Final Report: Analyzing Heart Disease Statistics

Aziz Shaik
Virginia Tech
azizs@vt.edu

Ashutosh Tiwari
Virginia Tech
ashutosh@vt.edu

Abstract

At the conception of this project, we set out to examine heart disease statistics provided by the University of California, Irvine in an effort to analyze and figure out the characteristics in patients that contribute the greatest to the likelihood of heart disease, and additionally, create accurate models to predict the chance of heart disease with health data [1]. We initially made use of a Naive Bayes classifier, and used techniques such as SMOTE and variance thresholds to various degrees of success with another Naive Bayes classifier, an SVM classifier, a logistic regression classifier, and a Perceptron classifier, with the reduced feature Naive Bayes having the highest rate of accuracy at around 89%, only marginally better than the full-feature Naive Bayes.

1 Full-Feature Naive Bayes Classifier

Initially, we constructed the Naive Bayes classifier, making use of all 13 features provided in the dataset. This provided a decent result, but attempts to improve the result of this classifier will be discussed in latter sections.

1.1 Data Pre-processing

There was a lot of pre-processing to do when implementing the Naive Bayes classifier, especially considering there were 13 features identified as potentially having an effect on a patient's likelihood of contracting heart disease. For the reader's convenience, the features are listed here, along with the shortened name in parentheses [2]:

- Age (age)
- Sex (sex)
- Chest Pain Type (cp)
- Resting Blood Pressure (trestbps)
- Serum Cholesterol (chol)
- Fasting Blood Sugar (fbs)
- Resting Electrocardiograph Results (restecg)
- Maximum Heart Rate Achieved (thalach)
- Exercise Induced Angina (exang)
- ST Depression Induced by Exercise Compared to Rest (oldpeak)
- Slope of the Peak Exercise Stress Test Segment (slope)
- Number of Major Vessels Colored by Flourosopy (ca)
- Results from Thallium Stress Test (thal)

Many of these features, such as age or chol, were continuous, and so we binned the data into quartiles assigning each quartile a value from 0-3 for easier calculations when it came to computing the conditional probabilities as dictated by the classifier.

Our data set had 303 data points, so we divided it such that about 85% was devoted to training the model while the remaining was used to test towards the end.

1.2 Classifier Results

The model returned a final test accuracy of 88% with the data prepossessing described above. Figure 1 shows the test accuracy over each test.

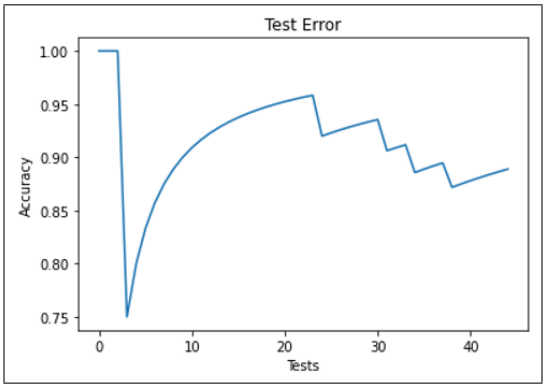

Figure 1: Test Accuracy of Full-Feature Naive Bayes Classifier

Additionally, one of the main issues that we wanted to solve in our proposal was to figure out which of the features contributed the most to deciding the likelihood of heart disease in a given patient. We found that a patient's fasting blood sugar, had the highest probability in approximately 86% of data points where the patient tested positive.

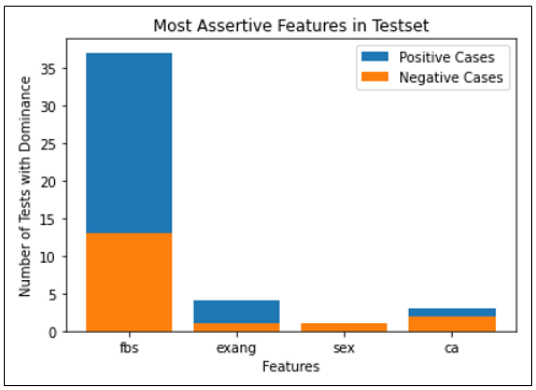

Figure 2: Features Most Influential in Classification

However, we dismissed this feature as a major source because both patients who tested positive and negative had similar rates of blood sugar over 120 mg/dl. The next most significant probability was whether or not the patient had exercise-induced angina, where an affirmative result for this increased the likelihood of heart disease. As seen in Figure 2, there were only four features that asserted themselves as the most influential. Most of these features were very common across all samples, whether a positive case or not, indicating that these features, especially blood sugar, did not play as much of a role in classification as it was present in most cases.

2 Variance Thresholding and SMOTE

2.1 Variance Thresholds

Based on the results of the previous section, we thought that reducing the numbers of features that we had could improve classifier performance. In order to do this, we measured the variance of each feature, with results shown in Figure 3.

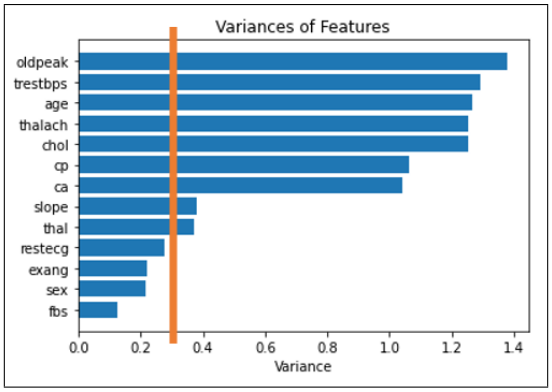

Figure 3: Variances of Each Feature

As seen in Figure 3, there is a large disparity between the lower six features and the higher ones, so we originally ran tests with classifiers with the top seven features with a variance threshold of 0.,6 but this resulted in low test accuracies. Therefore, we lowered the variance threshold to 0.3, so all features above the orange line in Figure 3 were used for classification. Note that of the four features that were removed, three of them were present in Figure 2.

