Heart Disease Dataset: Analysis and Report

Zehadi Alam, Edric Owusu, & Cam Piacentini

University of Georgia

CSCI 3360: Data Science I 37808

Dr. Hao Peng

December 3, 2021

Introduction

This Heart Disease Dataset is a subset of a larger database that consists of only fourteen of the seventy-six total attributes whose purpose is to determine whether a given patient has heart disease. The presence of heart disease is represented as a categorical binary attribute called "target" in which 1 indicates those who have heart disease while 0 indicates otherwise (Heart Disease UCI, 2018). Whether one has heart disease is determined by how much their blood vessel narrowed in diameter; those whose diameters narrowed by more than 50% were classified as having heart disease (Heart Disease Data Set, n.d.). There are 303 samples in total.

Table 1

Head of Pandas Dataframe for Heart Disease Dataset

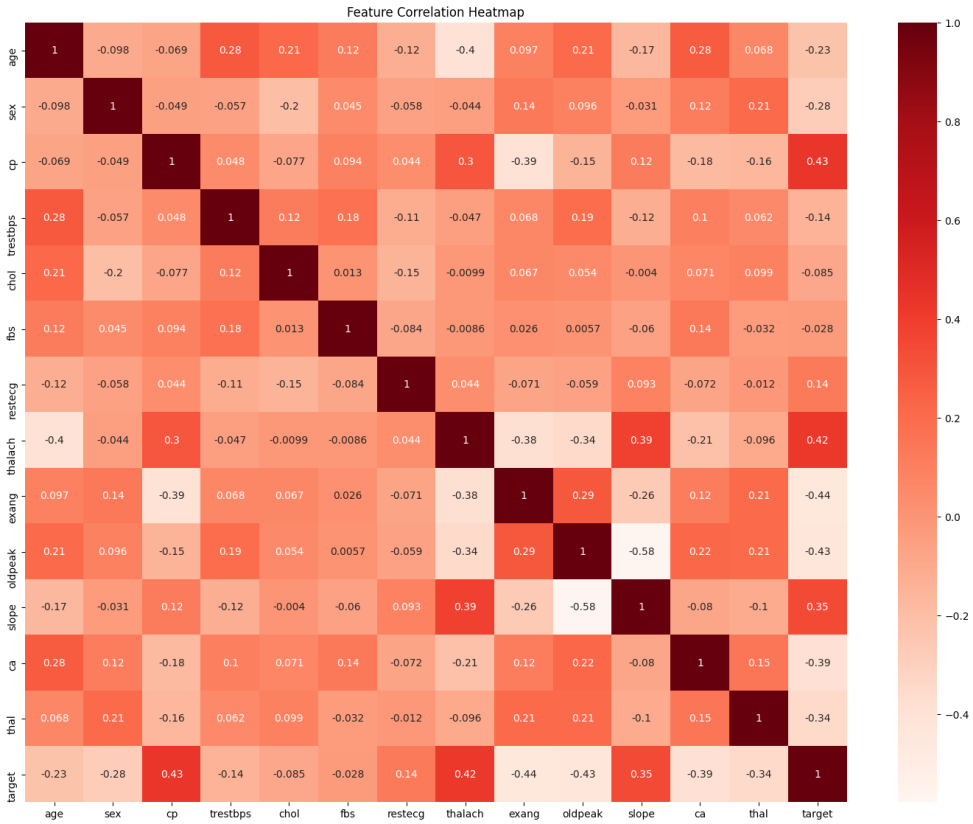
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

The input attributes in Table 1 from left to right refer to the follow: age (in years), sex (0 = female, 1 = male), chest pain (scale of 0-3), resting blood pressure, serum cholesterol (in mg/dl), fasting blood sugar (1 if > 120 mg/dl), resting electrocardiographic results (values of 0-2), maximum heart rate achieved, exercise induced angina (1 if true), ST Depression induced by exercise relative to rest, the slope of peak exercise ST segment, the number of blood vessels colored by fluoroscopy, and thalassemia (scale of 0-3) (Heart Disease UCI, 2018). For the sake of efficiency, these attributes will be referred to by their abbreviations for the remainder of this report.

Data Exploration

Figure 1

Feature Correlation Heatmap



Note. Correlations between all possible pairs of features are shown. Categorical features are treated as numerical features that can either have an integer value of 0 or 1.

