Prediction of Heart Disease Using Different Machine Learning Methods

Jazmine Boloor

Dr. S. Latifi, ECG 703 – Machine Learning

Electrical and Computer Engineering Dept.

University of Nevada, Las Vegas, USA

***Abstract*-- The goal of this research is to analyze different methods of machine learning that can be used to accurately predict the chance that a patient has a heart disease based on their specific symptoms and overall health. Support vector machines, logistic regression, and decision tree based methods will be analyzed to show their accuracy as well as their pros/cons in the prediction of heart disease. The effectiveness of the methods using different numbers of components will be compared to each other, where components with the least impact on the output will be ignored one by one to analyze the effect in accuracy. Then, the optimal number of components will be discussed. There is great value behind this research because of the alarmingly high numbers of heart disease in the world, particularly the United States. The future of machine learning presents a method of diagnosis for this issue that is completely noninvasive. This paper will discuss, analyze, and compare different machine learning methods that can be used in the prediction of cardiovascular disease.**

***Index Terms*-- Machine Learning, Heart Disease, Principal Component Analysis, Decision Trees, Support Vector Machines, Logistic Regression**

I. INTRODUCTION

The goal of machine learning is to discover patterns that may exist in data and implement learning capabilities into computer systems to recognize them [6]. Applying this concept to the medical field can be powerful in prediction of certain diseases. According to the Center for Disease Control (CDC), heart disease is the leading cause of death in the United States, with one out of every four deaths being associated with it [3]. This is clearly an issue, and techniques to better assess, predict, and manage heart disease is an ongoing research topic. Thus, the discovery of accurate prediction models can be useful, as it could mean that a noninvasive method of evaluation that is purely based on previous accounts is available. Thus, machine learning will be applied to a dataset describing the health statistics of patients to show the accuracy of prediction that can arise.

The following sections discuss the use of support vector machines (SVM), logistic regression, and decision tree machine learning methods that can be used as an aid in heart disease prediction. The use of different amounts of inputs are examined to show the optimal number of inputs that a machine would ideally want to accept for the dataset, and the results are compared.

II. DATASET SPECIFICATIONS

The University of California, Irvine created a machine learning repository that includes truthfully recorded data on a wide range of topics. The dataset used in this research was the heart disease dataset. The file provided in the repository includes several different items, including health statistics from three different hospitals. Heart disease patients in Switzerland, Hungary, and the United States were studied; the following research mainly focuses on those studied in the American hospital, located at the Cleveland Clinic in Ohio [4]. This clinic was used because it was the only one of the three with published studies already done on it, and comparison to other results was an integral part of the study. Note that it was important that real and accurate data was used when conducting this research so as to come to accurate and reasonable conclusions. Thus, the conclusions made in the following sections are the ground truth for this dataset, where 303 patients are studied. For all three of the methods tested, an eight to twenty ratio of training and testing models was used – thus 242 training points and 61 testing points.

The information collected by the Cleveland Clinic's surrounded 76 different variables, fourteen of which have been sorted for study [4]. These fourteen variables include: age, sex, chest pain type, resting blood pressure, fasting blood sugar, resting ECG reading, cholesterol, exercise induced angina, max heartrate, ST depression, ST slope, vessels colored, thalassemia, and diagnosis [4]. The other variables collected were discarded for the purpose of this research for a plethora of reasons, such as an insufficient amount of responses or general lack of influence on the data.

The dataset revolves around accurate data that was collected from real test subjects, meaning that the patients should reflect the general population in some capacity (though the relatively small number of subjects should be kept in mind). This allows for inferences that are based on medical and biological standpoints to be made on the data, as is done so throughout the rest of this paper. For this reason, the general guidelines that are associated with the features are listed below for better analysis in future sections. Additionally, some general facts are listed that may contribute to the results of the data, as well as some clarifications for a better understanding of the way the data was processed.

- Age: Risk factors for heart disease tend to start around age 35, and begin to increase more rapidly around age 50. After age 65, the risk is the highest. A higher age is associated with a higher risk [7] [3].
- Sex: Men of almost every race are statistically more likely to develop heart disease than women [3]. Note that this value was recorded in a binary fashion, where men are listed as one and women are listed as zero.
- Chest Pain Type: The values recorded for chest pain were a number one through four, where [4]:
  1 – typical angina
  2 – atypical angina
  3 – non-anginal pain
  4 – asymptomatic

Note that anginal pain occurs when the heart is not receiving enough oxygen in the blood [2]

