EEE 591
Project 1
6th October, 2023

Name: *Ipsit Sahoo*
1224872415

Project 1

Problem 1: Exploratory Data Analysis (EDA)

In our analysis of the dataset, we focused on understanding the covariance and correlations among the variables. The dataset comprises 14 columns, with 13 columns containing pertinent features and attributes related to human anatomy. The 14th column, denoted as **a1p2**, serves as the target feature that we aim to predict using the other attributes.

First, we examined the correlations between these 13 attributes and the target feature a1p2. The top 10 features displayed correlations within the range of **[0.18, 0.52]**. Among them, the top three highly correlated features were **thai**, **nmvcf**, and **eia**. We also explored the correlations between individual features. The top 10 such correlations ranged from **[0.39, 0.61]**. Notably, the features thal, nmvcf, eia, mhr, opst, cpt, dests, sex, age, and rer exhibited significant correlations with the target feature a1p2.

To determine whether to retain the features rbp, sc, and fbs, which showed weaker correlations, we examined their relationships with other attributes in the dataset. rbp demonstrated a strong correlation with age, suggesting its relevance, and thus, we included it in the final dataset.

We also considered covariance as a metric for feature selection. Both **rbp** and **sc** exhibited high covariance with each other and with the target feature a1p2, reinforcing their significance. Consequently, we incorporated rbp and sc into the final dataset. Although **fbs** did not display very high correlation or covariance, its association was still substantial, leading us to include it as well.

Problem 2: Machine Learning Algorithms

In this report, we present the outcomes of evaluating six distinct Machine Learning Algorithms. Prior to applying these algorithms, the data needs to be divided into training and testing sets. To achieve this, we employ a validation size of 0.2, which results in an 80% allocation to the training set and a 20% allocation to the testing set.

To ensure consistency and comparability across the datasets, we apply a Standard Scaler. This process scales both the training and testing sets to a uniform set of values, simplifying the application of our chosen algorithms.

The **Perceptron** algorithm has been run over tolerances ranging from 10^{-4} to 1. The algorithm achieved the best accuracy of 90.74% at tol=1.

The **Support Vector Machine** algorithm has been run over multiple kernels. The algorithm achieved the best accuracy of 94.44% by using the 'rbf' Kernel.

The **K-Nearest Neighbors** algorithm has been run from 1 to 99 different k values. The algorithm achieved the best accuracy of 92.59% at k=13.

Machine Learning Algorithms	Min. Accuracy	Max. Accuracy
Perceptron	75.92%	90.74%
Logistic Regression	-	90.74%
Support Vector Machine	92.59%	94.44%
Decision Tree Learning	-	77.78%
Random Forest Classification	-	92.59%
K-Nearest Neighbor	81.481%	92.59%

IMPORTANT OBSERVATIONS

- In the context of the provided dataset, we have made the decision to retain all available features. This choice is underpinned by the observation that these features exhibit substantial covariances and correlations with one another, collectively contributing to a meaningful influence on the predicted results.
- The best algorithm that should be opted for based on the results in **Support Vector Machines** (SVM) with an 'rbf' kernel which achieved an accuracy of 94.44%.
- To avoid the chance of overfitting, we should also consider Random Forest and K-Nearest neighbors
- Due to the small dataset size of around 270 rows, it is recommended to seek additional data to improve the effectiveness of machine learning algorithms.