2.2 SMOTE Analysis

We made use of SMOTE analysis from the imblearn Python library, and found that it did its purpose of increasing the training set size from 258 to 276, but failed in terms of improving the accuracy of the model. For example, when using the original, scaled dataset, our accuracy was around 82% with the SVM, whereas using the dataset with SMOTE applied, reduced it to around 71%. Upon doing further research, we found that SMOTE analysis works only for continuous data, which made sense since our data was binned and categorical, so we continued with our classifiers using our normal dataset.

3 Additional Classifiers

3.1 Reduced-feature Naive Bayes

We reduced our feature set and, as shown in Figure 4, the accuracy was still around the same at about 89%. This value is still much higher than when the variance threshold was at 0.6.

3.2 Perceptron Classifier

For the Perceptron classifier and the other ones, we used the StandardScaler from the sklearn Python library in order to standardize the values. In our case, we had a tolerance of 0.1 and also scale on each iteration by 0.1. This resulted in an accuracy of 84%.

3.3 SVM Classifier

For our SVM classifier, we added on cross validation in order to improve the classifier. We used a linear kernel function in the SVM, and with a squared hinge-loss function for the loss function. These decisions, along with a tolerance of 10^{-5} were made through trial and error, eventually resulting in a test accuracy rate of 82%.

3.4 Logistic Regression Classifier

The final classifier that we used was a logistic regression classifier, which proved to be a bit more challenging to decide by trial and error due to how many parameters there were. We ended up using a "Library for Large Linear Classification," or 'liblinear' algorithm as this seemed to be the most optimal for smaller datasets like the one we were working with. We allowed it to run for 100 iterations, and used L1 regularization for the penalty. Despite our efforts we were only able to get 80% accuracy, and that was with this configuration.

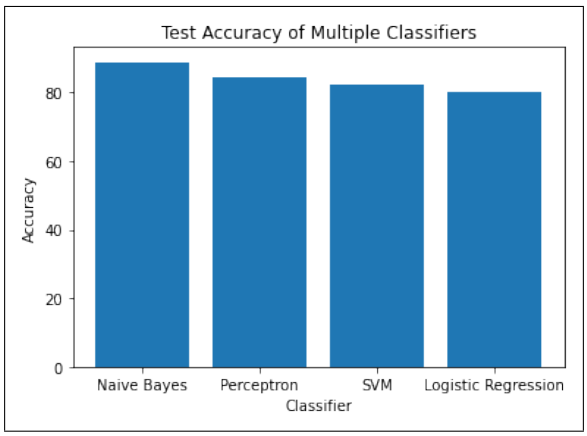

Figure 4: Test Accuracy of Each Classifier

4 Conclusion

Interestingly, machine learning for medical applications has long been a point of research, and one paper performed experiments similar to us, except they had a much larger sample size, and were interested in classifying cardiovascular disease [3]. The results of this paper were that Naive Bayes with a produced featureset produced the best indicator of cardiovascular disease, which was the same result produced by our experiments.

Another paper focused more on using Naive Bayes for medical problems, and argued that based on the nature of medical data mining, Naive Bayes was the optimal classifier for medical issues, even though it may have its flaws [4]. One of the key reasons for its success in this area is how it handles noisy data, which we observed in our dataset, where data points were inconsistent or leaned towards certain features more, with an uneven distribution.

Keeping in mind external research as well as the results of our own experiments, we were reasonably successful in achieving our goal of building a decent classifier to determine the diagnosis of a patient with around 90% accuracy, which can improve health services by reducing the necessity of invasive heart disease testing methods, or by understanding it earlier. We tested with three other classifiers to find that this one was the best. Another goal of this project was to determine the most impactful features, which we found to be age, maximum heart rate, and cholesterol, from our variance thresholds.

5 Contributions

In the programming portion of the progress made so far, Ashutosh Tiwari was responsible for setting up the environment, as well as managing the data and handling the parsing of the files. The data processing and classifier construction was primarily done by Aziz Shaik. The writing of the report was also a joint effort with Aziz responsible for much of the writing of it, while Ashutosh looked over and made revisions as necessary.

References

[1] Ronit. “Heart Disease UCI.” Kaggle, 25 June 2018, www.kaggle.com/ronitf/heart-disease-uci.

[2] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, “ Heart Disease Data Set,” Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease. [Accessed: 23-Oct-2020].

[3] P. Golpour, M. Ghayour-Mobarhan, A. Saki, H. Esmaily, A. Taghipour, M. Tajfard, H. Ghazizadeh, M. Moohebati, and G. A. Ferns, “Comparison of Support Vector Machine, Naïve Bayes and Logistic Regression for Assessing the Necessity for Coronary Angiography,” International Journal of Environmental Research and Public Health, vol. 17, no. 18, p. 6449, Sep. 2020.

[4] K. Al-Aidaroos, A. Bakar, Z. Othman, "Medical data classification with Naive Bayes approach," Information Technology Journal 2012, 11 (9), 1166.