Predicting Heart Disease Using Support Vector Machines

Sydney Orrison

7 April 2023

1 Purpose of Study

This study aims to determine the most important data features that contribute to concluding whether or not a person has heart disease, without making a tradeoff for accuracy by removing features. Based on the Stalog (Heart) Data Set, containing information from 303 hospital patients, this study uses a Support Vector Machine to train and test models with linear and non-linear kernel functions and compares their accuracy. Based on the results, the most important features are then tested and show that a much smaller subsection of the data set can be used to classify people with and without heart disease with nearly 97% accuracy. Conclusively, this study shows what factors are the most important to consider when a person suspects that they may be suffering from heart disease.

2 The Stalog (Heart) Data Set

The data set used in this study is the Stalog (Heart) Data Set, which is a part of the University of California, Irvine's data repository. This data set contains thirteen data features with information about 303 hospital patients, both male and female, between the ages of 29 and 77. This information was taken from each of the patients upon being admitted to the hospital. The fourteenth column in the data set, the **target** column, contains a 0 if that patient did not have heart disease at all, a 1 if they had heart disease, a 2 if they had low risk heart disease, a 3 if they had moderate risk heart disease, and a 4 if they had high risk heart disease. The chart below shows a data sample.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45.0	1.0	1.0	110.0	264.0	0.0	0.0	132.0	0.0	1.2	2.0	0.0	7.0	1
299	68.0	1.0	4.0	144.0	193.0	1.0	0.0	141.0	0.0	3.4	2.0	2.0	7.0	2
300	57.0	1.0	4.0	130.0	131.0	0.0	0.0	115.0	1.0	1.2	2.0	1.0	7.0	3
301	57.0	0.0	2.0	130.0	236.0	0.0	2.0	174.0	0.0	0.0	2.0	1.0	3.0	1
302	38.0	1.0	3.0	138.0	175.0	0.0	0.0	173.0	0.0	0.0	1.0	?	3.0	0

Figure 1: Sample of Heart Data

Each of the thirteen columns have short - and not very descriptive - column IDs. To make sense of these column names, a helpful chart is shown below. For the first run of the Support Vector Machine, all thirteen of these columns contain relevant feature information. Later on, the number of features used will decrease.

Feature Name	***Description***
age	Ranges from 29 to 77.
sex	Either 0 (Female) or 1 (Male).
cp	Represents the type of chest pain the person experiences, either 1 (typical angina), 2 (atypical angina), 3 (non-anginal pain), or 4 (asymptomatic/no pain).
trestbps	Resting blood pressure (in mm Hg on admission to the hospital).
chol	Serum cholesterol level (in mg/dl).
fbs	Fasting blood sugar level ($>$ 120 mg/dl, 1 = true; 0 = false).
restecg	Resting electrocardiographic results. Either 0 (normal), 1 (having ST-T wave abnormality), or 2 (showing probable or definite left ventricular hypertrophy by Estes' criteria).
thalach	Maximum heart rate achieved in beats per second (bps).
exang	Exercise induced angina? Either 0 (no) or 1 (yes).
oldpeak	Extend of ST depression induced by exercise relative to rest. ST depression is the decrease in the amplitude (height) of the ST segment of the electrocardiogram (ECG) waveform, a graphical representation of heart activity.
slope	The slope of the peak exercise ST segment. Either 1 (upsloping), 2 (flat), or 3 (downsloping).
ca	Number of major vessels (0-3) colored (partially or completely blocked) by fluoroscopy.
thal	The presence of a blood disorder called thalassemia. Either 3 (normal), 6 (fixed defect), or 7 (reversible defect).
target	Indicates the absence (0) or presence (1-4) of heart disease in each person. The values 1-4 represent the severity from low to high of the person's heart disease condition.

Figure 2: Column ID Chart

3 Data Processing

3.1 Removing Rows with Missing Data

Throughout the data set there are six rows, or six hospital patients, that contain a '?' in the place of information for one or more of the features. These rows were removed from the data set so that they do not cause issues later when the data is run through the Support Vector Machine. Figure 3 below is an example of what a row with missing data looks like.

302	38.0	1.0	3.0	138.0	175.0	0.0	0.0	173.0	0.0	0.0	1.0	?	3.0	0

Figure 3: Sample of Heart Data with Missing Column

The example above is a line with missing data in the **ca** column, which refers to the number of major vessels in the patient's heart that are either partially or completely blocked. Later testing will conclude that this feature is highly important in determining whether or not a person has heart disease, meaning that rows that have this column missing should definitely be excluded. The same is true of the **thal** column, which indicates the presence or absence of a rare blood disorder called thalassemia, as there are a few rows with that information missing as well.

