Project Report – Greg Martirosov, Marcos Sanchez, Herschelle Bumrah

Abstract:

Upon choosing our dataset we first began to explore the different features and see how they would affect the target variable. To do this, we created various plots and created an rpart class model to check what features correlate with the target variable. Checking specific models like frequency of heart disease by age we saw a traditional bell curve shape. This is expected but tells us age is a good general feature to measure and include in our models. Once we had an understanding of the more important features we began to build our rpart model and naive Bayes model. Both the rpart and naive Bayes model resulted in around an 80% accuracy and upon reading the plotted model we drew multiple conclusions. An accuracy rating of 80% gave us enough confidence to check our plotted tree again to see what features impact a heart disease diagnosis.

Dataset: https://www.kaggle.com/datasets/mexwell/heart-disease-dataset

Results:

From our plotted tree we reviewed each branch to see what features were correlated with a diagnosis. The highest impact feature was chest pain rated greater than 4. Resting heart rate was the second leading correlated feature to being diagnosed with heart disease. The third highest correlated feature to being diagnosed with heart disease is being a man.

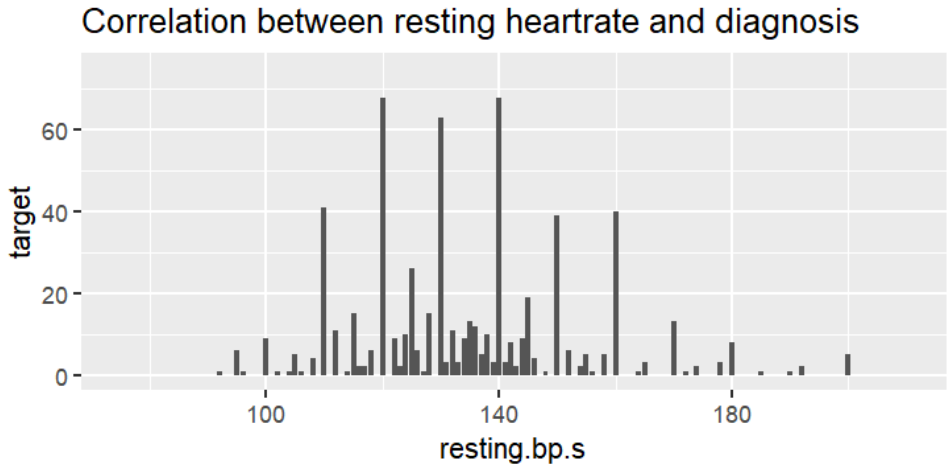
Discussion:

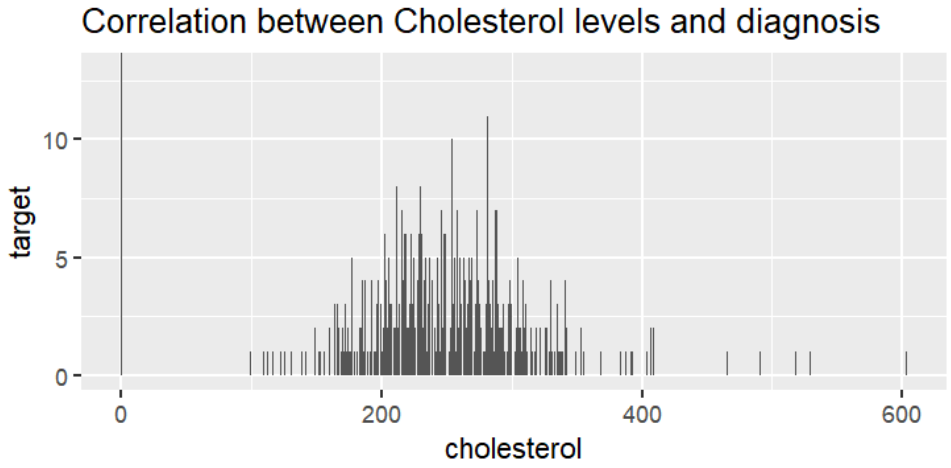
From the most highly correlated features to the target variable we can tell that high chest pain with a high resting heart rate is a sign that there is potential for heart disease to occur. This is a given, but more interestingly this dataset tells us that men are more likely to have heart disease. However, even if you are a male with both the symptoms higher cholesterol levels put you at a higher risk. Exercise-induced angina puts you at a higher risk of heart disease,

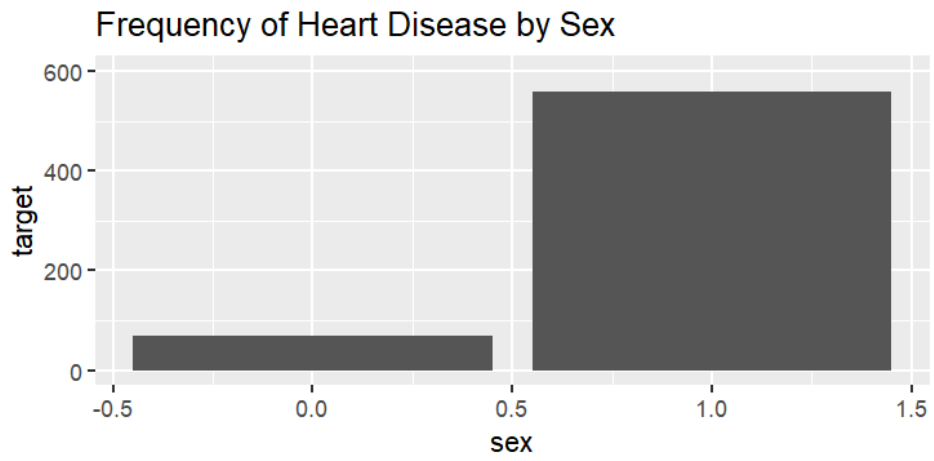
but at such a low level it could be misinterpreted as a guaranteed correlation possibly being uninvolved in a diagnosis.