Correlation with the response variable (target) was highest for cp, thalach, and slope with values of 0.43, 0.42, and 0.35 respectively. From this, we can conclude that these attributes are the highest indicators for whether a patient has heart disease. It should also be noted that thalach specifically also has high correlations with slope at 0.39 and cp at 0.3, indicating some sort of relationship between the two pairs of attributes. The attributes exang and oldpeak have a relatively high correlation value of 0.29, which yet again hints at some of sort of relationship between these two features which may prove useful when determining the presence of heart disease in a patient.

Table 2

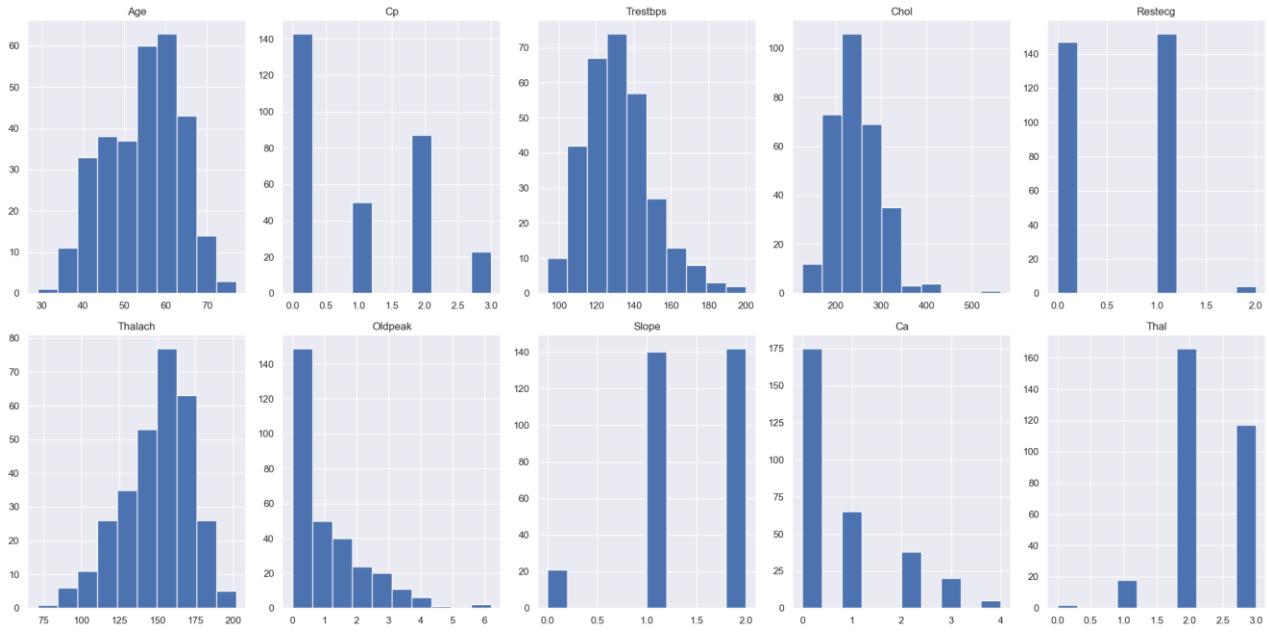
Summary Statistics for Numerical Attributes

Attribure	Minimum Value	Maximum Value	Median	Mean	Standard Deviation
age	29	77	55	54.36633663	9.08210099
cp	0	3	1	0.9669967	1.032052489
trestbps	94	200	130	131.6237624	17.53814281
chol	126	564	240	246.2640264	51.83075099
restecg	0	2	1	0.528052805	0.525859596
thalach	71	202	153	149.6468647	22.90516111
oldpeak	0	6.2	0.8	1.03960396	1.161075022
slope	0	2	1	1.399339934	0.616226145
ca	0	4	0	0.729372937	1.022606365
thal	0	3	2	2.313531353	0.612276507

Note. The minimum, maximum, median, mean, and standard deviation are given for all numerical attributes in the dataset. This table gives an indication as to what the overall distribution for each of these attributes are across the entire dataset.

