Cardiovascular Disease Prediction

1st Jaykumar Patel
Data Science
Stevens Institute of Technology
NJ, USA
jpatel5@stevens.edu

2nd Juilee Thakur
Data Science
Stevens Institute of Technology
NJ, USA
jthakur3@stevens.edu

3rd Samyak Upare
Data Science
Stevens Institute of Technology
NJ, USA
supare@stevens.edu

***Abstract*—The prevalence of cardiovascular disease is rising quickly every day, making it crucial and alarming to anticipate such illnesses. It is challenging to make this diagnosis since it must be done quickly and precisely. Based on a number of medical criteria, the main focus of this research is on which patients have a higher risk of developing cardiovascular disease. We have created a technique to predict the likelihood that a patient would be diagnosed with cardiovascular disease based on their medical history. In order to forecast and categorize cardiovascular disease patients, we applied a variety of machine learning methods, including logistic regression. We modified the model's application to enhance the predictability of a heart attack for each individual using a very helpful approach.**

keywords - cardiovascular, logistic regression

I. Introduction

Cardiovascular disease has gotten a lot of attention in medical research among the various disorders that are life-threatening. A difficult and expensive task, the diagnosis of cardiovascular disease. This program automatically forecasts the patient's heart status and allows for additional treatment. Signs, symptoms, and a physical examination of the patient are frequently used to make the diagnosis of cardiac disease. To anticipate the severity of the patient's chance of acquiring cardiovascular disease, we would like to gather data pertaining to every aspect of our field of study and train the data in accordance with the machine learning algorithm suggested by. You should do it. It is frequently used to categorize whether a patient has cardiovascular disease or is healthy using data mining techniques like logistic regression. The suggested system can extract precise hidden data patterns, relationships, and patterns related to cardiovascular disease from historical cardiovascular disease databases. It can also provide complex solutions to problems involving cardiovascular disease diagnosis. Making wise clinical decisions could therefore be advantageous for healthcare professionals.

II. Related Work

Mohan et al states that in order to process the raw healthcare data and provide a fresh and original discernment towards heart disease, machine learning techniques were used. If the disease is discovered in its early stages and preventative measures are implemented as soon as possible, the mortality rate can be significantly reduced. The characteristics of the suggested hybrid Random Forest (RF) and Linear Method technique are combined (LM). When it came to predicting heart disease, HRFLM showed to be quite reliable [1].

When combined with PCA, alternating decision trees have demonstrated exceptional performance, however in some other situations, decision trees have demonstrated exceptionally poor performance, which may be caused by overfitting. Because they employ numerous algorithms to address the issue of overfitting, Random Forest and Ensemble models have fared very well. When combined with PCA, alternating decision trees have demonstrated exceptional performance, however in some other situations, decision trees have demonstrated exceptionally poor performance, which may be caused by overfitting. There is still much need for study about how to manage high dimensional data and overfitting, however Random Forest and Ensemble models have done quite well since they address the issue of overfitting by using many algorithms [2].

Srinivas et al states that based on the estimated significant weightage, an effective method was developed for the extraction of significant patterns from heart disease data warehouses for the valuable prediction of heart attack. The frequent patterns with a value larger than a predetermined threshold were selected. The definition of three mining objectives is based on data exploration. All of these models were capable of providing complicated predictions for heart attacks[3].

III. Our Solution

For now we have performed Logistic Regression on the data set.

A. Description of Dataset

The data set that we are planning to use is from UCI Machine Learning Repository that contains 14 physical attributes based on the physical testing of a patient. In addition to having blood drawn, the patient also undergoes a quick exercise test.

The 14 attributes are:

age: age in years

sex: sex (1 = male; 0 = female)

cp: chest pain type
Value 1: typical angina
Value 2: atypical angina
Value 3: non-anginal pain
Value 4 asymptomatic

trestbps: resting blood pressure (in mm Hg on admission to the hospital)

chol: serum cholesterol in mg/dl

fbs: (fasting blood sugar >120 mg/dl) (1 = true; 0 = false)

restecg: resting electr-cardiographic results
Value 0: normal
Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV)
Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

thalach: maximum heart rate achieved

exang: exercise induced angina (1 = yes; 0 = no)

oldpeak = ST depression induced by exercise relative to rest

slope: the slope of the peak exercise ST segment
Value 1: upsloping
Value 2: flat
Value 3: downsloping

ca: number of major vessels (0-3) colored by flourosopy

thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

num: diagnosis of cardiovascular disease (angiographic disease status)
Value 0: <50% diameter narrowing
Value 1: >50% diameter narrowing

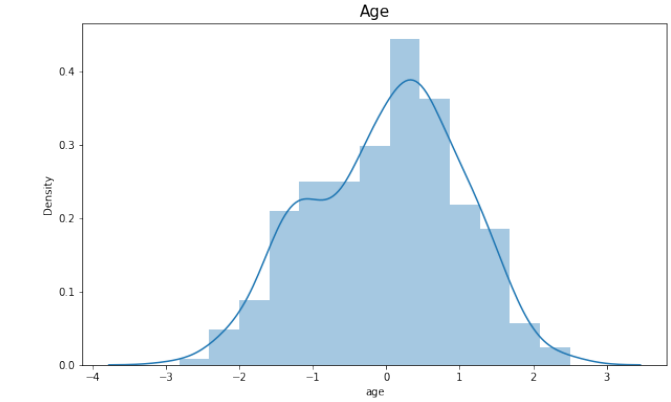
(in any major vessel: attributes 59 through 68 are vessels) The website has data obtained from Hungarian Institute of Cardiology, University Hospital (Zurich), University Hospital (Basel) and V.A. Medical Centre, Long Beach and Cleveland Clinic Foundation.

