Aditya Ranade and Austin Wang
Ghosh
CSCI183 - Spring 2022
17 May 2022

HW3 - Report

Project Brief:

In this homework assignment, we were tasked with analyzing various health data points of individuals to see if they had any relation or could predict if an individual given a certain set of features would have a heart attack or not.

Variables in Dataset:

To clear up the some of the variable names you will see used throughout the rest of this report, below is a list of variables, what they mean, their type, which ones we dropped, and why:

Variable Name: age
1) age - NUMERICAL

Variable Name: sex
2) sex - CATEGORICAL
(dropped - only binary values for male and female)

Variable Name: cp
3) chest pain type (4 values) - CATEGORICAL
(dropped - only 4 values showing type of chest pain)

Variable Name: trestbps
4) resting blood pressure - NUMERICAL

Variable Name: chol
5) serum cholesterol in mg/dl - NUMERICAL

Variable Name: fbs
6) fasting blood sugar > 120 mg/dl - CATEGORICAL
(dropped - only binary values 0/1 which show whether fasting blood sugar is over 120 mg/dl)

Variable Name: restecg
7) resting electrocardiographic results (values 0,1,2) - CATEGORICAL

(dropped - only values 0, 1, and 2 which show results)

Variable Name: thalach
8) maximum heart rate achieved - NUMERICAL

Variable Name: exang
9) exercise-induced angina - CATEGORICAL
(dropped - only binary values 0/1 which show whether there is or isn't exercise-induced angina)

Variable Name: oldpeak
10) oldpeak = ST depression induced by exercise relative to rest - NUMERICAL

Variable Name: slope
11) the slope of the peak exercise ST segment - CATEGORICAL
(dropped - slope values are only 0, 1, and 2 which show slope of certain segment)

Variable Name: ca
12) number of major vessels (0-3) colored by fluoroscopy - CATEGORICAL
(dropped - certain number of vessels only which becomes categorized)

Variable Name: thal
13) thal: 0 = normal; 1 = fixed defect; 2 = reversible defect - CATEGORICAL
(dropped - values only tell us whether it is normal, has a fixed defect, or a reversible defect)

Variable Name: target
14) target: 0= less chance of heart attack 1= more chance of heart attack - CATEGORICAL

Procedure:

1. Put data into a dataframe using the pandas library.
2. Drop the non-numerical values that we cannot use for analysis (for this homework, we were only asked to use numerical values, so we dropped those that were non-numerical immediately. However, in a real-world example, we would analyze those variables as well to understand if they have any effect on if an individual would have a heart attack or not).
3. Using the features identified as numerical, create all the permutations of plots using the matplotlib library.
4. Figure out the plots with numerical features that can be used for good classification models.
5. Split the dataset into 70% train and 30% test.

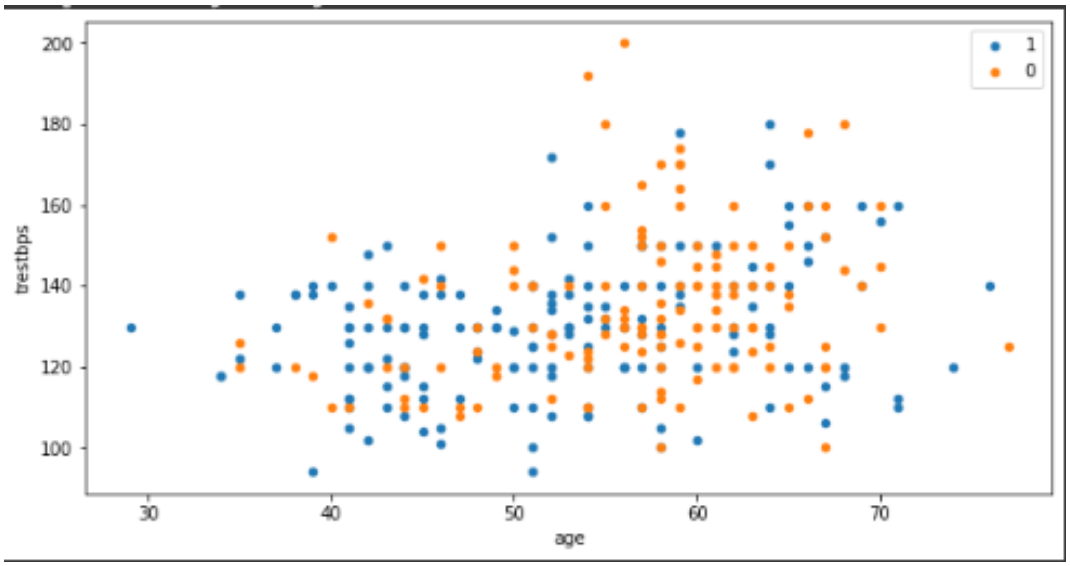
6. Implement the Logistic Regression classification algorithm.
7. Plot the linear boundary of classification for the plots that can be used for classification (logistic regression).
8. Use the evaluation metrics (precision, recall, accuracy, F1 score) to evaluate how well the models we made have performed for the given dataset.