Figure 2

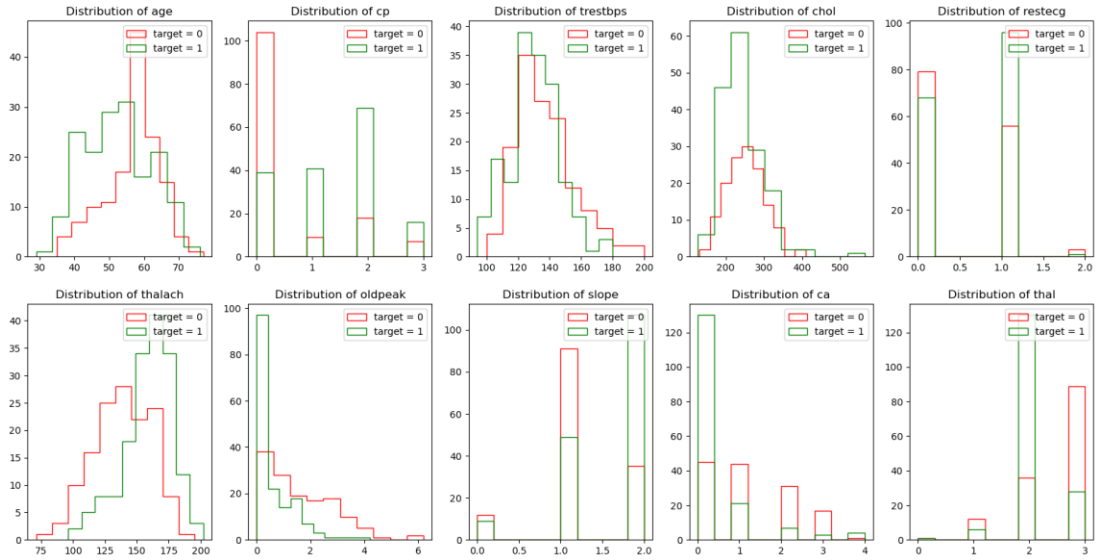
Histograms for Numerical Attributes



Note. These histograms are a visual representation of Table 2.

Figure 3

Grouped Histograms for Numerical Attributes

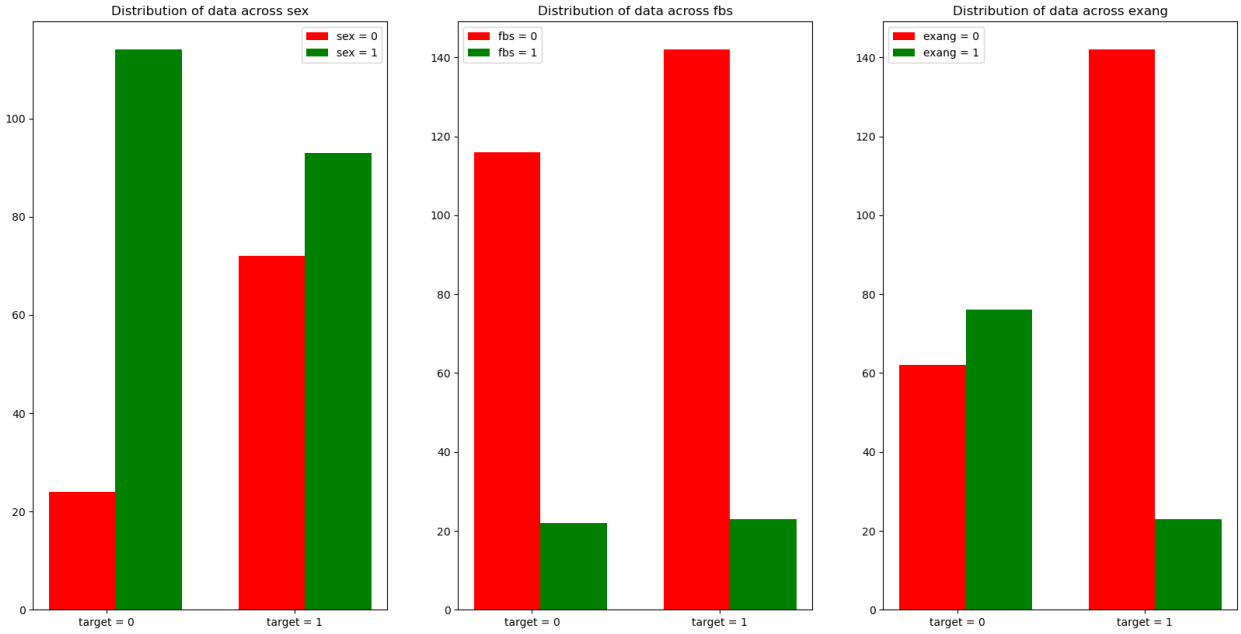


Note. These are histograms grouped by the response attribute target. They visually indicate how each numerical attribute is distributed for each of the two possible response values.