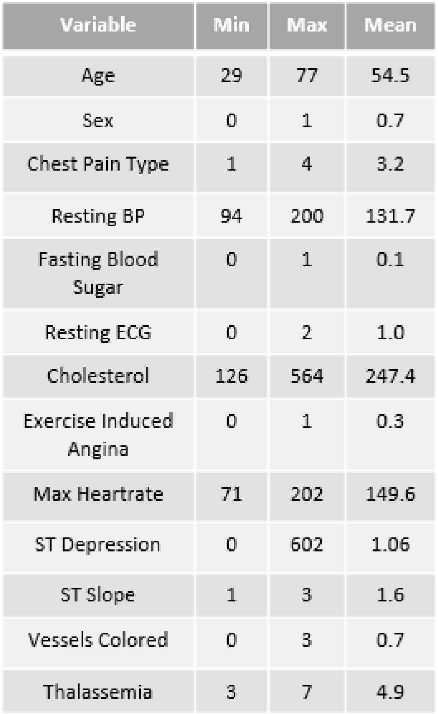
- Resting Blood Pressure: Only systolic blood pressure was recorded for this study, where recommended values range from 90 mm Hg to 120 mm Hg; 120-129 mm Hg is elevated blood pressure, 130-139 mm Hg is stage 1 hypertension (or high blood pressure), 140-179 mm Hg is hypertension stage two, and above 180 mm Hg is hypertensive crisis (requiring immediate medical attention) [3].
- Fasting Blood Sugar: Recommended between 70 mg/dL and 100 mg/dL; over 126 mg/dL indicates diabetes [3]
- Resting ECG Reading: The values recorded for the resting ECG reading were a number zero through two, where [4]:
  0 – normal
  1 – ST-T wave abnormalities
  2 – showing signs of left ventricular hypertrophy (an enlargement of the chamber) [2].
- Cholesterol – The values recorded for cholesterol correspond to total cholesterol, where recommended values range from 120 mg/dL to 200 mg/dL.
- Exercise Induced Angina: The values recorded were either a one or a zero, where one indicated a patient did have exercise induced angina and zero indicated that a patient did not have exercise induced angina.
- Max Heartrate: Recommended values for maximum heartrate are dependent on age, where thirty year old's should see around 190 beats per minute (beats per minute), 55 year old's should see around 165 bpm, 70 year old's should see around 150 bpm, and ages in between those listed should also see heart rates in between those listed.
- ST Depression – Recommended between zero and 0.1 mV
- ST Slope – This value was recorded as a value between one and three, where [4]:
  1 – upsloping
  2 – flat
  3 – down sloping

  The ST slope was recorded was taken while a patient was exercising [4].
- Vessels Colored: This was recorded as a number zero through three, where the number listed is the amount of major vessels that were colored by a fluoroscopy [4]. Note that the colored vessels are highlighted as possible need for attention.

- Thalassemia: This is a blood disorder that causes the body to have a lowered amount of hemoglobin, which can lead to several different heart abnormalities. The number recorded corresponded to the either a three, which was normal, a six, which was a fixed defect, or a seven, which was a defect that was reversible [4].

The last feature associated with the diagnosis was the diagnosis. UCI made this column of the dataset a number that was either zero (indicating no presence of heart disease), or greater than zero (indicating presence of heart disease) [4]. For a more simple analysis, the data in this column was converted into a binary 0 or 1 (with the same descriptions).

It is important to note that most of the guidelines above are directly listed from American associations, such as the Centers for Disease Control and Prevention (CDC), so these numbers may differ slightly from other countries recommendations. Some general statistics about the dataset, such as minimum, maximum, and mean values, are shown in Table 1. There are some outliers within many of the features, and analysis on the effect of these cases may be valuable in future work.



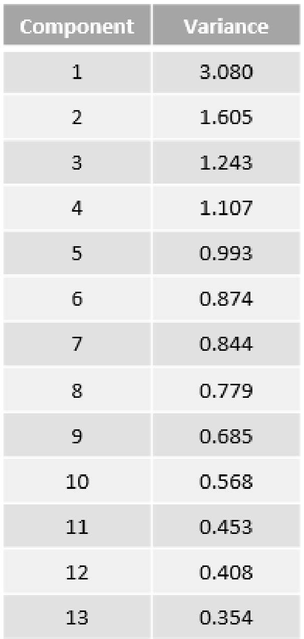
Variable	Min	Max	Mean
Age	29	77	54.5
Sex	0	1	0.7
Chest Pain Type	1	4	3.2
Resting BP	94	200	131.7
Fasting Blood Sugar	0	1	0.1
Resting ECG	0	2	1.0
Cholesterol	126	564	247.4
Exercise Induced Angina	0	1	0.3
Max Heartrate	71	202	149.6
ST Depression	0	602	1.06
ST Slope	1	3	1.6
Vessels Colored	0	3	0.7
Thalassemia	3	7	4.9

Table 1 – Describing the dataset features

Note that the values for the features, such as sex, are described in the above section (because they are a number value rather than a description).