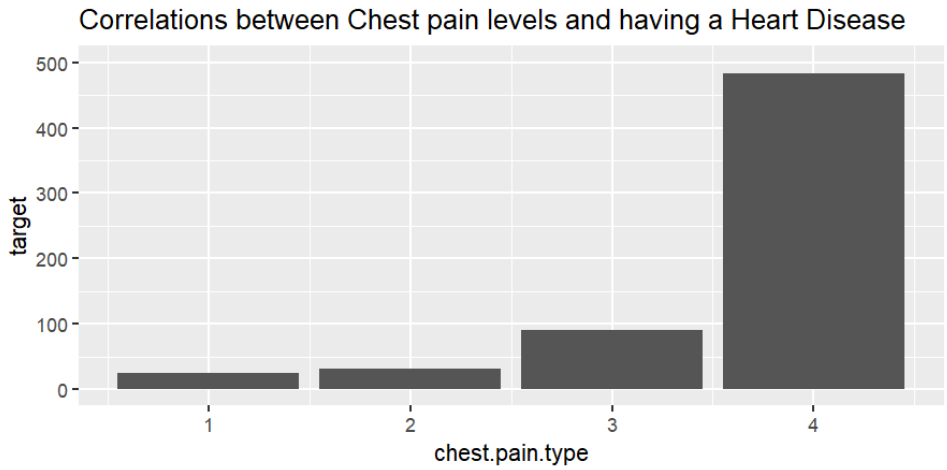
Use of AI:
Used AI to bugfix for naive Bayes function because it wanted the target column as a factor.

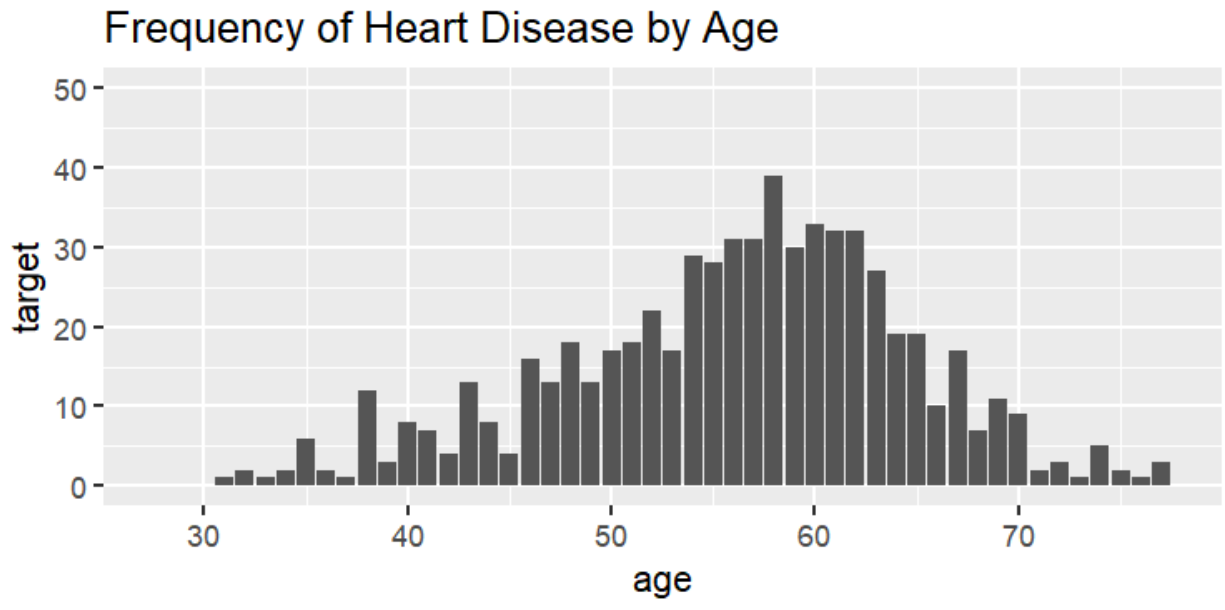
Plots used:






Frequency of Heart Disease by Sex
target
600
400
200
0
-0.5
0.0
0.5
1.0
1.5
sex


Correlations between Chest pain levels and having a Heart Disease
target
500
400
300
200
100
0
1
2
3
4
chest.pain.type


Frequency of Heart Disease by Age
target
50
40
30
20
10
0
30
40
50
60
70
age