Surprisingly, the range for age is wider when target = 1. This runs counter to the general understanding that those with heart disease tend to be older. However, the distributions for cp, chol, and thalach map out as expected; the values for these attributes are higher when target is 1.

Figure 4

Grouped Bar Graphs for Categorical Attributes



Note. These graphs display the distribution of the three categorical attributes across the two values for target.

Interestingly, the ratio of those without heart disease and those with heart disease is lower when sex is 0 (female), which goes against the notion that those with heart disease are most likely to be male. Similarly, the ratio of those with heart disease and those with it is lower when exang is 0 even though one would expect there to be more instances of exercise-induced angina in the

presence of heart disease. Both anomalies can be explained by the potential of bias. The study inherently calls for those who are already at risk for heart disease and if the sample was more reflective of the general population, the distribution of these attributes among the target would most likely more closely align with the general trends.

Logistic Regression, KNN, and Clustering machine learning models were used to better asses the data.

Logistic Regression

Using the scikit-learn package, we trained a logistic regression model that would predict whether a patient is will most likely have heart disease or not. This model was trained on 75% of all the data, with elements being selected at random. The model has an intercept of 4.1867127, and its coefficients are -0.0079878, -1.37223233, 0.80185237, -0.0204009, -0.00138602, 0.06924716, 0.68170084, 0.01236913, -0.75512184, -0.87633602 , 0.50294952, -0.78640781, and -0.70703948 for their respective attributes in the order of where they are placed in the dataframe.

Table 3

Classification Report for Logistic Regression Model on Training Set

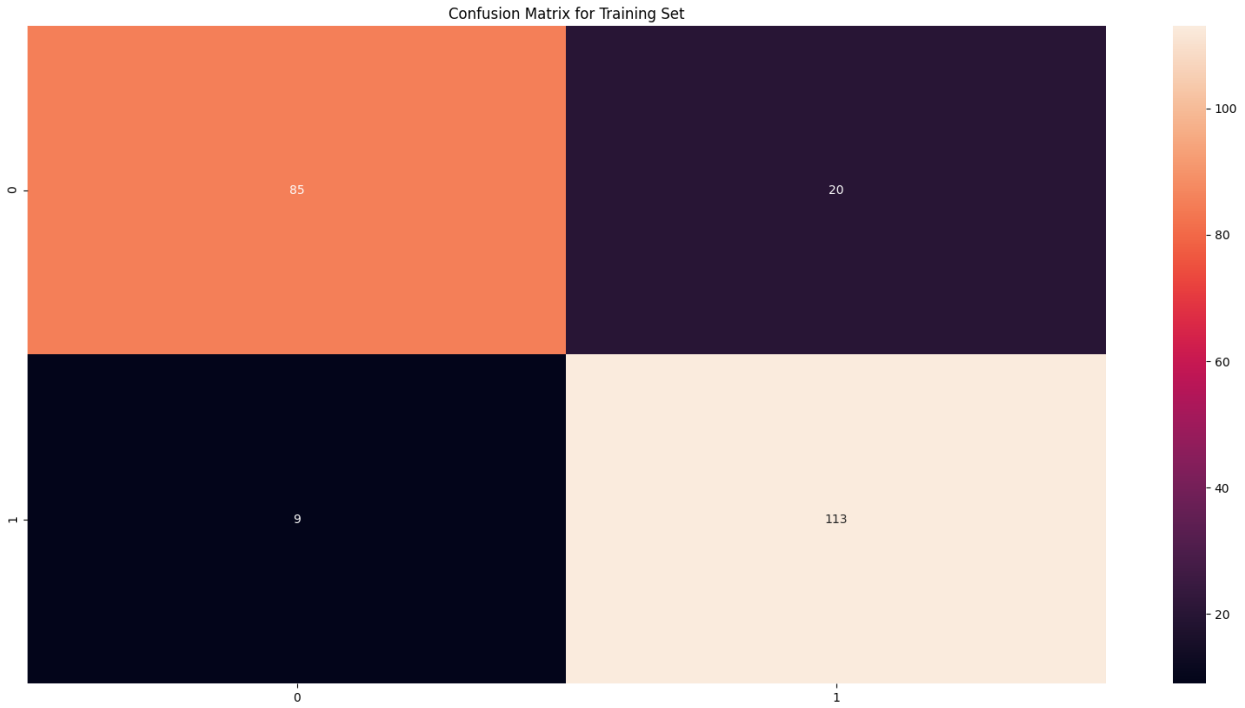
	precision	recall	f1-score	support
0	0.90	0.81	0.85	105
1	0.85	0.93	0.89	122
accuracy			0.87	227
macro avg	0.88	0.87	0.87	227
weighted avg	0.87	0.87	0.87	227

Note. This is the classification matrix that shows how the model performs when ran on the original training set.