III. ANALYZING THE DATASET USING PCA

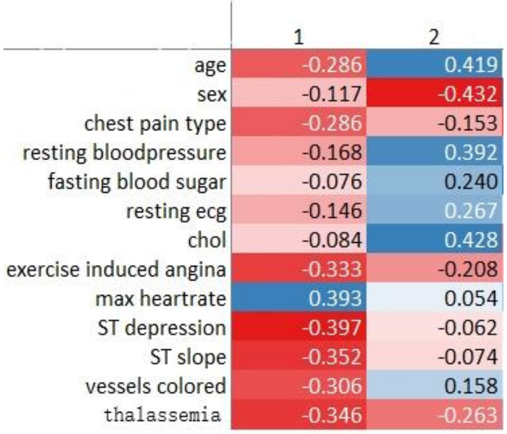
The first test that was run on the dataset was the Principal Component Analysis (PCA). This allowed for the components to be listed in order of the weight they held on the diagnosis. Table 2 was created using Excel's PCA extension, and it shows the variance of each of the components that the PCA constructed. It is clear from the table that the first principal component had the most impact on the variance, followed by the second, third, ect. Note that this is to be expected, and confirms that the analysis was run correctly. The first principal component (PC) has almost double (at least) of the variance seen in any other component, meaning it has a more significant impact on the output in comparison to the rest of the features. In general, the table makes it clear that the variances of the features decrease nonlinearly (but still trend downward), which is to be expected.



Component	Variance
1	3.080
2	1.605
3	1.243
4	1.107
5	0.993
6	0.874
7	0.844
8	0.779
9	0.685
10	0.568
11	0.453
12	0.408
13	0.354

Table 2 – Principal Components Vs. Variance

Figure 1 shows the individual impacts that each of the features has on the first two principal components, which will be analyzed for clarification. The chart is color coded, and the items that are shown in darker red or darker blue have a higher value. Note that the positive/negative values just show correlation, so their absolute values are analyzed in this case.



	1	2
age	-0.286	0.419
sex	-0.117	-0.432
chest pain type	-0.286	-0.153
resting bloodpressure	-0.168	0.392
fasting blood sugar	-0.076	0.240
resting ecg	-0.146	0.267
chol	-0.084	0.428
exercise induced angina	-0.333	-0.208
max heartrate	0.393	0.054
ST depression	-0.397	-0.062
ST slope	-0.352	-0.074
vessels colored	-0.306	0.158
thalassemia	-0.346	-0.263

Figure 1 – Analysis of first two principal components

Beginning with PC one, it can be seen that the features with the greatest association are max heartrate, ST depression, ST slope, and thalassemia, while features such as blood sugar and cholesterol have a very small impact on the component. Moving on to PC two, this component shows completely different associations. PC two has strong impacts from the sex, age, cholesterol, and blood pressure of a patient, while it has very small impacts from most of the features listed to impact PC one. The entire chart is shown in Figure 2, and can be zoomed in on and analyzed in the same way as stated above. Again, according to the variances, the first four components show variance values larger that one and the rest of the values then start to decline. This should also be considered when analyzing Figure 2, as the features more heavily associated with the first few principal components will have more pull on the output (and the last few will have the smallest pull on the output).

The principal component analysis was used along with all three of the machine learning algorithms that the next sections address – support vector machine, logistic regression, and decision tree. This allowed for an analysis on the optimal number of PC's to use given this dataset

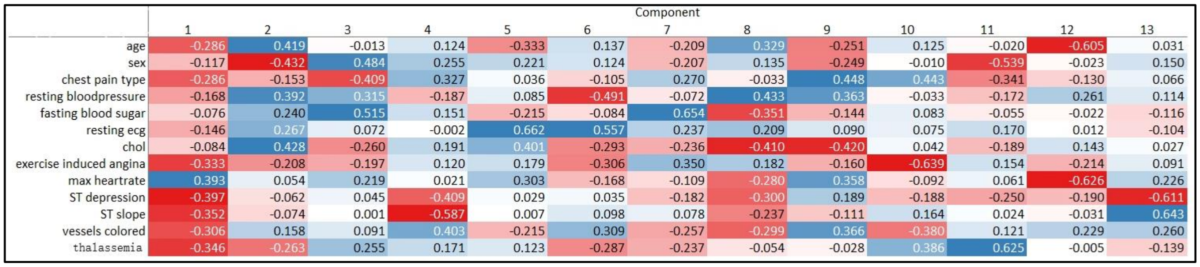
IV. SUPPORT VECTOR MACHINE RESULTS

To begin analysis on the dataset using the SVM approach, the algorithm was run using all thirteen input features and the singular output. The machine was trained using the 242 training values and then tested using the 61 test values discussed previously to show what the likelihood that it would make a correct decision (for accurate data) would be. Table 3 shows the results of this test. The next columns in Table 3 show the results of teaching the machine using one less principal component per test. The results of tests for 3-13 PCs are listed.

Number of PCs	SVM Accuracy
13	83.2 %
12	83.2 %
11	83.5 %
10	81.2 %
9	81.5 %
8	81.8 %
7	78.5 %
6	79.2 %
5	76.9 %
4	68.3 %
3	68.0 %

Table 3 – Results of the SVM tests

It is clear from the table above that the result of the SVM learning algorithm is not linear