Results:

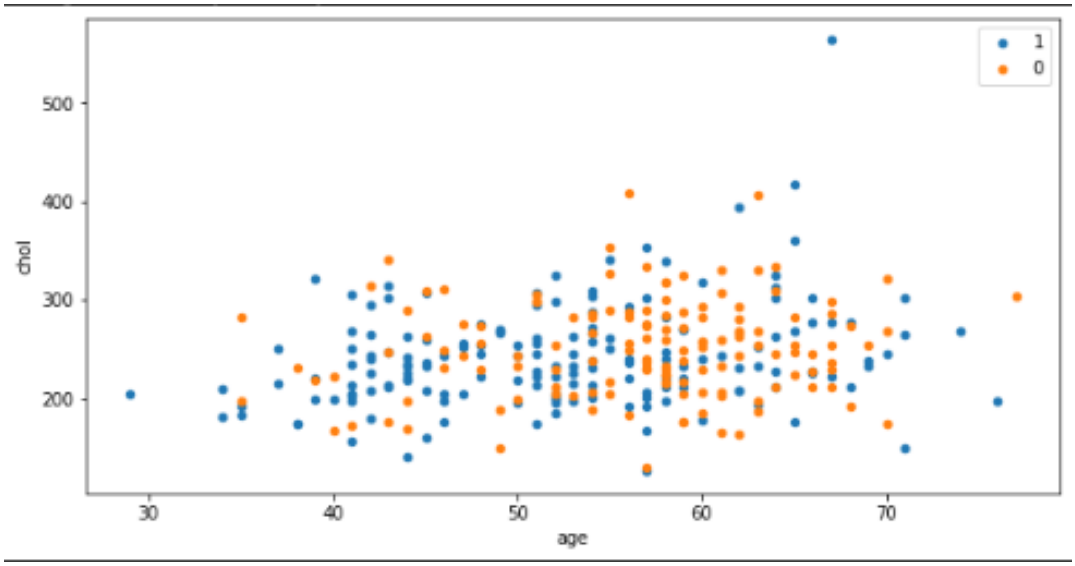
After dropping all the non-numerical values in the dataset, the values we came up with that we could use for plotting were: age, trestbps, chol, thalach, oldpeak, and target. Using these variables, we were able to come up with ten different graphs, four of which (that you will see later) we determined would be good for implementing classification and doing logistic regression.

Below are the 10 various graphs that we plotted using the matplotlib library:

