matplotlib for visualization and essential libraries to implement clustering such as standard scaler and Kmeans from Scikit-learn

**Exploratory data analysis**

At this phase of the task, I had used statistical and graphical method to identify patterns with the dataset and provide gainful insights in understanding the underlying relationship existing with inherent features of the dataset

**Data preprocessing**

The datasets where adequately normalized to ensure it fits properly and biased insight is brought to its barest minimum. Their silhouette and ARI(adjusted rand index metric used to evaluate the quality of cluster algorithm) scores were determined. Silhouette(determines how well datapoints fits within assigned clustered compared to other clusters ) Score ranges from -1 to 1, where a higher value indicates better-defined and well-separated clusters. ARI ranges from -1 to 1, where a value close to 1 indicates a strong agreement, a value close to 0 indicates random labelling, and a value close to -1 indicates a strong disagreement. For the features scaled using Standard Scaler, the K-means models achieved a silhouette score of 0.205 . The ARI scores for the models was 0.034 .The Scatter plot provides a visual representation of the clusters formed by K-means.

Challenges and solution

The task of implementing clustering using KMeans revealed vital patterns within the win dataset however but not without its daunting challenges of knowing an optimal (K)cluster and having to work with data of multiple dimensions.with a good values derived in silhouette analysis it is achievable