	Component												
	1	2	3	4	5	6	7	8	9	10	11	12	13
age	-0.286	0.419	-0.013	0.124	-0.333	0.137	-0.209	0.329	-0.251	0.125	-0.020	-0.605	0.031
sex	-0.117	-0.432	0.484	0.255	0.221	0.124	-0.207	0.135	-0.249	-0.010	-0.539	-0.023	0.150
chest pain type	-0.286	-0.153	-0.409	0.327	0.036	-0.105	0.270	-0.033	0.448	0.443	-0.341	-0.130	0.066
resting bloodpressure	-0.168	0.392	0.315	-0.187	0.085	-0.491	-0.072	0.433	0.363	-0.033	-0.172	0.261	0.114
fasting blood sugar	-0.076	0.240	0.515	0.151	-0.215	-0.084	0.654	-0.351	-0.144	0.083	-0.055	-0.022	-0.116
resting ecg	-0.146	0.267	0.072	-0.002	0.662	0.557	0.237	0.209	0.090	0.075	0.170	0.012	-0.104
chol	-0.084	0.428	-0.260	0.191	0.401	-0.293	-0.236	-0.410	-0.420	0.042	-0.189	0.143	0.027
exercise induced angina	-0.333	-0.208	-0.197	0.120	0.179	-0.306	0.350	0.182	-0.160	-0.639	0.154	-0.214	0.091
max heartrate	0.393	0.054	0.219	0.021	0.303	-0.168	-0.109	-0.280	0.358	-0.092	0.061	-0.626	0.226
ST depression	-0.397	-0.062	0.045	-0.409	0.029	0.035	-0.182	-0.300	0.189	-0.188	-0.250	-0.190	-0.611
ST slope	-0.352	-0.074	0.001	-0.587	0.007	0.098	0.078	-0.237	-0.111	0.164	0.024	-0.031	0.643
vessels colored	-0.306	0.158	0.091	0.403	-0.215	0.309	-0.257	-0.299	0.366	-0.380	0.121	0.229	0.260
thalassemia	-0.346	-0.263	0.255	0.171	0.123	-0.287	-0.237	-0.054	-0.028	0.386	0.625	-0.005	-0.139

Figure 2 – Complete View of the PCA Feature Percentages Per Component

with respect to the decline in the number of PCs. Linearity was not to be expected due to factors such as the curse of dimensionality; however, the table does show an overall trend downwards with the decline in PCs used. Thus, the optimal number of PCs may differ depending on the use. For instance, if one was looking for the absolute best accuracy that was found during these experiments, it would be found with eleven PCs. However, given that there is only a 1.4% decrease in accuracy when testing eight PCs as opposed to thirteen, this could be a tradeoff that is desirable to narrow the data with a fairly low decrease in accuracy. Another observation that can be made about the data is that there is a rather significant decrease in accuracy after five PCs are tested – any number lower than this resulted in a much lower accuracy. The figure below shows one of the many plots that were created in association to the data – the "x" markers show the predicted data. Note that the graph shows the age vs. blood pressure statistics, but a plot using any of the given variables would show the same points on different axis's. Note the outlier on this dataset – the patient with a cholesterol over 100 mg/dL larger than any other patient, and over 300 mg/dL larger than the recommended value given by the CDC [3]. This insinuates research on how outliers in this dataset may be valuable to future work.

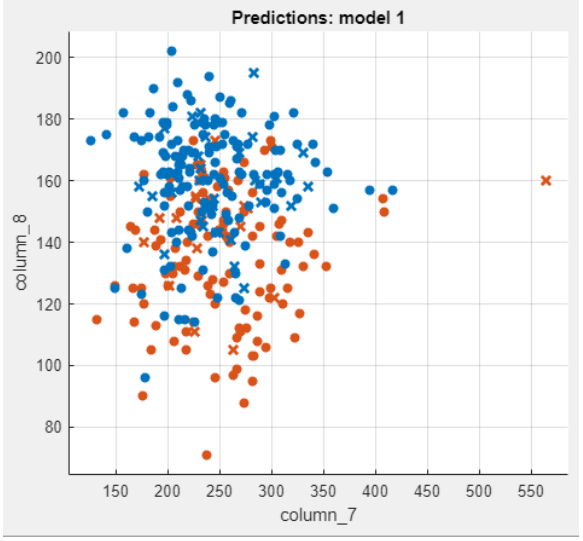


Figure 3 – Showing the distribution of learned datapoints; x-axis shows cholesterol and y-axis shows max heartrate

The nature of most algorithms that predict healthcare require a very low amount of false negatives. This is due to obvious reasons – it can be detrimental to misdiagnose someone as healthy when they actually have a heart disease. Though the percentage of false positives would ideally also be low for accurate and reliable results, the impact of false negatives is more palpable and should therefore be addressed first. Figure 4 shows the SVMs confusion matrix, which shows the amount of true/false predictions. The results are presented in number of patients; the accuracy was calculated by dividing the amount of correct cases (true positive, TP, and true negative, TN) by the amount of total cases (TP,TN, false positive, FP, and false negative, FN), thus:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

As shown in Figure 4, the algorithm accurately decides on 32 negative cases and 18 positive cases; this leaves 3 false positives and 8 false negatives. Again, future work should focus first on lowering these false negatives, as this would create the biggest issue in any real life application.

