Heart Attack Prediction

1st Anthony Spencer
dept. Computer Science
California State Polytechnic University, Pomona
Pomona, California
aspencer1@cpp.edu

2nd Arsham Mehrani
dept. Computer Science
California State Polytechnic University, Pomona
California, USA
amehrani@cpp.edu

3rd Sara Nersisian
dept. Computer Science
California State Polytechnic University, Pomona
California, USA
saran@cpp.edu

4th Michael Melkonian
dept. Computer Science
California State Polytechnic University, Pomona
California, USA
mamelkonian@cpp.edu

5th Vincent Verdugo
dept. Computer Science
California State Polytechnic University, Pomona
California, USA
veverdugo@cpp.edu

***Abstract*—According to the CDC, one person dies every 36 seconds from some type of cardiovascular disease and costs the United States about $363 billion each year. In this document we cover 5 different machine learning models to understand which method is best in predicting heart disease. Our data, Heart Failure Prediction Data-Set, was put through the following models: Random Forests, AdaBoost, Neural Networks, Support Vector Machine's, and Logistical Regression to obtain the highest overall accuracy of 88.4%.**

I. Introduction

Every year hundreds of thousands of people experience heart attacks. According to the CDC heart disease is the leading cause of death in America, killing almost seven hundred thousand people a year. That number will keep increasing with the rise of obesity in America. Heart attacks need to be treated fast in order to reduce the risk of being fatal. It's because of this necessity that doctors need a more proactive approach in identifying heart attacks victims. The CDC also states that about 20 percent of all heart attacks are silent. A silent heart attack occurs without the person feeling it happen, which means the damage is done and the person does not know it. Better more proactive methods of predicting heart attacks can solve both issues.

In this project we intend to use a publicly available data-set of patient data in order to train machine learning models to predict a heart attack. We plan to use three different machine learning algorithms (Neural Networks, Random Forrest, Support Vector Machines , Logistic Regression, and Ada Boost) and compare the results of all three. All three algorithms will be further discussed in methodology. Determining which algorithm is the most successful can help medical professionals understand which patients are most at risk, so the patient can be monitored accordingly.

Machine learning is breaking into almost every industry and improving performance, Health care is no different. We intend to build machine learning systems that can help a cardiologist. Much like the way machine learning is revolutionizing radiology with cancer detection, we hope this technology can be utilized to help drive down the death rate of America's leading cause of death. Please observe the conference page limits.

II. Data Set

For this project we are using "Heart Failure Prediction Dataset" [1] publicly available on the Kaggle website. This dataset is currently (Oct 2021) the largest heart disease dataset available which was created by combining five different international heart datasets. It contains around 1000 records providing the patients' demographic and medical information. There are 12 attributes:

1) Age: [years]
2) Sex: [M/F]
3) ChestPain: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4) RestingBP: resting blood pressure [mm Hg]
5) Cholesterol: serum cholesterol [mm.dl]
6) FastingBS: fasting blood sugar [1: if FastingBS ¿ 120 mg/dl, 0: otherwise]
7) RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8) MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]

9) ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10) Oldpeak: oldpeak = ST [Numeric value measured in depression]
11) ST-Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12) HeartDisease: output class [1: heart disease, 0: Normal]

After some exploratory data analysis, there were some notable observations. The dataset mostly consists of males, possibly leading to some bias. Another thing to note is that despite the fact that there are mostly males in the dataset, the proportion of people having heart disease for the entire dataset is relatively balanced. Unsurprisingly, the proportion of men suffering from heart disease exceeds that of women.

Upon observing the Age feature of the dataset, it appeared to have a normal distribution. The ages range from 30 to 75, with the majority of instances being in the age range of 45 to 65. A question that naturally arises is how age is correlated to heart disease. As expected, as a person's age increases, the higher the chance of them getting heart disease. This was most often true when the age of the individual exceeded 55. From this same observation, it can be said that no one in this dataset under the age of 30 is likely to get heart disease, whereas older men have the highest chance of obtaining it.

Lastly, after creating a histogram of maximum heart rate in those who are likely vs. not likely to get heart disease, it was apparent that the maximum heart rate of people with heart disease appeared to be below 140. This was one of the most important visual depictions in this dataset. Showing us that people without heart disease seem to have a higher heart rate, likely because they are more active individuals. The probability of active people getting heart disease is lower than inactive people.

III. Methodology

We use five machine learning algorithms: Random Forests, Neural Networks, Ada boost, Support Vector Machines, and Logistic Regression. All of which are supervised machine learning algorithms. We compare the prediction of each method for all the instances and measure the accuracy.

A. class label

The goal is to predict a class label "likeliness of heart attack" for each instance/person based on the given background data. The prediction is going to be a binary value of 1 = likely to have a heart attack, and 0 = not likely to have a heart attack. The data set is labeled so the prediction is compared to the class label for accuracy. At the end we are also going to take the majority vote on each prediction from each method to see if combining these methods together provides a better accuracy overall.

B. training vs testing

For each algorithm the data set is split 70% for training the model and 30% for testing. Every time the data used for each algorithm is shuffled so the models do not simply memorize the data set. Also we created a standard 70-30 split data set without shuffle for each model to train on at least once. This control data insured that every model has the same conditions. The difference between the accuracies obtained from each model is a result of the model itself not the data.

IV. Results

A. Neural Networks

The main focus for training the Neural Network was fine-tuning the architecture to extract the best accuracy possible. The best accuracy reached was 87.84%, with an average accuracy of 83.27%. With each epoch, the accuracy scores changed due to shuffling the data set and randomizing initial weights for a better trained model. After experimenting with ranges of epochs, the model now uses a total of 1175.