The accuracy is relatively high at 87%. As this is a medical dataset in which a value of 1 represents the presence of heart disease, it is also important to look at the recall value for target values of 1. It is 93%, a relatively high recall that indicates that 93% of all positive cases are correctly identified. These numbers are not the best metric for determining how good the model is since the model is being ran on the original testing data, meaning there is most likely some overfitting.

Figure 5

Confusion Matrix for Logistic Regression Model on Training Set



Note. Take note of how 113 of the 122 total positive cases are correctly identified as true positives by the model.

Table 4

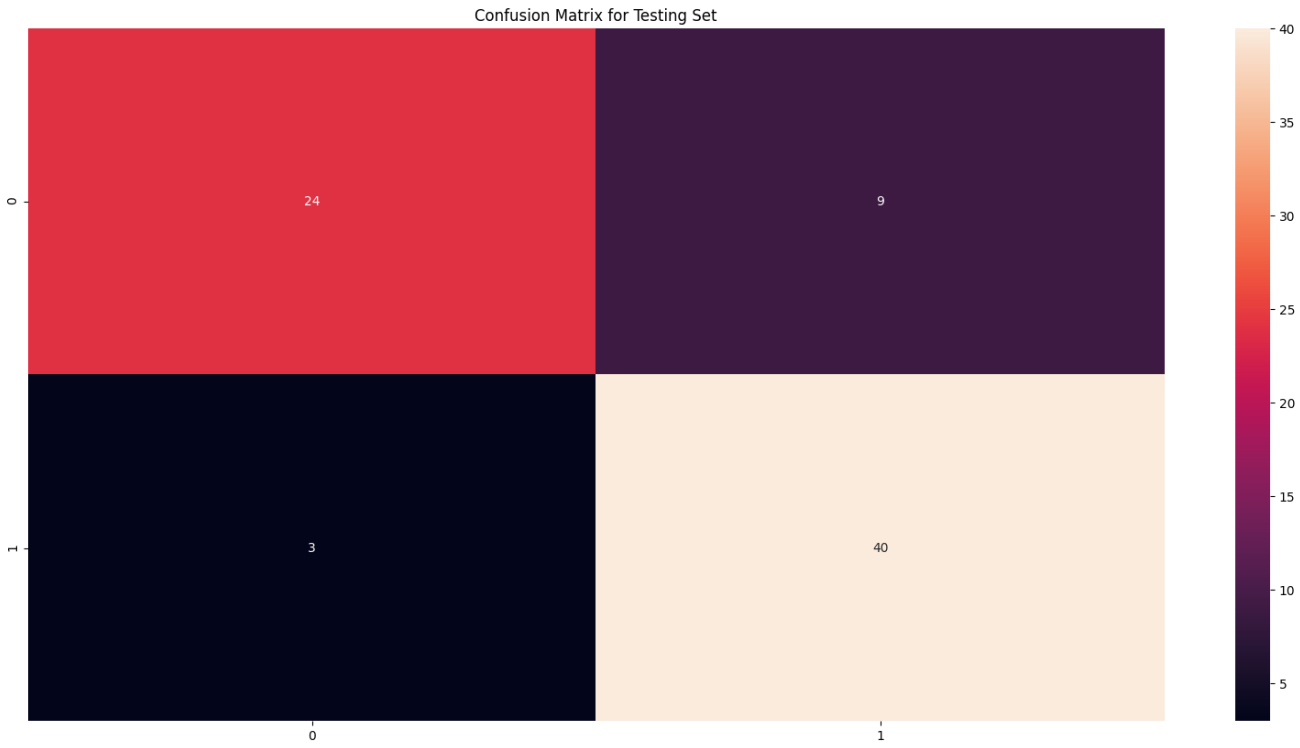
Classification Report for Logistic Regression Model on Testing Set

	precision	recall	f1-score	support
0	0.89	0.73	0.80	33
1	0.82	0.93	0.87	43
accuracy			0.84	76
macro avg	0.85	0.83	0.83	76
weighted avg	0.85	0.84	0.84	76

Note. The classification matrix that shows how the model performed on the testing set, which is a better metric for how good the model is. There is nonsignificant drop in accuracy from 87% to 84%, suggesting that the model is good at predicting whether a patient has heart disease. When target is 1, recall is 93%, which is identical to the training set. It can be concluded that this model is very good at detecting heart disease.