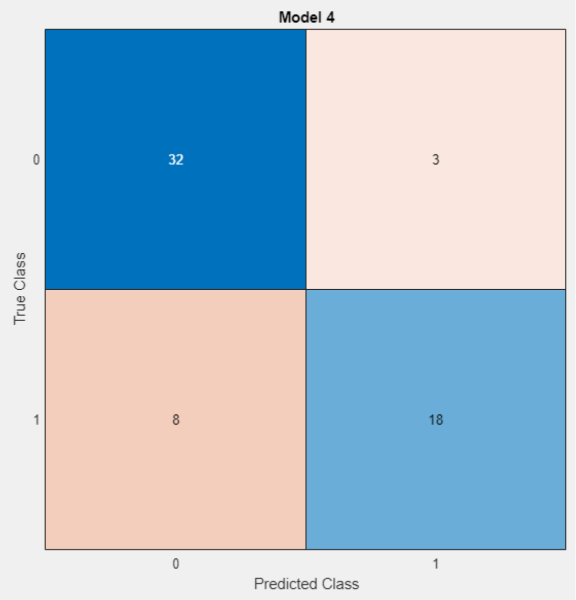


Figure 4 – Confusion Matrix for SVM Showing Issue with False Negatives(8 PCs)

Overall, the SVM method had reasonable results, though future work would want to address the issue of false negatives and inaccurate predictions in general. Additionally, more training/test data may be helpful in better addressing the full extent of the experiments (to a certain degree, as too much information could also cause issues).

V. LOGISTIC REGRESSION RESULTS

Similarly to the previous section, the logistic regression algorithm was trained using the 242 training values and then tested using the 61 test values, and then this was repeated with a one less PC for each test. Again, the accuracy calculation comes from the data in the confusion matrix (and the accuracy equation stated previously). The results of tests for three through thirteen principal components are listed in Table 4. From first glance, is clear that these results are comparable to those found for the support vector machine method previously tested.

Number of PCs	Logistic Regression Accuracy
13	84.2 %
12	82.8 %
11	83.5 %
10	81.2 %
9	82.2 %
8	82.5 %
7	80.2 %
6	80.2 %
5	75.5 %
4	69.0 %
3	69.3 %

Table 4 – Results of the logistic regression tests

The results shown for the logistic regression approach proved to be relatively similar to the SVM method. The highest accuracy found throughout all of the tests done was the first logistic regression test run, the one using all thirteen of the PCs. This experiment found 84.2% accuracy. There is an obvious tradeoff to using this value, as it takes into account all 13 of the PCs. If there is some room for error, the amount of PCs can be more than cut in half for approximately four percent less accuracy (or two percent less in comparison to the twelve PC test accuracy). Thus, the optimal number of PCs in this case may be thirteen, if accuracy is the most important factor, while six PCs would be optimal for the most least amount used with reasonable accuracy.

Figure 5 shows the distribution of the patients, where the x axis shows age and the y axis shows max heart rate; the “x” markers correspond to predictions, given the ground truth data collected. Note the color distribution in the figure – patients with higher ages and lower max heart rates tended to diagnose positive (red), while patients with lower ages and higher max heart rates tended to diagnose negative (blue). This intuitively makes sense given the CDCs guidelines, and suggests that many of the patients with heart disease were not able to get their heartrates very high in comparison to those with healthy hearts. The clustering of positive/negative diagnosis’s is fairly clear in this distribution of the data.

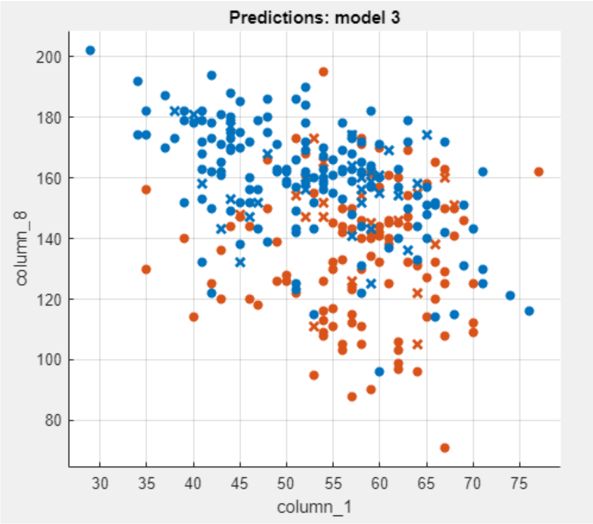


Figure 5 – Showing the distribution of learned datapoints; x-axis shows age and y-axis shows max heartrate

Figure 6 shows the confusion matrix for the data when seven principal components are used (or 80.2% accuracy). This particular amount of PCs was chosen at random because almost every amount chosen highlighted the biggest issue with this research – the strikingly high amount of false negatives. As explained previously, false negatives pose a huge issue in medical predictions. It is clear that there are more false negatives (7) than false positives (5) and entirely too many in general. Note that this number fluctuates as the number of PCs and the accuracy fluctuates, but the false negatives consistently outweigh the false positives. This should be the first issue tackled with future work in this realm – the false negatives must be shrunk to a much smaller percentage, preferably effectively zero, for this algorithm to be realistically applicable. Due to time constraints, this was not attempted for this paper, but filtering codes that would double check each feature for better association could be implemented.