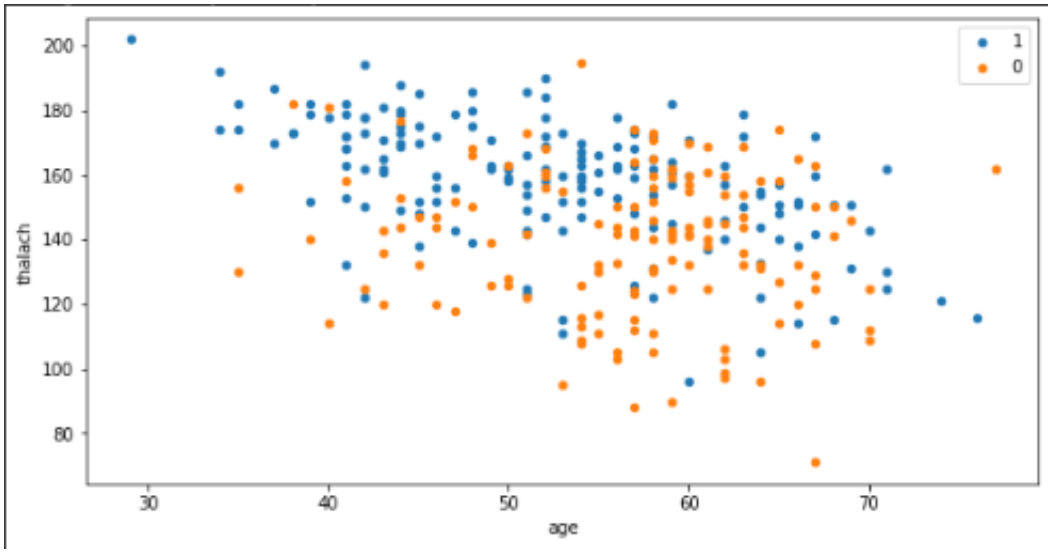
Graph 1: age x trestbps



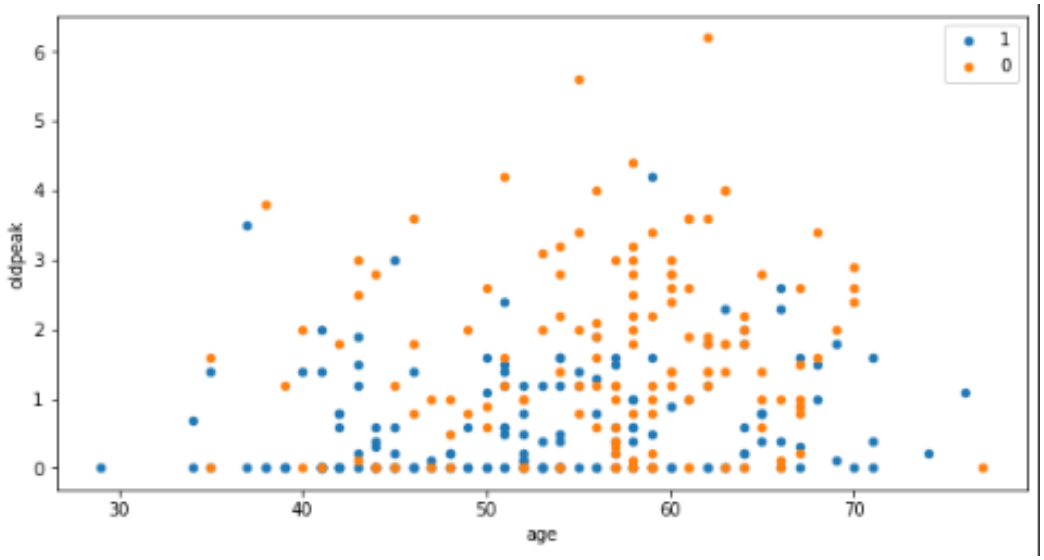
Graph 2: age x chol



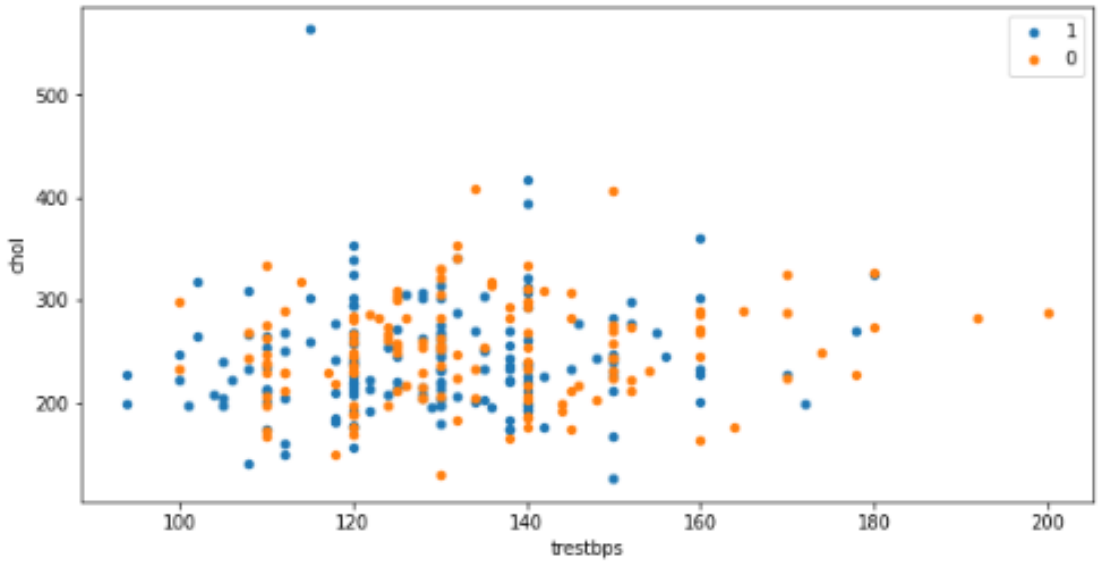
Graph 3: age x thalach



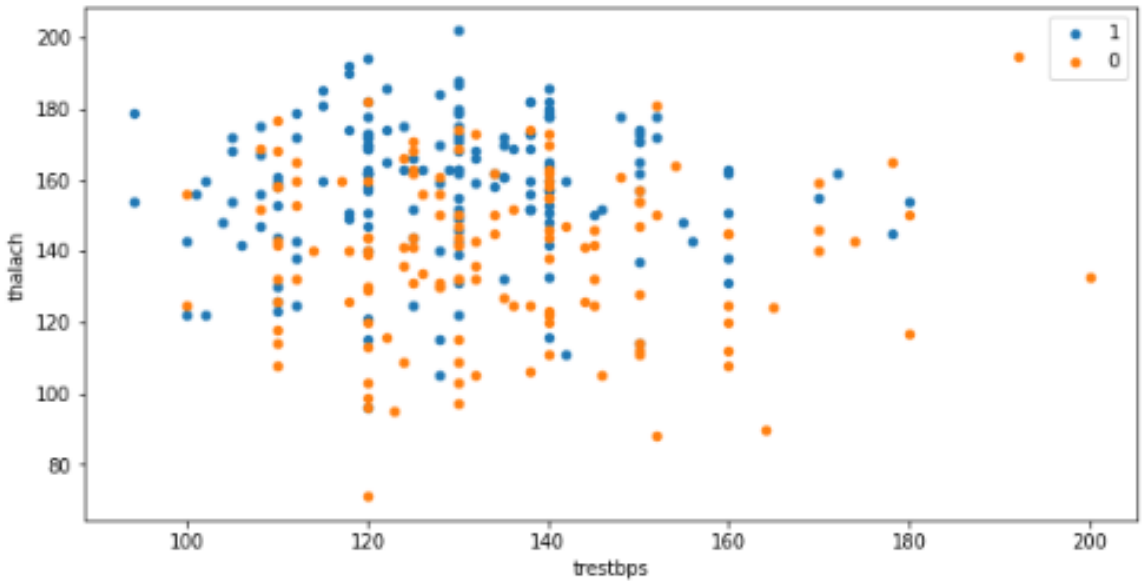
Graph 4: age x oldpeak



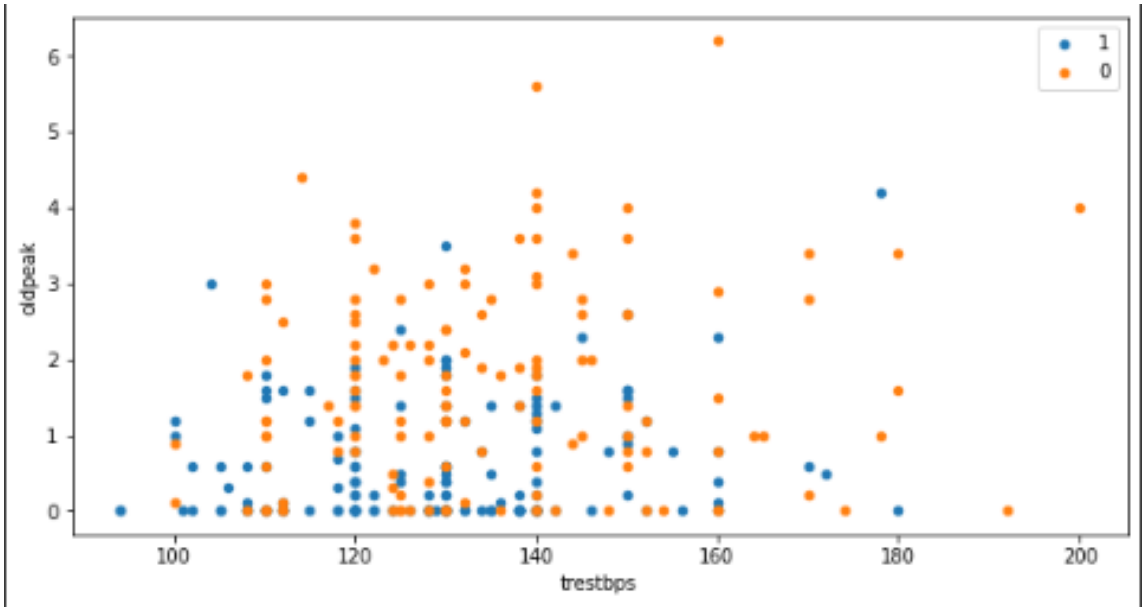
