Name: Srikanth Reddy Dongala
Reg No: INBT0707
Batch: June 2023
LinkedIn: https://www.linkedin.com/in/srikanth-dongala/
GitHub: https://github.com/Srikanthdongalajsr

Task 1 - Questions

1. **Which Python library did you find most useful in loading and exploring the dataset?**

   The libraries that are vital for loading and exploring the dataset are:
   Numpy: To perform numeric calculations in a better an efficient way
   Pandas: To load data and analyze the data with some powerful methods available.
   Matplotlib, seaborn, plotly express: To visualize the data, which enables us to gain more insights from the data and draw conclusions effectively
   Sklearn: This is a powerful tool to apply ML algorithms on the dataset and also evaluate the performance of the model.

2. **What preprocessing steps did you find necessary to apply to the heart dataset?**

   The dataset is almost perfect for feeding to the model, it doesn't have any null values, wrong datatypes, or wrong data, but the necessary steps are transforming the data of skewed features (cholesterol and st_depression), both these features are right skewed hence applied the root transformation to make the data to a normal distribution.

   Outlier treatment is also needed but since the original source of the data is not known we cannot be so sure if they are true outliers, still we found the best model as Decision Tree Classifier and Random Forest Classifier which are robust to outliers hence, in this particular analysis outliers are not treated.

3. **What metrics were used to evaluate the classification problem and why?**

   A confusion matrix is used to check the values of FP and FN since both of them are equally important in this dataset, F1 score is also checked along with the ROC curve to see how the model performs at varying thresholds i.e. the AUC score.

4. **How to detect an overfitting problem in the model and what strategies to mitigate it?**

Overfitting is a problem where the model performs well on the trained data but fails to replicate the same level of performance on the test data. If we see an overfitting problem in the model (which is occurred due to the complexity of the model) we can use regularization techniques where we add a penalty term to the loss function. Also, we can try to reduce the complexity of the data, by performing feature selection. If we still see overfitting then we need to do hyper-tuning for the model and check the best parameters that result in a more generalized model.

5. **How do you choose the number of clusters for the K-Means algorithm? Explain why?**

   K-Means algorithm is a type of unsupervised algorithm, where there won't be any labeled data (dependent variable) but uses the relation among the features in the data and tries to form clusters so that every item in the cluster will be similar to the other items in the same clusters but differ from other clusters.

   The main challenge in this type of problem is to find the optimal number of clusters for the chosen dataset. The methods that are available to find the optimal number of clusters are:

   a. Elbow Curve Method: Using this we try to calculate the optimal number of clusters, where we see a bend in the curve i.e., after this point, the change in inertia/distortion is very small compared to the change with the clusters below the optimal value.

      However, this method has the drawback of considering only the within the cluster distance in evaluating the optimal number of clusters i.e., only cohesion will be considered.

   b. Silhouette Score: This is an alternative to the Elbow curve method (WCSS v/s cluster size) where both cohesion and separation are considered in deciding the optimal number of clusters, the range of the score is [-1,1]. The higher the score the better the model, hence we need to observe which number of clusters is resulting in the best score.