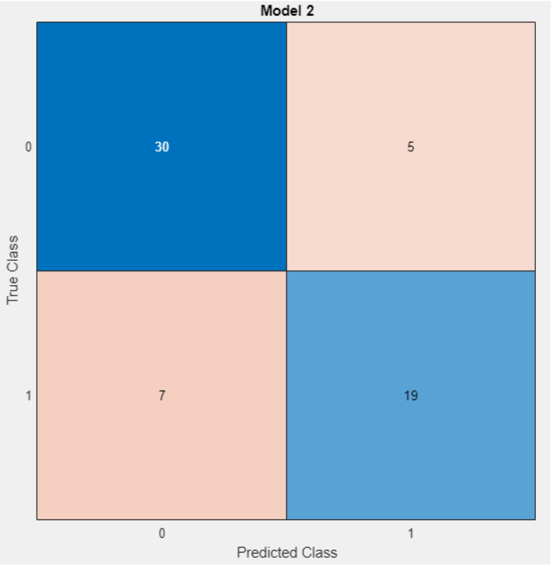


Figure 6 – Confusion Matrix for LR Showing Issue with False Negatives (7 PCs)

VI. DECISION TREE RESULTS

Tests for the decision tree method were run in the same fashion described in the previous two sections. The results for this method for three through thirteen principal components are shown in Table 5. It should be noted that only a fine decision tree was tested – this used a maximum of 100 splits in the data. Again, not that the data was trained using 242 data points, and then tested using 61 (for around an eighty to twenty ratio).

The results for accuracy were obviously lower for this test when compared to the support vector machine and logistic regression methods. This could be attributed to many different factors. The first factor that may be explored in order to make this method have more precise accuracy would come with changing the order of the features, which is discussed more in the next section. Another reason this test may not have performed as well as well as the SVM and LR ones did is because of the nature of the algorithm. SVM and LR classify the patients through a separation (where SVM uses a line for this case and LR uses a section on a function) – thus allowing for clusters of patients with similar features to be "next to" each other. The SVM algorithm classifies each patient while the LR one assigns a percentage. The decision tree analyzes each component per patient and splits it according to the results. This may not be the best method for a dataset with the amount of features specified here (with as small a dataset as used).

Number of PCs	Decision Tree Accuracy
13	77.2 %
12	75.4 %
11	74.6 %
10	74.6 %
9	74.6 %
8	74.6 %
7	76.2 %
6	76.6 %
5	73.9 %
4	68.0 %
3	64.4 %

Table 5 – Results of the logistic regression tests

The following images show a scatter plot and confusion matrix for the decision tree, as was presented with the previous methods.

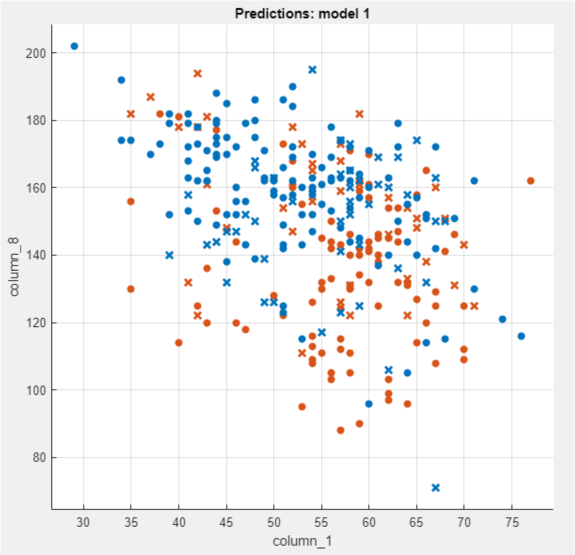


Figure 7 – Showing the distribution of learned datapoints; x-axis shows age and y-axis shows max heartrate

One particularly striking issue that this method posed was its number of false negatives. As elaborated on previously, the false negatives would be the most damaging issue that these predictions could cause with real use. Around twenty percent of the predictions were false negatives, making this test the worst for this issue as well. Future work on research using this method would definitely want to focus on this issue.

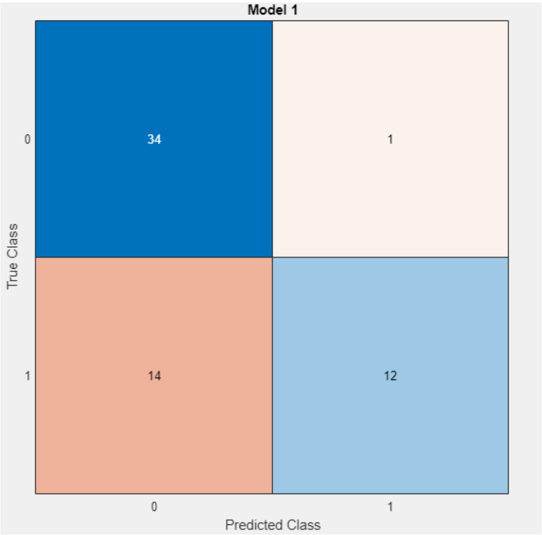


Figure 8 - Confusion Matrix for Decision Tree Showing Issue with False Negatives (12 PCs)