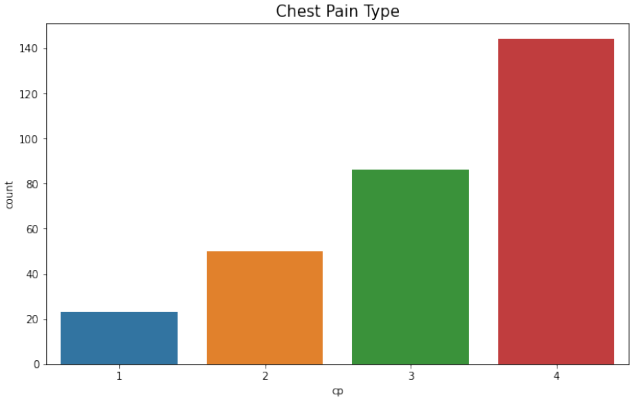
B. Machine Learning Algorithms

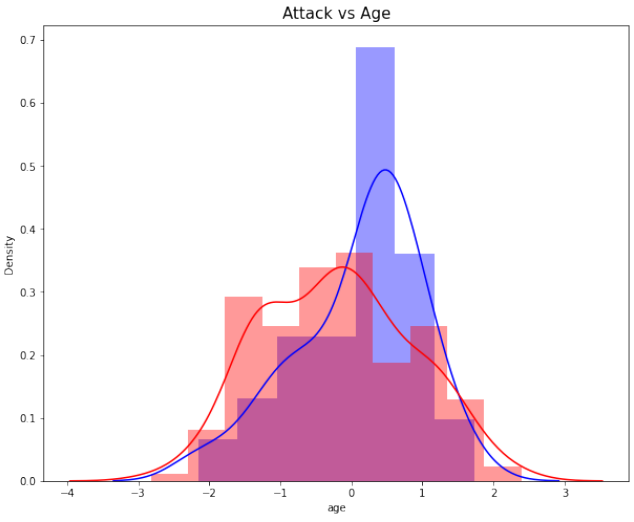
1) Logistic Regression: With the use of certain criteria, this proposed method can classify patients as having cardiovascular disease or not. This information might be used by the suggested system to build a model that forecasts if a patient has this illness. The suggested method makes use of the Sklearn library to generate the score using a logistic regression technique. Recent research have shown that the machine learning discipline of logistic regressions is the most well-liked and rapidly developing. A supervised learning classification approach called logistic regression is used to forecast the likelihood of a target variable. Data are described and the link between dependent binary variables and one or more nominal, order, interval, or ratio-level independent variables are explained using logistic regression.

C. Implementation Details

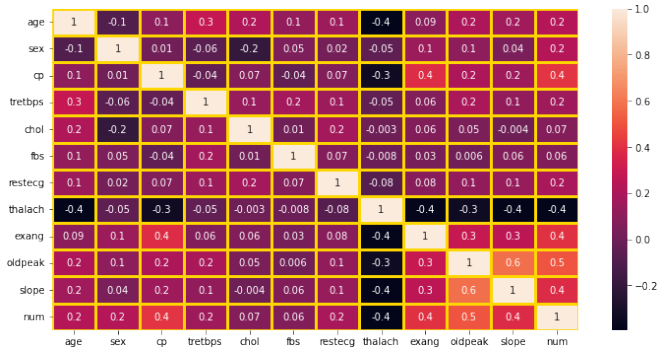
This project was carried out in a total of four sections. Pre-processing is the first step, during which all the missing and pointless data is eliminated. Some features were being incorrectly shown as a result of missing values. This issue was taken into account, and all the missing or null values were successfully erased. For the sake of future algorithm implementation, some of the independent variables were not relevant. Then, using the provided dataset, we examine our feature data. The dataset also includes some continuously available data. Therefore, we use normalization techniques to scale down characteristics to a standard value. (We solely used approaches for normalizing continuous features.) To better comprehend the dataset, we plot several graphs over the features.







To display the correlations between the dataset's mean variables, we generated a heatmap



According to our Logistics regression model we get 82% accuracy.

```
print(accuracy_score(y_test, y_pred_lr) * 100,'%')
```

```
81.66666666666667 %
```

IV. Comparison

There is now only one algorithm in use, logistic regression, and it has been fully implemented. As a result, we are unable to compare their accuracy at this time, but we will be able to do so once other algorithms have been included.

V. Future Directions

Two additional machine learning algorithms will be used in the future, according to our plans.

VI. Conclusion

We are currently working to put our machine learning algorithms into practice. We presently use logistic regression, but we'll switch it out for more precise machine learning algorithms. We are not finished yet, thus we are unable to offer a thorough conclusion at this time.

References

[1] Senthilkumar Mohan, Chandrasegar Thirumalati, Gautam Srivastava,"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access.

[2] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja, "Heart disease prediction using machine learning techniques: a survey", International Journal of Engineering Technology, 7 (2.8) (2018) 684-687

[3] K.Srinivas, Dr.G.Raghavendra Rao, Dr. A.Govardhan, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", The 5th International Conference on Computer Science Education Hefei, China, August 24–27, 2010.