Figure 6

Confusion Matrix for Logistic Regression Model on Testing Set



Note. 40 of all 43 cases of the total positive cases are correctly identified as true positives by the model.

K-Nearest Neighbors (KNN)

Using the `scikit-learn` module, we trained a KNN model to predict whether an individual has heart disease or not. To ensure that the KNN model could aptly be applied, we ensured that the data underwent preprocessing by means of feature scaling. Feature scaling refers to bringing the features in the data to the same scale through normalization. In this context, the magnitudes of some of the features must be normalized, so that the distance is not skewed towards a particular feature. We used the `StandardScaler` from the `sklearn` module to

address this. The way it works is that for each feature, the mean is subtracted, and the result is divided by the standard deviation.

After the data preprocessing through feature scaling, the KNN model was applied to the training set using K values of 1, 3, 5, 7, 9, and 50. The evaluation metrics of this are as follows:

Figure 8

Classification Reports for Training Sets

K = 1

	precision	recall	f1-score	support
0	1.00	1.00	1.00	105
1	1.00	1.00	1.00	122
accuracy			1.00	227
macro avg	1.00	1.00	1.00	227
weighted avg	1.00	1.00	1.00	227

```
[[105   0]
 [  0 122]]
```

K = 3

	precision	recall	f1-score	support
0	0.86	0.85	0.85	105
1	0.87	0.88	0.87	122
accuracy			0.86	227
macro avg	0.86	0.86	0.86	227
weighted avg	0.86	0.86	0.86	227

```
[[ 89  16]
 [ 15 107]]
```

K = 5

	precision	recall	f1-score	support
0	0.85	0.83	0.84	105
1	0.86	0.88	0.87	122
accuracy			0.85	227
macro avg	0.85	0.85	0.85	227
weighted avg	0.85	0.85	0.85	227

```
[[ 87  18]
 [ 15 107]]
```

K = 7

	precision	recall	f1-score	support
0	0.87	0.81	0.84	105
1	0.84	0.89	0.87	122
accuracy			0.85	227
macro avg	0.86	0.85	0.85	227
weighted avg	0.86	0.85	0.85	227

```
[[ 85  20]
 [ 13 109]]
```

K = 9

	precision	recall	f1-score	support
0	0.89	0.76	0.82	105
1	0.82	0.92	0.86	122
accuracy			0.85	227
macro avg	0.85	0.84	0.84	227
weighted avg	0.85	0.85	0.84	227

```
[[ 80  25]
 [ 10 112]]
```

K = 50

	precision	recall	f1-score	support
0	0.89	0.70	0.79	105
1	0.78	0.93	0.85	122
accuracy			0.82	227
macro avg	0.84	0.82	0.82	227
weighted avg	0.83	0.82	0.82	227

```
[[ 74  31]
 [  9 113]]
```

Note. There was perfect accuracy when K = 1. The lowest accuracy (0.82) was when K = 50. Meanwhile, recall when target is 1 was also perfect for L=1 and lowest (0.88) when K = 7 or 9.

Figure 9

Classification Reports for Testing Sets

K = 1

	precision	recall	f1-score	support
0	0.73	0.82	0.77	33
1	0.85	0.77	0.80	43
accuracy			0.79	76
macro avg	0.79	0.79	0.79	76
weighted avg	0.80	0.79	0.79	76

```
[[27  6]
 [10 33]]
```

K = 3

	precision	recall	f1-score	support
0	0.82	0.82	0.82	33
1	0.86	0.86	0.86	43
accuracy			0.84	76
macro avg	0.84	0.84	0.84	76
weighted avg	0.84	0.84	0.84	76

```
[[27  6]
 [ 6 37]]
```

K = 5

	precision	recall	f1-score	support
0	0.85	0.70	0.77	33
1	0.80	0.91	0.85	43
accuracy			0.82	76
macro avg	0.82	0.80	0.81	76
weighted avg	0.82	0.82	0.81	76

```
[[23 10]
 [ 4 39]]
```

K = 7

	precision	recall	f1-score	support
0	0.90	0.79	0.84	33
1	0.85	0.93	0.89	43
accuracy			0.87	76
macro avg	0.87	0.86	0.86	76
weighted avg	0.87	0.87	0.87	76

```
[[26  7]
 [ 3 40]]
```