VII. SUMMARY AND ANALYSIS OF TEST RESULTS

A. *Sensitivity to Change in Order of Features*

Due to the nature of machine learning, it is valuable to analyze the data given the inputs in a different order, as depending on the algorithm, this can change the results of the accuracy that the machine can predict at. Thus, tests were run to analyze how a different order of features would change each of the test results. The previous tests were run with the same order of components as they were presented for all three of the tests; for the following tests, the order of the features was modified randomly. It is to be assumed that the support vector machine and logistic regression results would not change, as they are independent of variable order. The decision tree could be predicted to change slightly. The results using all thirteen features is shown below, and confirm these predictions. It can be assumed that the results of the accuracies for the rest of the PCs tests may also change slightly if the order of the input features changes.

Test	**Previous Accuracy**	**New Accuracy**
SVM	83.2 %	83.2%
LR	84.2%	84.2%
Decision Tree	77.2 %	74.6%

Table 6 – Changing the Order of Features (Thirteen Components)

Thus, it may be valuable to inspect different variations of the decision tree model if that algorithm was chosen despite its lower accuracy. More about this in the Decision Tree Results section of the paper.

B. *Analyzing the Optimal Algorithm*

The results of these methods can be analyzed in different ways when attempting to find the optimal method and optimal number of components. If attempting to find the absolute best accuracy that can be found within all of these tests among every number of principal components, then using the logistic regression approach with all thirteen components would be the best choice given the experimental results. On the other hand, if the amount of principal components was being optimized to find the smallest number of PCs with a comparable accuracy, both the LR and SVM methods can cut their PCs from thirteen to eight with only about a two percent decrease in accuracy. This may be optimal in applications that don't want to use thirteen principal components. Additionally, this may be even more optimal if an application did not want to record certain data, such as the resting ECG reading, as this component is not heavily influencing any of the first four principal components. Being able to possibly not include a feature like this could drastically decrease the cost of finding all of the components, as well as increase the accessibility. Finding the exact effect of excluding a feature like this would be powerful future work.

Figure 9 shows the accuracies throughout the number of PCs. It is clear that SVM and LR have very similar results. It is also cleat that the decision tree method performed noticeably worse than the other two (for reasons probable to those discussed previously).

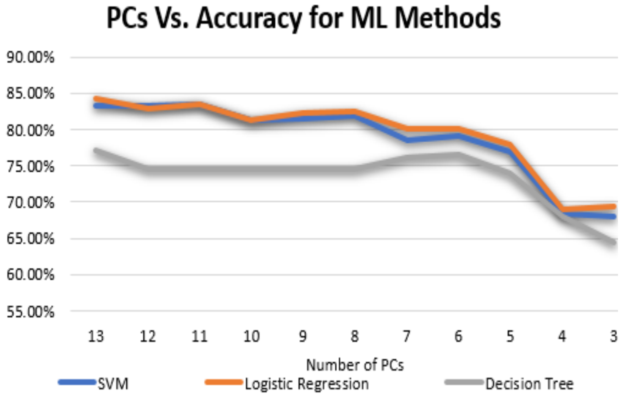


Figure 9 – Graph of ML Methods Accuracies

Another thing to note about Figure 9 is that none of the methods operate in a linear manner. Though all three trend down in accuracy as the amount of PCs decrease, the accuracy seems to stay rather stable until around seven components (and less) for all three of the methods. At this point, the accuracy steadily declines as the amount of PCs decrease.

VIII. DISCUSSION OF RELATED WORK

As mentioned previously, this dataset has had previous studies performed on it. This section of the report will examine other research that compares to the work done in this one. Note that none of the research presented in this section will analyze the dataset in terms of principal components, as done in this research. Rather, only all fourteen features will be addressed.

Author Rawat did a study in 2021 comparing seven different ML algorithms to test this dataset. Though no conclusions were made testing PCs, the data using the thirteen features proved to have similar results using their algorithms. Their machines were trained/tested with the confusion matrix [8]. When they testing the data using the SVM method, they received 80.3% accuracy; using the LR method, they received 80.3% accuracy (matching SVM); using the decision tree method, they received 77% accuracy [8]. Not only do these findings compare to the ones found during this research, but they are trend in the same way (SVM and LR are similar, while DT is considerably less accurate). It should be noted that while these accuracies are similar, they are slightly smaller than the ones found in this research. This could be due to many different things – the most probable of which being that a different test set was used for both sets of research. The research in this report was done using test data that was the ground truth – data from the machine learning repository, which consisted of 303 patients [4]. The research in the study done by Rawat was tested using 61 of these cases, similar to the research presented, however the exact cases used were not known so they may have been different than those used in the presented work [8]. Thus, these accuracies may reflect a different training/testing set, which would make sense given how close they are to the ones found in the research presented.

