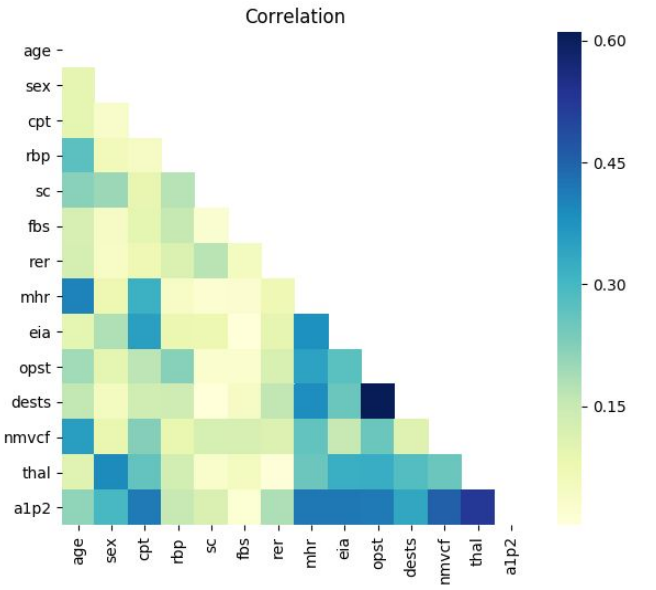
Acme Medical Analysis and Prediction Enterprises

Patient Data Analysis Report and Review of Developed Prediction Models

This report analyzes heart related data of 270 patients. In the given database, a total of 13 parameters are provided and information regarding presence or absence of heart disease is provided. In this analysis, the most important parameters influencing occurrence of heart disease in a patient are found out. Along with this analysis, the report also reviews prediction models (of varying accuracy) developed using the given database, for predicting the presence or absence of heart disease.



Various expected and unexpected insights are observed while analyzing the data. Firstly, a descriptive table of parameters reveal that the average age of patient is approximately 54 years, yet age has a low correlation (0.2123) with the occurrence of heart disease. A similar anti-intuitive case is observed (0.1554) for resting blood pressure (rbp). The most influencing parameters for presence of heart disease is observed to be thal (thalassemia - an inherited blood disorder) with correlation of 0.5250 and (nmvcf) number of major vessels colored by flourosopy having correlation 0.4553. Others are eia (0.4193), mhr (-0.4185), opst (0.4180). A study of correlation matrix of the parameters reveal that opst and dests are largely correlated (0.6097), possibly because both have to do something with ST segment of the ECG signal, however, both have low correlation with occurence of heart disease. Pairplots of five most correlated variables with occurrence of heart disease can be used to observe their inter- and intra-distribution.

Five models have been developed to predict the presence or absence of heart disease using the following modelling algorithms: Perceptron, Logistic Regression, Support Vector Machine, Decision Tree Learning and K-Nearest Neighbor. Perceptron model is the simplest of all in terms of complexity and expectedly, gives the lowest accuracy. Logistic regression, despite being a fairly

simple model, gives a good accuracy because logistic regression model is focused towards binary classification predominantly and the problem in our hand is a binary classification problem (disease absent / disease present). However K nearest neighbors algorithm gives the maximum accuracy for classification (at k = 3). Space Vector Machine (with a linear kernel) and decision tree learning give a similar accuracy of 86%. Thus, K nearest neighbors based model is the best model to use for predicting presence or absence of heart disease.

Model	Test Data Accuracy	Total Data Accuracy
Perceptron	75%	80%
Logistic Regression	84%	87%
Space Vector Machine (SVM)	83%	86%
Decision Tree Learning	79%	86%
K Nearest Neighbors	81%	89%