This model's architecture consists of 11 densely connected input neurons with a ReLU activation function. There are a total of four hidden layers, each with 20 densely connected neurons with a ReLU activation function. And finally, the output layer consists of a single densely connected neuron with a linear activation function.

The model was compiled using the mean squared error loss function and adam optimizer, and the learning rate was 0.001. Each of these factors was fine-tuned for maximum accuracy possible while keeping the time spent training minimal.

B. Random Forests

Random Forest is a popular supervised machine learning algorithm. The technique considers multiple decision trees and outputs the highest vote obtained by all the trees. In this research Random Forest performed well. The highest accuracy score reached by the RF model was 87.31%. The model was trained with variety of max_depth (maximum depth of each tree in the forest) and n_estimators (the number of trees to be used in the forest) values ranging from 5 to 1000. Ultimately, the best settings producing the highest accuracy was obtained when max_depth was in default mode (having value of NONE, meaning each tree is expanded until every leaf is pure) and n_estimators was 20.

C. Support Vector Machines

With the Heart Failure Prediction data-set extracted from Kaggle we were able to train and test through a variety of "setting scenarios." In other words, a variation of C values (ranging from sizes of .5 incrementally to 20); variations of degree value (1 to 3), kernel type (linear, poly, and rbf), and finally decision function shape (ovo and ovr) were tested in every possible combination.

This in conjunction to the use of 80/20 and 70/30 split data ultimately is what lead towards a maximum accuracy of approximately 84.06%. It is important to note that SVM's work relatively well with data that is linearly separable. This was not necessarily the case when observing most of the features which only allowed for two features to really work well in creating a linear SVM plot. Overall, the SVM model yielded one of the lowest accuracies among the 5.

D. Logistic Regression

Logistic Regression is a supervised learning classification algorithm used to predict the probability of a target variable. This target variable is binary, which makes this algorithm a good choice for our dataset. With the base Logistic Regression set to 1000 max iterations, an accuracy of 85% was immediately achieved. One of the reasons for this is likely because the features are mostly independent of each other. The sample size was not the greatest, but it does seem sufficient. Utilizing hyperparameter tuning with Scikit-learn's GridSearchCV method, some more optimal parameters were obtained for this model. After using these newfound parameters, the accuracy improved to 88%. After graphing the ROC curve for this model, it was found that the AUC (Area Under Curve) was 95%, which is an excellent result indicating that the model will likely not get much better from here.

E. Ada Boost

At first for our Ada Boost implementation, we used the base estimator for Ada Boost, which is a decision tree with max depth of 1. By implementing a grid search with the number of estimators and the learning rate we were able to reach an accuracy of 85.6%. This was achieved by testing the number of estimators from 2 to 1000 and leaning rate from .01 to 1. The optimal settings that achieved the highest accuracy were estimators set to 1000 and the learning rate set to .01. After looking into random forest, we decided to try a different method to test and see if we get better results. Changing the method, we decided to implement a random Forrest with Ada boost. Keeping the same grid search from the first test and adding max depth from 1 to 100 and bootstrap True/False we managed to achieve our highest accuracy yet at 88.4%. The correct settings for this accuracy would be number of estimators set to 50, max depth set to 100, learning rate set to 1 and bootstrap set to False. With this combination our model was able in increase the accuracy of out first attempt at Ada Boost by 3%. Combining methods yielded positive results and possibilities for future research in the subject.

V. Related Work

In our research, we have found the article and experiments done by Davide Chicco and Giuseppe Jurman.[3] In an article titled "Machine learning can predict survival of patients," The authors explain whereby using a similar data set to the one presented in this paper, they were able to extract the two crucial features of serum creatinine and ejection fraction. By using only these features, they obtained high accuracy with their model. However, the data set only included 300 samples which is less than a third of the samples used in our experiments.

The authors implemented ten different machine learning algorithms to predict patient survival. The classifiers included Linear Regression, Random Forests, Decision Tree, Artificial Neural Network (perceptron), two Support Vector Machines (linear, and with Gaussian radial kernel), an ensemble boosting method (Gradient Boosting), and more.

This paper closely follows our experiment, including the use of algorithms such as Neural Networks and Ada Boost. The authors used grid search to find the models that generate the highest Matthews correlation coefficient. Our team has also used a variety of machine learning algorithms. Still, due to a lack of resources on our end, we could only experiment with the five previously mentioned algorithms. Finally, the results closely match our results with an accuracy of close to 80% for the majority of the algorithms.

VI. Conclusion

We compared the classification results of five machine learning algorithms: SVM, Neural Network, Random Forest, Logistic Regression and Ada Boost. Due to clean data, all the models obtained by each algorithm had high accuracy of above 80%. However, Ada Boost provided the most accurate model having accuracy of 88.4% We attempted to take it further and ensemble (SVM, Random Forest, Logistic Regression and Ada Boos), but this did not lead to a higher accuracy. This model was set up giving one vote each algorithm leaving us to believe the algorithms were getting the same incorrect predictions on the same instances. Possibly this means there is a unknown attribute we are missing from our data, or that we need more data to better our accuracy.

VII. References

[1] Federico Soriano, *Heart Failure Prediction*, Kaggle, 2021. [Dataset]. Available: https://www.kaggle.com/fedesoriano/heart-failure-prediction. [Accessed: Oct 11,2021]

[2]"Heart disease facts," Centers for Disease Control and Prevention, 27-Sep-2021. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm. [Accessed: 12-Oct-2021].

[3]D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Medical Informatics and Decision Making, 03-Feb-2020. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5#Sec8. [Accessed: 02-Dec-2021].

VIII. Git link

https://github.com/Arsham1024/ML_Project