K = 9

	precision	recall	f1-score	support
0	0.89	0.76	0.82	33
1	0.83	0.93	0.88	43
accuracy			0.86	76
macro avg	0.86	0.84	0.85	76
weighted avg	0.86	0.86	0.85	76

```
[[25  8]
 [ 3 40]]
```

K = 50

	precision	recall	f1-score	support
0	0.91	0.61	0.73	33
1	0.76	0.95	0.85	43
accuracy			0.80	76
macro avg	0.83	0.78	0.79	76
weighted avg	0.82	0.80	0.79	76

```
[[20 13]
 [ 2 41]]
```

Note. Accuracy is highest when K = 7 (0.87) while recall is highest when K = 50 (0.95). Accuracy is lowest when K = 50 (0.80) and recall is lowest when K = 1 (0.77)

K-Means Clustering

Finally, we implemented K-Means Clustering in order to group the data points and try to find trends using the proper library from scikit-learn. The goal of this model was to try and

compare target (response variable) with other attributes in order to try and find patterns in recognizing heart disease. To do this, the target and one other attribute were used at a time and were grouped into two clusters to separate those with heart disease and those without it. The attributes cp, thalach, and oldpeak were chosen as they had the highest correlation values with target.

Figure 10

K-Means Cluster for target and cp

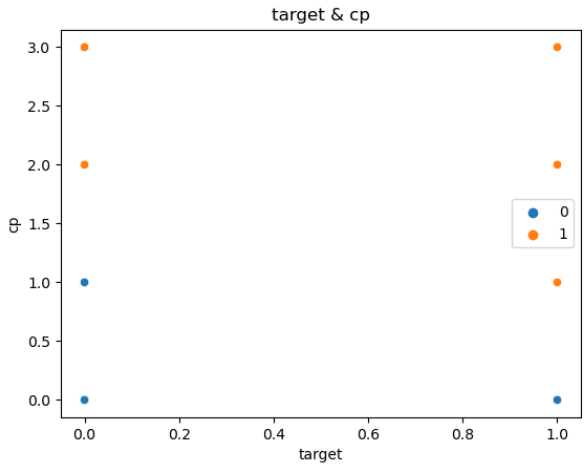


Figure 10

K-Means Cluster for target and slope

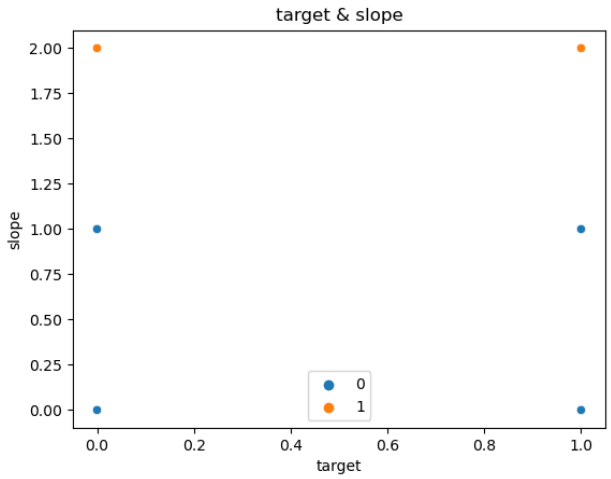
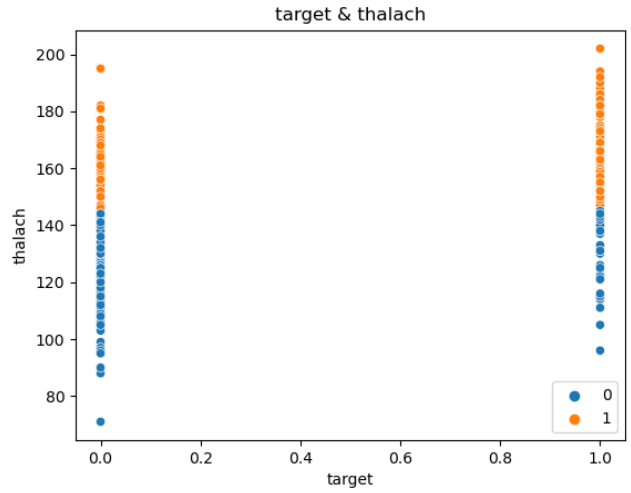


Figure 11

K-Means Cluster for target and thalach



For comparing target & cp and target & slope, these scatterplots proved to be unhelpful. Since these attributes have such a small range of values, the data points stacked on top of each other making it seem like there were only a few points. The plot of target & thalach was more

useful since thalach has a wider range of values. However, as target only has two possible values, all three of these clusters did not provide too much insight into how much of an influence the attributes have on heart disease. They just note that there is an influence of some sort.