3.2 Representing Heart Disease Severity

As is shown in the Figure 2 chart, the **target** column not only keeps track of whether or not a person *has* heart disease, but it also tracks the severity. The distribution of the **target** column is shown in Figure 4 below. The 0 column indicates the people that did not have heart disease, and columns 1, 2, 3 and 4 indicate that a person had heart disease in increasing severity respectively.

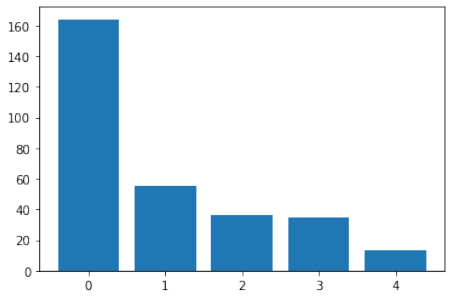

Figure 4: Heart Disease Presence Distribution

The histogram above visually displays the distribution of the **target** column. In total, there are 160 patients that did not have heart disease at all (0), 54 that had heart disease (1), 35 that had low risk heart disease (2), 35 that had moderate risk heart disease (3), and 13 that had high risk heart disease (4). Overall, of the 297 usable rows of patient information, 160 of them did not have heart disease and 137 did at some level of severity.

Early testing showed that it is not likely that a Support Vector Machine model will be able to accurately predict both whether or not a person has heart disease and the specific severity. In fact, the prediction was right around 50% of the time on average. Therefore the data is altered to fit the research purpose of this study. Every place in the **target** column that has a 1, 2, 3, or 4 is changed to only contain a 1. To accomplish this, the data set undergoes the following instruction.

```
df['target'] = df['target'].replace([1, 2, 3, 4], 1)
```

3.3 Data Normalization

Because the data is represented in a number of different units, e.g. age in years, heart rate in beats per second, etc., the data needs to be normalized before testing it with the Support Vector Machine. To do this, the data is run through a function which normalizes the data, scaling it down into values according to the Standard Scaler. The Standard Scaler changes the data so that the mean of the values is 0 and the standard deviation is 1. The normalization function, shown below, takes in the data frame containing the heart data and returns the normalized data.

```
def normalize(df):

    features = df.drop(['target'], axis=1)          # select the features to normalize
    scaler = StandardScaler()                       # initialize the standard scaler
    scaler.fit(features)                            # fit the scaler to the data
    normalized = scaler.transform(features)         # normalize the features

    return normalized
```

The result of this function is a matrix, contained in a list of lists format, with the normalized data. This will then be passed into the Support Vector Machine.

4 Testing the Support Vector Machine

Now that it has been processed, the data can be passed into the function that runs the Support Vector Machine. A Support Vector Machine functions by mapping the given data into a high feature space, and can then categorize the data, even in the case that it is not easily linearly separable. To do this, it finds the margin, which is the largest radius of the cylinder that is formed about the decision line in a classification problem. In other words, the margin is the largest region that can be placed and still separate two classes of points without containing any of the data points inside. When the margin placed around a decision line is the largest possible margin for any potential decision line for a particular data set, it is the maximum margin. Support vectors are the data points that lie the closest to the maximum margin, and they are used to find the ideal decision boundary with which to classify points.

4.1 Setting up the Testing Function

The **testModel** function that is used to run the Support Vector Machine, is relatively simple. The function takes in three parameters, **kernelType**, the kernel function that will be used for that particular SVM, **x**, the normalized data, and **y**, the target column containing the correct classifications of whether or not patients have heart disease. The data is split up into testing and training sets and a random state is chosen so that the random results do not vary from test to test. From there, the SVM is initialized, fit to the data, and prediction tests are run. Measuring the testing set's results against the SVM's predictions provides the measure of accuracy that we will use to determine how well the model functions.

4.2 Choosing a Kernel Function

In a support vector machine, kernel matrix represents the inner dot product of the original vectors, or data points. Data is transformed into a higher dimensional space but computations only need to occur in the original lower dimensional space. The kernel matrix is used to determine the decision boundary that maximizes the separation between the classes in the transformed dimensional space.

In order to determine the best kernel function to use for this particular data set, four different kernel functions are tested and their accuracy is compared. This is accomplished through multiple runs of the **testModel** with four different inputs in the **kernelType** parameter. The first kernel function is **linear**, which resulted in 96.7% accuracy. Immediately, this value is surprising, as this indicates that the data with thirteen features is already highly linearly separable. In other words, there is a straight line that can be drawn between data points to separate and classify them with very high accuracy. The second kernel function is **polynomial**, which is a non-linear learning model that casts data points in a feature space of polynomials of those points, which is 98.3% accurate. As there is already higher accuracy with a kernel function that is meant for non-linear data, it indicates that linear is not likely the best kernel function for this data. However, it should be noted that the accuracy of the linear kernel function SVM is still very high. Third, the **radial basis function** kernel type, which is another popular choice for non-linear data, has 96.7% accuracy, which is the same as the linear kernel. Finally, the fourth kernel function, **sigmoid**, resulted in 98.3% accuracy, similarly to the radial bias function kernel. Although the non-linear kernel functions seem to be the better option for this data, the linear function still performed very well. In this case, all four kernel functions will continue to be tested on any manipulated versions of the data to see what the differences may reveal about the data.