Graph 5: trestbps x chol



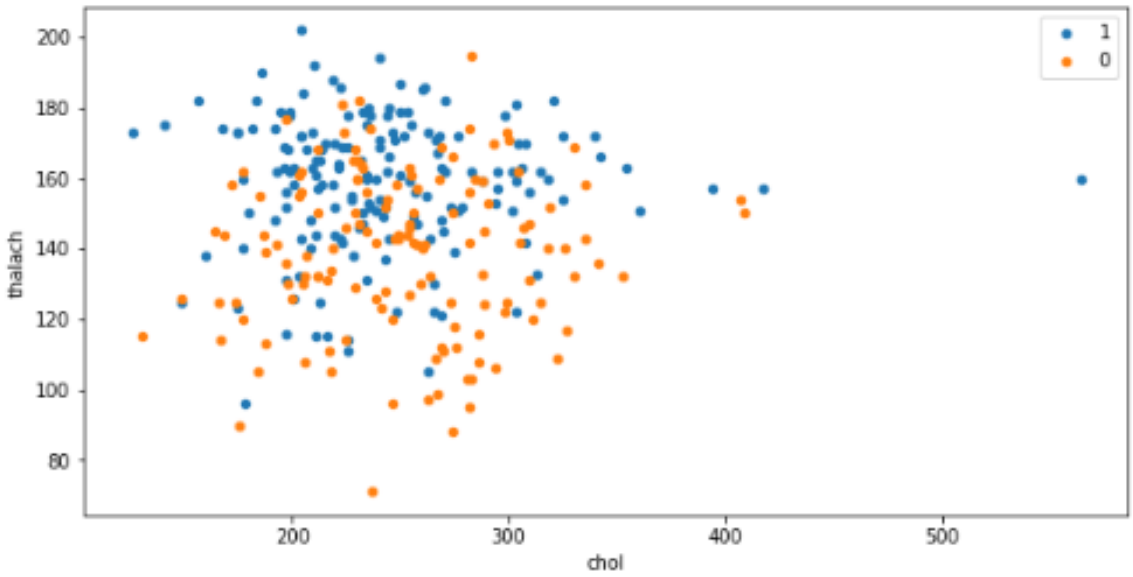
Graph 6: trestbps x thalach



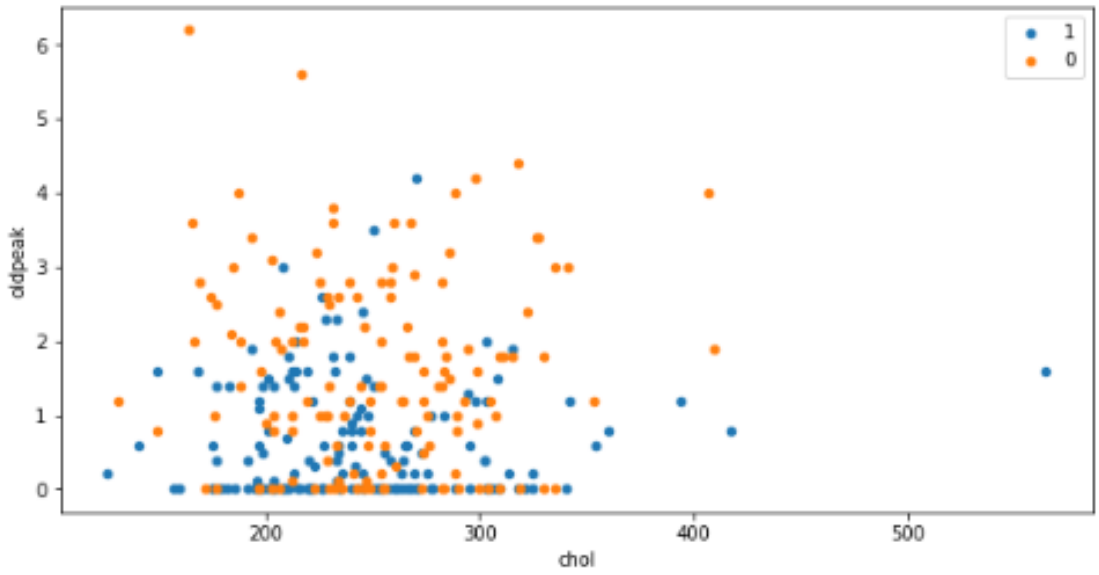
Graph 7: trestbps x oldpeak



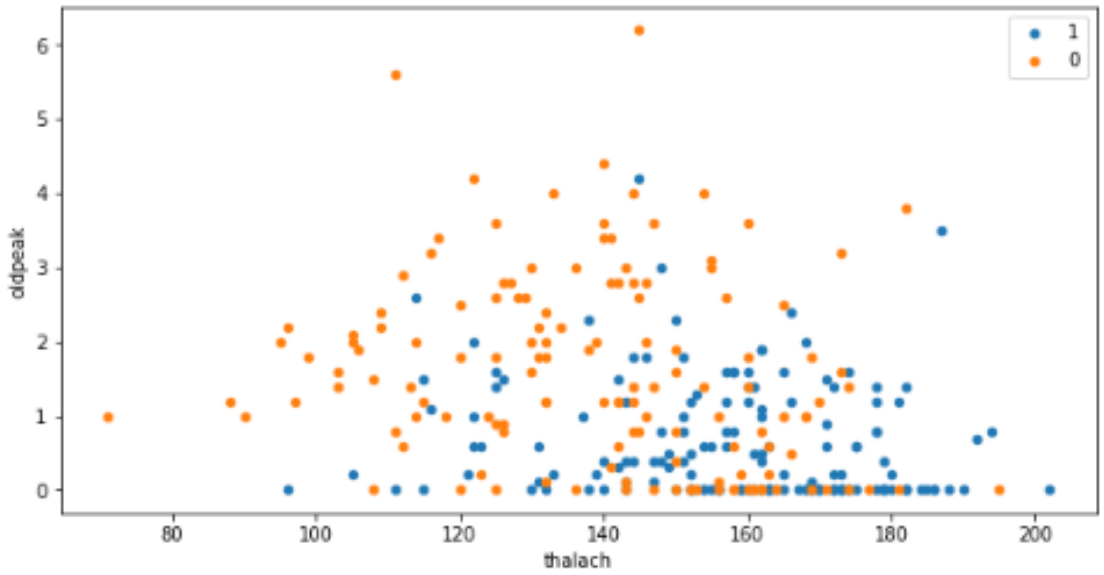
Graph 8: chol x thalach



Graph 9: chol x oldpeak



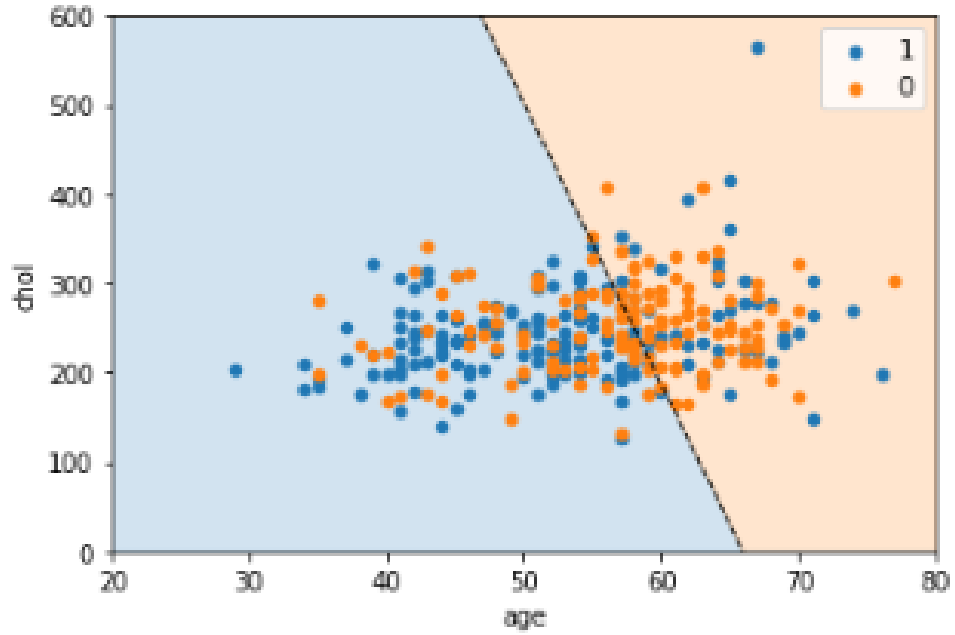
