PROJECT-1 REPORT

ANALYSIS FROM PROBLEM-1:

Absolute values of correlation and covariance of each variable with all other variables is calculated in the python script. The variables which are most highly correlated with each other are 'dests' with 'opst' (correlation value: 0.609712) and 'a1p2' with the variables 'thal', 'nmvcf', 'eia', 'mhr', 'opst' and 'cpt' has very high correlation. These 6 variables which are highly correlated with 'a1p2' will play a significant role in predicting the heart disease as 'a1p2' is our target variable. 'Opst' is highly correlated with 'a1p2' and 'dests' hence only one of 'dests' and 'opst' is enough to have. 'mhr', 'sc', 'rbp', 'age' and 'thal' have high covariance or dependability on 'a1p2' hence they are key variables. But, 'sc' has high covariance/dependability on 'rbp' and 'age' hence one of the three can be chosen for prediction purpose. Let's choose 'sc'. Remaining 2, 'mhr' and 'thal' are already chosen as they have high correlation with 'a1p2'. Hence, 'thal', 'nmvcf', 'eia', 'mhr', 'opst', 'cpt', and 'sc' could be considered highly significant in the process of heart disease prediction.

ANALYSIS FROM PROBLEM-2:

Following table referring the output table of Problem-2, shows the percentage of combined (train data (70%) + test data (30%)) accuracy achieved for different Machine Learning methods:

Machine Learning Method	Perceptron	Support Vector Machine (kernel-linear)	Decision Tree Learning	Random Forest	K-Nearest Neighbors	Logistic Regression
Combined Accuracy (%)	86	87	92	94	87	87

After trying out different Machine Learning methods to create an effective heart prediction model, Random Forest seems to be the best method based on the prediction percentage shown above. For each Machine Learning method, the best combination of parameters was found by running loops and following trial and error process. For the Random Forest method, number of tress or 'n_estimators' value was set to 11 to achieve the best possible combined accuracy. After analyzing the best accuracy results for all Machine Learning methods, it was found that Random Forest had the highest combined accuracy percentage. Hence it is the best Machine Learning method for heart disease prediction.