Heart Attack Risk Prediction Using Machine Learning

Submitted by:
Sweety Kumari

Contents:

Introduction

Heart attack risk predictor is an online platform designed and developed to explore the path of machine learning . The goal is to predict the risk of heart attack in a patient from collective data, so as to be able to detect configurations at risk for the patient, and therefore, in cases requiring emergency medical assistance, alert the appropriate medical staff of the situation of the latter.

By analyzing the data we can predict the risk of heart attack in our project. Machine learning algorithms can also be helpful in providing vital statistics, real-time data and advanced analytics in terms of the patient’s disease, lab test results, blood pressure, family history, clinical trial data, etc., to doctors.

Motivation

The fact that human life is dependent on the proper functioning of the heart is the driving force behind this research. The heart is a crucial part of our bodies, and heart disease has become the leading cause of death globally. According to the WHO, an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

Objectives

- **Problem Statement**
  - Predicting the risk of Heart Attack using Machine Learning Classifier

- **Challenges**
  - Data collection
  - Irregularities in data

Tools and Technologies

- **Language**
  - Python
- **Libraries**
  - sklearn
  - borutapy
  - matplotlib
- **Platform**
  - Google Collab

Literature Survey

- I Ketut Agung Enriko, Muhammad Suryanegara, Dadang Gunawan “Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters” Google Scholar | Publisher Site

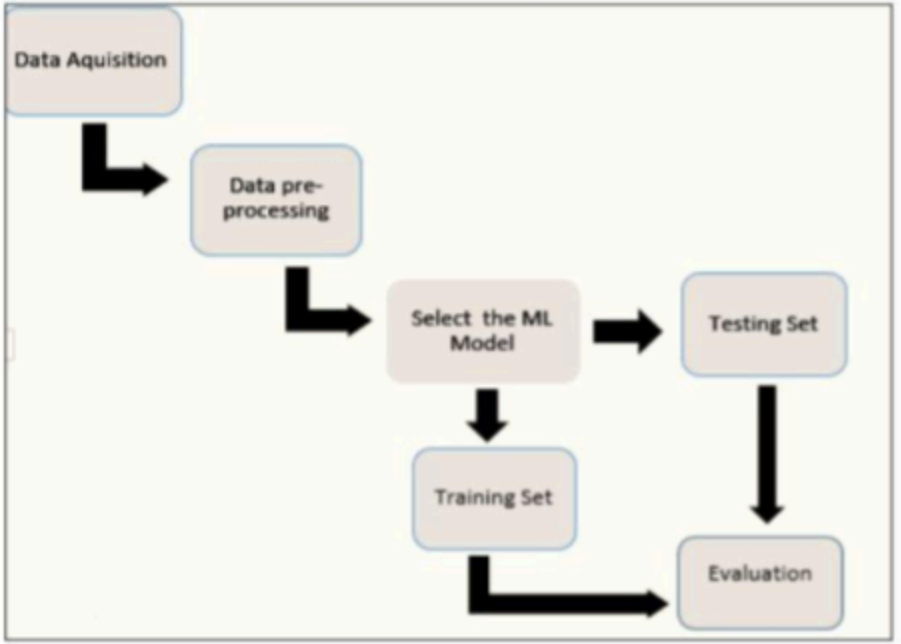
- M. Akhil jabbar, B.L.Deekshatulu, Priti Chandra “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm” Google Scholar | Publisher Site

- Gongde Guo, Hui Wang, David Bell , Yaxin Bi , and Kieran Greer “KNN Model-Based Approach in Classification” Google Scholar | Publisher Site

- Han, J., Kamber, M., & Pei, J. “Data mining: concepts and techniques”. Elsevier (2011) Publisher Site

Methodology

- Description of the dataset
- Preprocessing of the dataset
- Machine Learning classifier proposed
- Evaluation process used