Graph 10: thalach x oldpeak



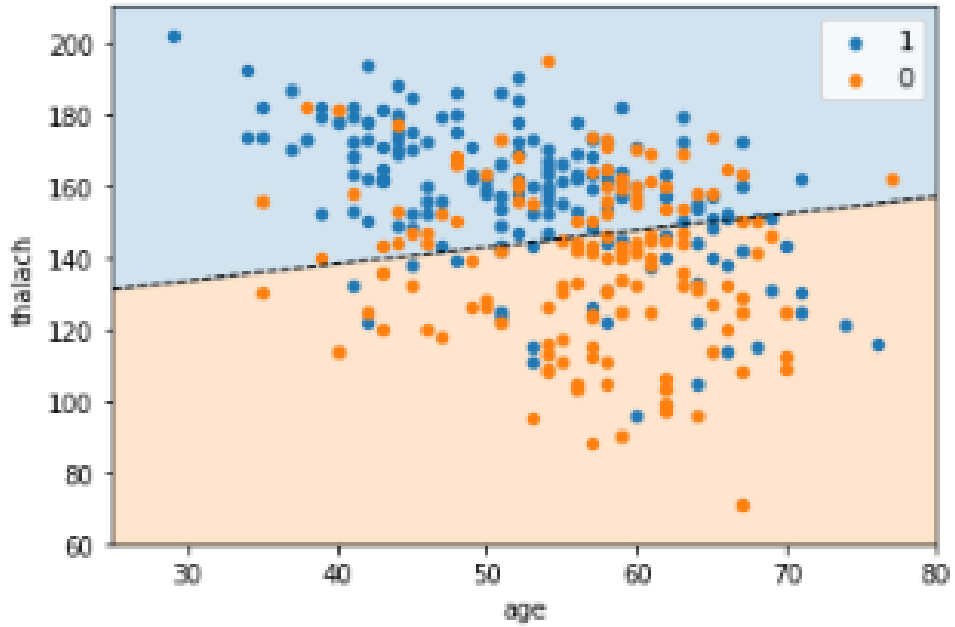
From these 10 graphs, we figured out that there were 4 graphs that we would possibly analyze further and use logistic regression to get a classification boundary. As far as the classification algorithm was concerned, we used the one from the sci-kit