Analysis and Modelling of Heart Disease

Our analysis includes a statistical study in the analysis and modelling of the given data to predict heart disease checked for the given variables: correlation of each variable, dependent or independent, with all the other variables. Determining which variables are most highly correlated with each other and which are highly correlated with the variable you wish to predict.

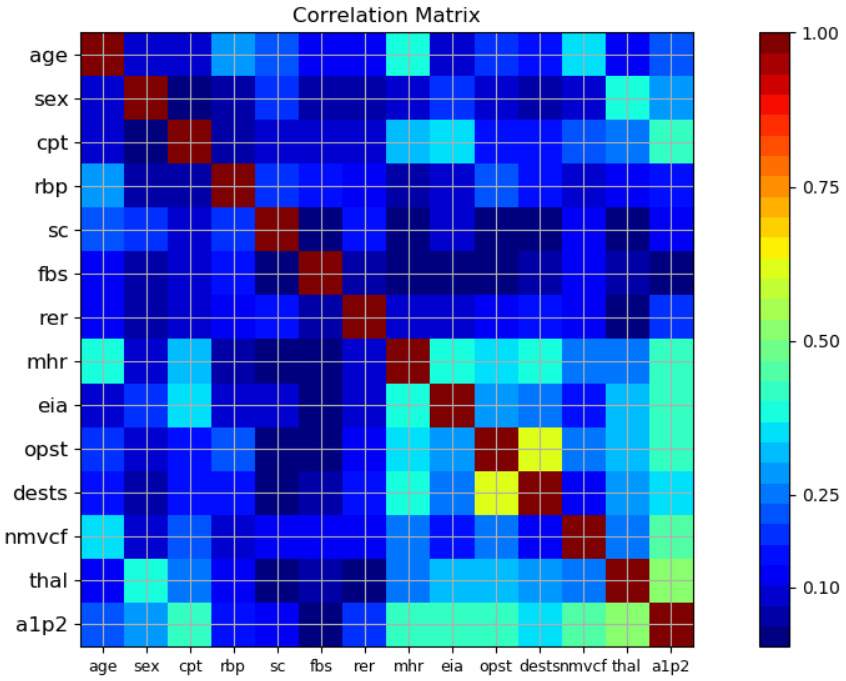
The statistical relationship between two variables is referred to as their correlation. Correlation can tell us how one variable moves with respect to other variable. Though the sign tells us the if the variables are moving in the same direction (+ve) or not (-ve), we are more interested in the correlation value. If the absolute value of the variables is 1, it means the variables share a linear relationship.

These are the variables in the dataset:

Name	Num	Description
age	0	age
sex	1	sex
cpt	2	chest pain type (4 values)
rbp	3	resting blood pressure
sc	4	serum cholestoral in mg/dl
fbs	5	fasting blood sugar > 120 mg/dl
rer	6	resting electrocardiographic results (values 0,1,2)
mhr	7	maximum heart rate achieved
eia	8	exercise induced angina
opst	9	oldpeak = ST depression induced by exercise relative to rest
dests	10	the slope of the peak exercise ST segment
nmvcf	11	number of major vessels (0-3) colored by flourosopy
thal	12	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
a1p2	13	absence of heart disease = 1, presence = 2

Since, in Machine learning, if two values share a very high correlation, they seem to represent the same information. Therefore, pairwise correlation amongst different features must be less, as then they seem to indicate different information. Lesser correlation means that statistically the features are independent, and greater correlation means they might be related to each other. Here, we see that most of the blocks in the covariance matrices and pairwise plots seems to indicate that 13 features are not very related to each other. Only (opst -) and (dests) have correlation value of 0.609. ['thal', 'nmvcf', 'eia', 'mhr', 'opst', 'cpt'] seems to be more correlated [>0.417] to a1p2. Therefore in our further calculation I have included all the features given as this

seems to indicate the best results. Also, leaving any feature out results in loss of data , therefore, I intend to use all the given feature.



Classifier Type	Test Accuracy %	Total Accuracy %
Perceptron	75.308	80.37
Linear Regression	83.95	86.66
Support Vector Machine	82.71	86.29
Decision Tree Classifier	79.01	85.92
K Nearest Neighbors	81.48	88.51

On testing our data from different patients over all the 13 features given, I found the following above results. I tested five different classifier such as perceptron, linear regression, Support Vector Machine, Decision Tree Classifier, K Nearest Neighbors. I also tested the models over different parameters and found that over the given dataset, Linear Regression performs the best though its one of the simpler classifiers . K Nearest Neighbor classifier too performs decently well over the complete dataset. Perceptron being one of the primitive classifiers performs the worst. Single decision tress could be slightly biased as it performs better when the training data is included with the test data as indicated in the total accuracy output. For the logistic regression, we seem to be getting the best results at the values (solver='liblinear', max_iter =1000,C=10,multi_class='ovr'). The tried different solvers such as 'sag', 'saga' and other parameters. The accuracy seems to decrease on changing any of these values. For the SVM , the optimal values were found to be (kernel='linear', C=1, random_state=10) . I tried other kernels too such as 'linear', 'poly', 'rbf', 'sigmoid'. But, accuracy only decreases. In fact an SVM with linear kernel is very similar to logistic regression. Similarly, I tested different values for other classifiers as well and narrowed on the following best values. Therefore, I am convinced that the model of K Nearest Neighbors is the best performing and can be used for future use. Total Accuracy is also not too high showing that the models are over fit, which is inactive of a better model.