The clusters were also created for the attribute pairs thalach /cp and thalach/slope since they also had relatively high correlation values.

Figure 12

K-Means Cluster for thalach and cp

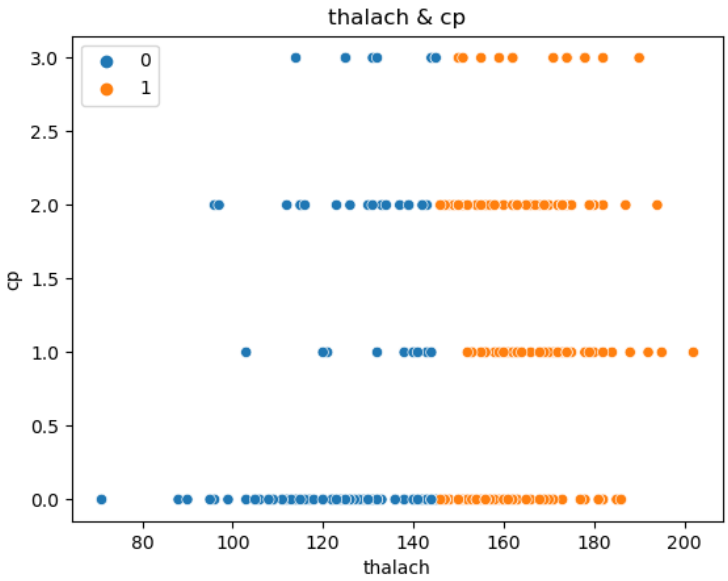
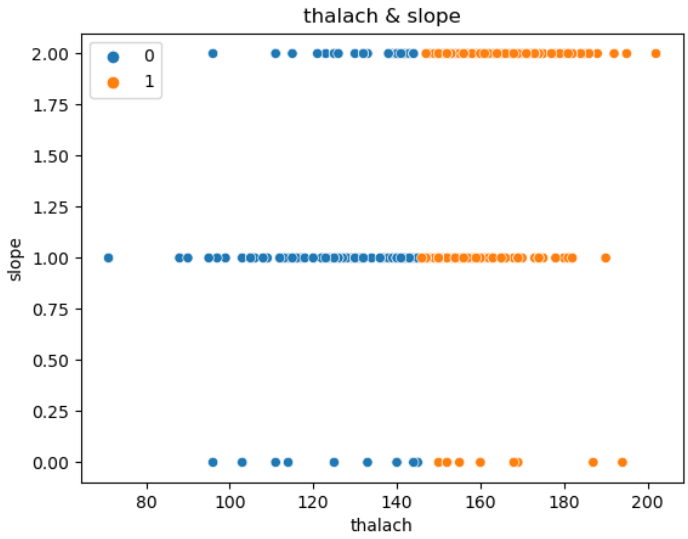


Figure 12

K-Means Cluster for thalach and slope



As cp and slope had a very limited range of possible values, thalach had a far greater influence on them. While these models do indicate how thalach relates to target, they do not say much about the influence cp and slope have on it.

Conclusion

When using accuracy as our primary metric, the best model is KNN when k is 7 with an accuracy of 87%, and the worst model is KNN when k is 50 with an accuracy of 80%. If we were to use the more contextually relevant recall as our primary metric, then the best model is interestingly KNN when k=50 with a recall of 95%, and the worst model is KNN when k=1 with a recall of 77%. If the goal is to correctly identify as many cases as possible, then KNN when k is 7 should be used. If the goal is to specifically catch all instances of heart disease as a means of prevention, then KNN when k=50 should be used.

Ultimately, the clustering models did not prove to be useful for the purposes of this analysis. They offered very few relevant insights into the data, and the insights they did could have most likely been better conveyed through other models. Due to the fact many of the numerical attributes are limited to a very small range of integers, the clustering models are not as insightful.

References

Jansoi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (2018, June 25). *Heart Disease UCI*. Retrieved from Kaggle: https://www.kaggle.com/ronitf/heart-disease-uci

Jansoi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (n.d.). *Heart Disease Data Set*. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Heart+Disease