learn library with a hinge-loss function to create a decision boundary for the target variable of whether a heart attack would occur or not given a set of two particular features. We used 70% of the given data to train the model and the remaining 30% to test. The four graphs we chose to further analyze were: age x chol, age x thalach, chol x thalach, and thalach x oldpeak.

Below are the 4 graphs that we chose with logistic regression implemented, and with the linear decision boundary displayed:

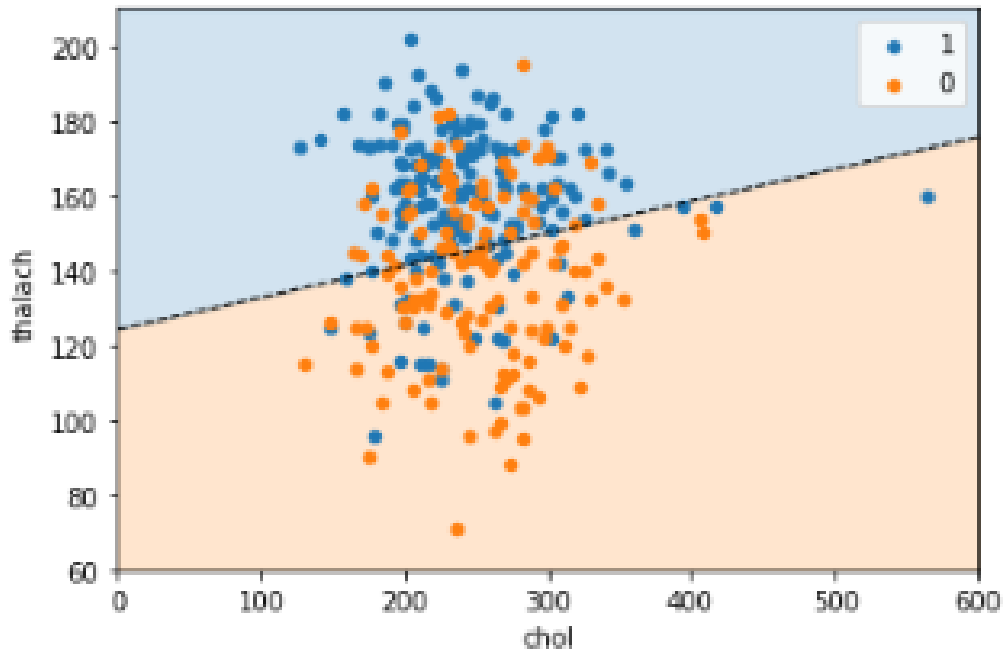
Graph 11: age x chol with logistic regression and linear decision boundary