5 Determining Feature Importance

After running the Support Vector Machine on the normalized data with thirteen features, it still is not clear which of the features are the most responsible for the high-accuracy classifications that were made. To find out, the **coef_** attribute, which is only applicable for linear kernel functions, can be used to determine which features are the most important in determining whether or not a given person has heart disease. The result can be displayed in a heat map, as is shown below.

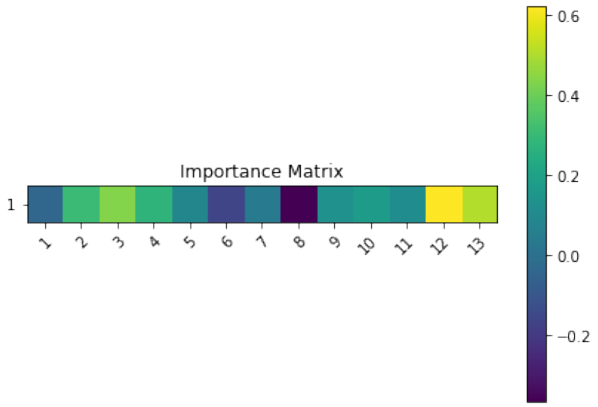

Figure 5: Importance Heatmap for All Thirteen Features

Looking at the heatmap in Figure 5 above, it appears that the features in the second, third, fourth, twelfth, and thirteenth columns are most important indicators of whether or not a given person has heart disease with accuracy in determining its severity. In other words, these features are the most influential in determining the final decision of the Support Vector Machine when classifying people into categories based on whether or not they have heart disease. These correspond with the person's sex, type of chest pain, resting heart rate, number of major vessels blocked, and presence of the blood disorder, thalassemia, respectively. If these really are the most important features, it can be hypothesized that removing all other features but these five will still result in a highly accurate prediction from the Support Vector Machine.

6 Testing the Support Vector Machine on Five-Feature Data

To test the model on data with less features, the unnecessary columns are dropped from the original data, the data is normalized, and then it is passed into the **testModel** function just as before. Similarly to the thirteen-feature data, this new five-feature version of the data will be run on all four of the kernel functions that were tested previously.

The results are very interesting, and can explain a lot about the nature of the data. For this five-feature data, the **linear** kernel function had 96.7% accuracy, the **polynomial** kernel function had 96.7% accuracy, the **radial basis kernel function** had 90% accuracy, and the **sigmoid** kernel function only had 83.3% accuracy. This shows that this smaller subset of the data is actually *more* linearly separable, if not just as linearly separable, as the larger thirteen-feature data set. Additionally, when using the linear kernel function, the results from the smaller data set appear to be just as accurate. The heatmap in Figure 6 below shows the importance heatmap for the smaller data set, which indicates that a person's sex and whether or not they have the blood disease, thalassemia, may be the most important indicators of whether or not a person has heart disease.

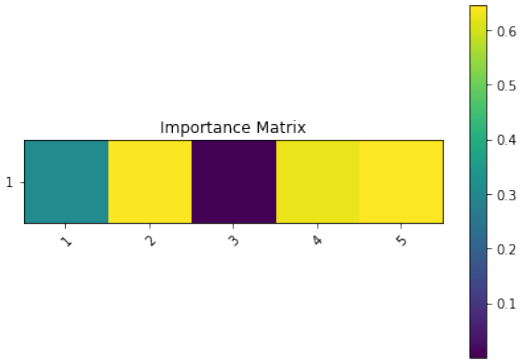

Figure 6: Importance Heatmap for Selected Five Features

7 Conclusion

Using a Support Vector Machine, it was determined that predicting whether or not a person has heart disease may be possible with less features, but shows that the more features provided, the more accurate the result will be. What were determined as the two most important factors, a person's sex and whether or not they have the blood disease, thalassemia, are not significantly *more*

important than other factors, so it can not be strongly concluded that they are direct indicators of a person's heart disease. While a Support Vector Machine is able to do a simple classification, a more complicated task, such as predicting the severity of a person's heart disease, did not appear to be highly accurate. Additionally, although surprising, this data appeared to be highly linearly separable, indicating that there are some clear differences between the data points for patients that do and do not have heart disease. This makes the data easier to classify using the SVM.

8 Sources

The Stalog (Heart) Data Set
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.