HEART DISEASE PREDICTION ANALYSIS

Heart disease is considered as one of the major causes of death throughout the world. An automated system in medical diagnosis would enhance medical efficiency and also reduce costs. We will design a system that can efficiently discover the rules to predict the risk level of patients based on the given parameters about their health. The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases. In this analysis, I will also use heart disease dataset to explore the highest important features that leads to heart disease.

Data At Hand:

I'll be working with the Cleveland Clinic Heart Disease dataset which contains 13 variables related to 303 patient diagnostics and one outcome variable indicating the presence or absence of heart disease.

This dataset is all about heart disease. It contains:
1)age, 2)sex, 3)chest pain type (4 values) , 4)resting blood pressure , 5)serum cholesterol in mg/dl , 6)Fasting blood sugar > 120 mg/dl , 7)resting electrocardiographic results (values 0,1,2) , 8)maximum heart rate achieved , 9)exercise-induced angina , 10)oldpeak = ST depression induced by exercise relative to rest , 11)the slope of the peak exercise ST segment ,12)number of major vessels (0-3) coloured by fluoroscopy ,13)thal (thalassemia): 0 = normal; 1 = fixed defect; 2 = reversible defect ,14)Diagnosis of Heart Disease :Indicates whether subject is suffering from heart disease or not.

Analysis:
Analysis methods includes Statistical Analysis, Feature importance/selection, Logistic regression modelling .

Here, in my data 'Diagnosis of Heart Disease' is my response(dependent) variable and it is qualitative .And all other variables are independent variables.

It is evident from my data that 164 patients out of 303 have not been diagnosed here disease and 139 patients have diagnosed with heart disease i.e. almost 46% of the total individuals are suffering from heart disease.

Here, I have my data on four different types of chest pain types and have found out that 8% of patients with typical angina chest pain,16% of patients with atypical angina,28% with non angina chest pain and 48% with asymptotic angina chest pain .

From our dataset I can say that 55.3% of males are diagnosed with heart disease where is only 26% of females are diagnosed with heart disease. Hence, males are more diagnosed with heart disease than females.

The correlation between ‘Resting blood pressure’ and ‘Age’ is high as Resting blood pressure tends to increase with age .Thus, it's essential for individuals, especially as they age, to monitor their blood pressure regularly and adopt healthy habits to help manage and prevent hypertension and its associated risks.

Also, it is observed that ‘Maximum heart rate achieved’ and ‘The slope of the peak exercise ST segment’ have positive correlation. The high correlation ‘Maximum heart rate achieved’ and ‘the slope of the peak exercise ST segment’ likely arises from the fact that both are influenced by similar physiological mechanisms related to cardiovascular fitness and function.

Now ,we know that a higher mean decrease in Gini suggests that the feature is more important for making decisions in the Random Forest model. Therefore, from my analysis it can be clearly observed that :
1) Thalassemia is the highest to cause heart disease for males and 2) Maximum heart rate achieved is the highest cause factor to cause heart disease for females. As in case of males Thalassemia has the highest mean decrease in Gini value of 19.88 and in case of females Maximum heart rate achieved has the highest mean decrease in Gini value of 51.95.

Then, consider a hypothesis testing to test the association between chest pain and heart disease diagnosis. As, chest pain serves as a key clinical indicator for evaluating and diagnosing heart disease. Healthcare providers often prioritize investigating chest pain symptoms to rule out or confirm cardiac causes and initiate appropriate management and treatment strategies to mitigate the risks associated with cardiovascular conditions.

For logistic regression , I have splitted the data into 75% training data and 25% testing data. And found that chest pain has the lowest p value

hence there is a strong association between chest pain and heart disease diagnosis. And this accepts our above null hypothesis.

Here, accuracy based on training data is 85% and accuracy of based on testing data is 86%.