Rawat also noted other machine learning algorithms, not presented in this paper, that performed with less accuracy than those presented. For instance, the random forest method was also tested, using the same criteria mentioned previously, and was found to have 75% accuracy [8]. The Naïve Bayes method was tested and found 78% accuracy [8]. LightGMB and XGBoost were tested as well, finding 77% and 75% accuracies, respectively [8]. From these results, it is clear that all of the machine learning algorithms behaved with similar accuracies, but SVM and logistic regression were found to have the highest accuracies, similar to the findings presented in this research.

An IEEE article written by Sutedja performed similar experiments to those done in the study presented in this report, though again focusing on all thirteen of the features without applying the principal component analysis (or any kind of data minimization) [5]. Their findings surrounded both machine learning and deep learning, the ladder of which will not be discussed. Their machine learning findings also composed of an eighty to twenty ratio of learning/testing data. Accuracy findings are summarized as SVM with 88%, and LR with 86% [5]. Thus, analysis using this approach outperformed the ones presented in this paper using thirteen components. Future work that implemented the approach using three through thirteen components would be valuable for further comparison to the data collected in this report.

Another study, performed by Burleigh and explored in Python in 2020, examined logistic regressions algorithm on the dataset (again, without using PCA or any kind of data slimming) [9]. This study found that their model was 75% accurate in predicting heart disease (46 of 61 test cases were predicted correctly) [9]. Also noted in this study was the amount of false positives/negatives – of which 5/61 were false positives and 10/61 were false negatives [9]. As mentioned previously, there is a very small threshold for false negatives in predictions of this nature – it seemed that this authors' test also experienced a high number of false negatives in comparison to the overall number of false cases.

Related work was also done in 2020 by Islam, where all features of the dataset were characterized using logistic regression [1]. This test used similar criteria to the previously mentioned ones and similar to the criteria used in the presented research – 61 test cases to calculate the results [1]. Their results found 89.19% accuracy using logistic regression, the highest of the analyzed research [1]. Beyond the accuracy being the highest, this study also found the lowest amount of false negatives. The confusion matrix that this study found for the 61 tests is shown in Figure 10. The study found only 1/61 false negatives according to the confusion matric they presented [1].

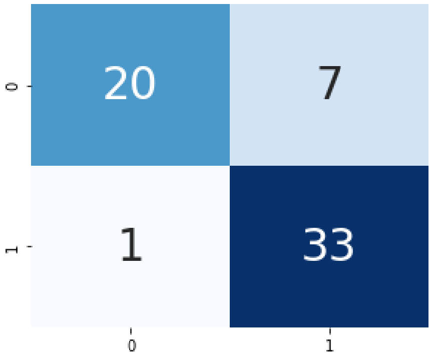


Figure 10 – Showing the Considerably Small Amount of False Negatives [1]

Future research may want to examine the algorithm used here to see if there is a similarly small amount of false negatives while also incorporating the PCA as was done in the presented research.

IX. MORE FUTURE WORK / CONCLUSION

A. Future Work

Though examples of future work have been detailed throughout the report, one thing that has not been addressed is using a different test dataset. Due to privacy laws, a dataset with all of this information could be difficult to come by, but its use would be interesting for this application. This could be incredibly valuable, as it would show the accuracy of the data based on new ground truth data.

B. Concluding Remarks

This paper has introduced a dataset with fourteen different parameters (thirteen inputs that create a diagnosis – the fourteenth feature). The training data consisted of 242 points and the test data consisted of 61 points; the total dataset was 303 points. Finding data that was known to be the ground truth was important for this research, so a dataset consisting of real patients' information was used. Support vector machine, logistic regression, and decision tree-based models were tested, analyzed, and compared for three through thirteen different principal components. Overall, results show that SVM and LR similarly had the highest accuracy throughout the number of principal components. Conclusions were made that though LR using all thirteen components had the highest overall accuracy, it may be more efficient to analyze the data using significantly less components, (a proposed value of eight was given), as the accuracy is not dramatically affected.

X. REFERENCES

[1] A. Islam, "Heart disease UCI - Eda and ML w/LR," *Kaggle*, 30-Jun-2020. [Online].

[2] "American Heart Association: To be a relentless force for a world of longer, healthier lives," *www.heart.org*. [Online].

[3] *Centers for Disease Control and Prevention*. [Online]. Available: https://www.cdc.gov/.

[4] W. Aha, *UCI Machine Learning Repository: Heart disease data set*, 07-Jan-1988. [Online].

[5] I. Sutedja, "Descriptive and Predictive Analysis on Heart Disease with Machine Learning and Deep Learning," 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS), 2021, pp. 1-6, doi: 10.1109/ICORIS52787.2021.9649585.

[6] J. hua, “What is Machine Learning?,” *Louisiana State University*. [Online].

[7] “Memorial Hermann Health System: Houston Hospitals, Institutes & Centers,” memorialhermann, 12-Apr-2022. [Online]. Available: https://www.memorialhermann.org/. [Accessed: Apr-2022].

[8] S. Rawat, “Heart disease prediction,” *Medium*, 28-Jun-2021. [Online].

[9] T. Burleigh, “Modeling the UCI heart disease dataset,” 20-Mar-2020. [Online].