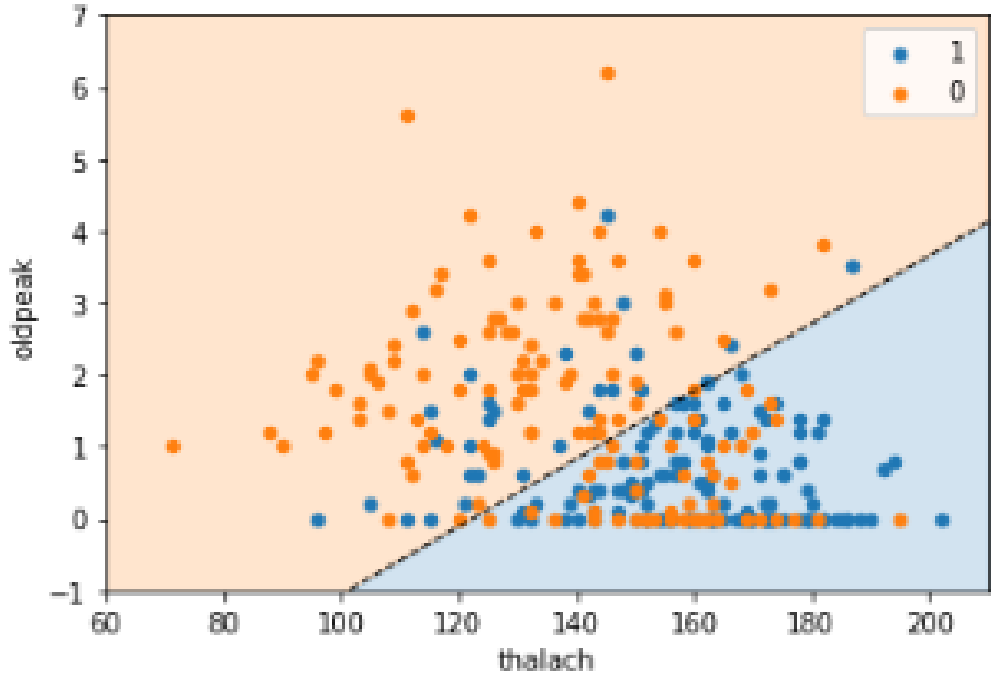
Graph 12: age x thalach with logistic regression and linear decision boundary



Graph 13: chol x thalach with logistic regression and linear decision boundary



Graph 14: thalach x oldpeak with logistic regression and linear decision boundary



After successfully plotting these four graphs with the linear decision boundary, we calculated the values of Precision, Recall, Accuracy, and the F1 Score to see how well these models were performing against the given data. We came up with the following values for each plot we performed logistic regression on:

Table 1: Values of Precision, Recall, Accuracy, and the F1 Score for each of the plots for which logistic regression was performed

	Feature	Precision	Recall	Accuracy	F1Score
0	age x chol	0.8936	0.6087	0.6483	0.7241
1	age x thalach	0.8297	0.6610	0.6923	0.7358
2	chol x thalach	0.8723	0.6833	0.7253	0.7664
3	thalach x oldpeak	0.8297	0.6964	0.7252	0.7572

From the values in Table 1, we can see that the chol x thalach has the highest F1 Score, meaning that it is the best model.