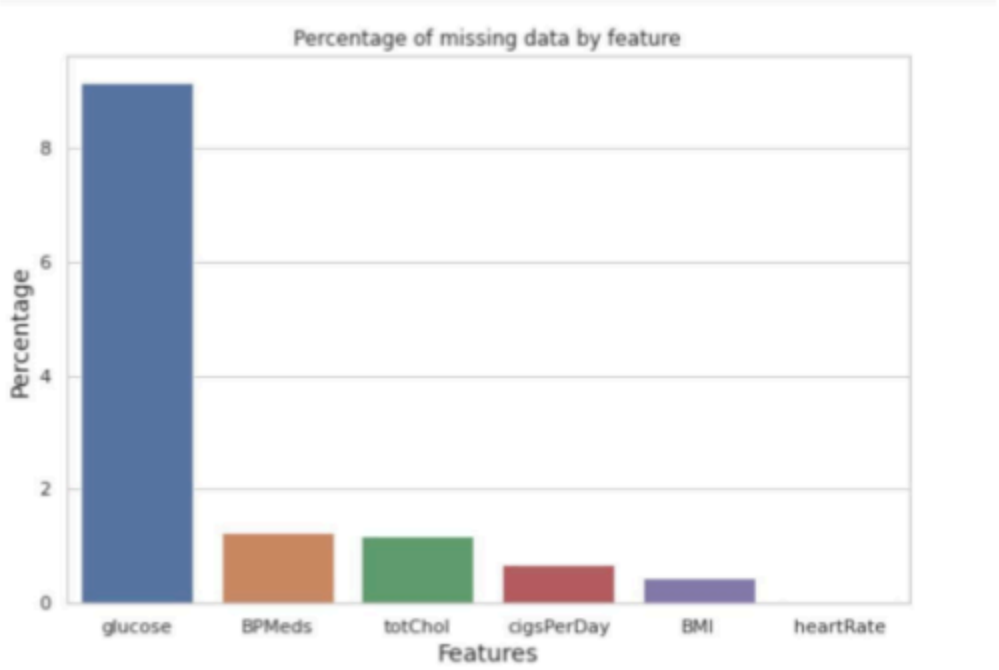
Dataset

The data set is publicly available on the biolincc website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

- Age
- Sex
- Total cholesterol
- Current Smoker
- Cigarettes Per Day
- BP Meds
- Glucose
- Prevalent Stroke
- Prevalent Hyp
- Diabetes
- Heart rate
- BMI (Body Mass Index)
- Dia BP (diastolic blood pressure)
- Sys BP (systolic blood pressure)

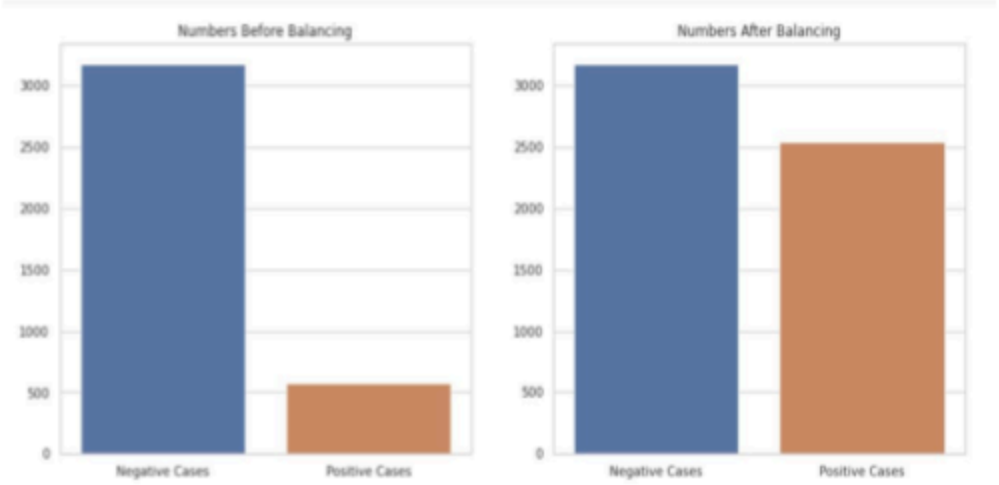
Preprocessing of Dataset

Here we have checked and dealt with missing and duplicate variables from the data set as these can grossly affect the performance of machine learning algorithm (many algorithms do not tolerate missing data).



Model development

- In our case, the number of negative cases greatly exceeds the number of positive cases.
- To address this problem. I balanced the data set using the Synthetic Minority Oversampling Technique (SMOTE).



Feature Selection

- Having irrelevant features in a data set can decrease the accuracy of the models applied, we used the Boruta Feature Selection technique to select the most important features which were later used to build our model.
- The Boruta Feature Selection algorithm which is a wrapper method built around the random forest classification algorithm. It tries to capture all the important, interesting features in a data set with respect to an outcome variable.
- After running the algorithm for 100 iterations the top selected features were: Age, total cholesterol, systolic blood pressure, diastolic blood pressure, BMI, heart rate and blood glucose.

Important Features

Most important features were found as age and sys BP

Top Features

- age, sys BP, BMI
- total cholesterol
- diabetes
- glucose
- heart rate
- cigs per day
- Prevalent Hyp
- sex
- current smoker
- diabetes
- BP meds
- Prevalent stroke

Machine Learning Classifier

- K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint.

- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

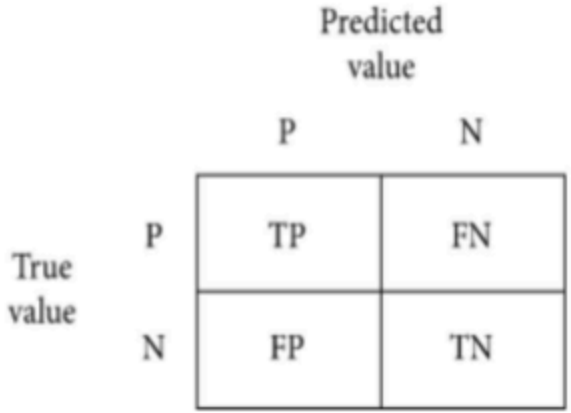
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Evaluation Process

For the evaluation process, confusion matrix, accuracy score, sensitivity and specificity are used. A confusion matrix is a table-like structure in which there are true values and predicted values, called true positive and true negative.

In Figure,

P = positive,
N = negative,
TP = true positive,
FN = false negative,
FP = false positive,
TN = true negative.



Evaluation Process Cont.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

$$\text{specificity} = \frac{TN}{TN + FP}.$$

$$\text{sensitivity} = \frac{TP}{TP + FN}.$$

Specifity is a measure of how well a classifier identifies negative cases.

Sensitivity is the proportion of actual positive cases that got predicted as positive (or true positive).

Result and Discussion

Accuracy = 84.45%.

Specificity = 76.51%.

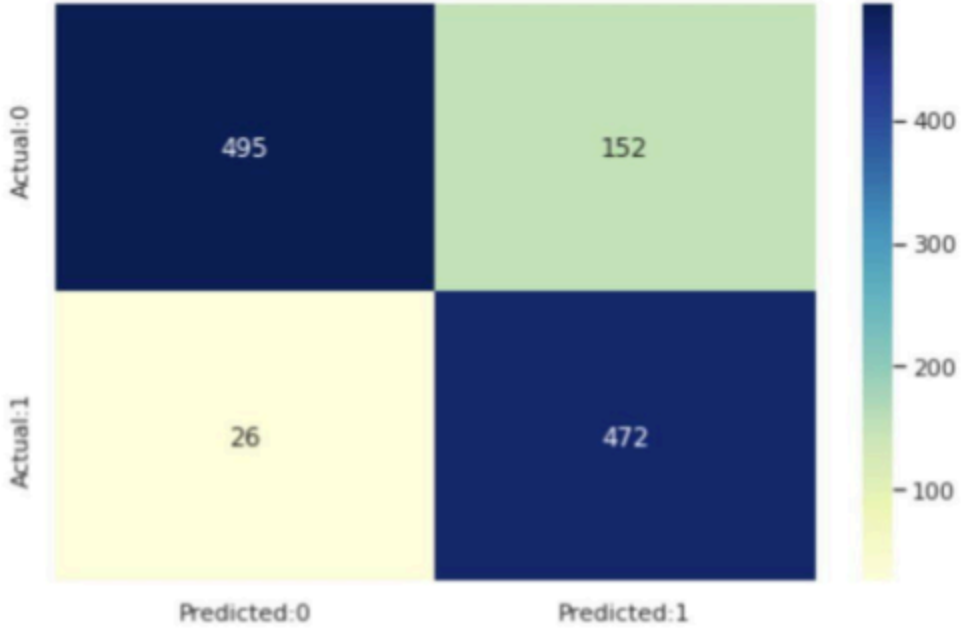
Sensitivity = 94.77%.

True Negative = 495

False Positive = 152

False Negative = 26

True Positive = 472



Future Scope

Today’s, world most of the data is computerized, the data is distributed and it is not utilizing properly. By analyzing the available data we can also use for unknown patterns. The future scope of the project is the prediction of heart attack by using advanced techniques and algorithms in less time complexity. The work can be enhanced by developing a web application based on the algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting heart attack effectively and efficiently.

Conclusion

With the increasing number of deaths due to heart attack, it has become mandatory to develop a system to predict heart attacks effectively and accurately. The motivation for the study was to find the most efficient way for the detection of heart attack. This study can then be used as a simple screening tool and all that we need to do is to input ones: age, BMI, systolic and diastolic blood pressures, heart rate and blood glucose levels after which the model can be run and it outputs a prediction.

References

[1]https://www.ijeat.org/wp-content/uploads/papers/v8i3S/C11410283S19.pdf

[2]https://www.nature.com/articles/s41598-020-76635-9

[3]https://www.datasciencecentral.com/select-important-variables-using-boruta-algorithm/

[4] https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learnin g/

[5]https://www.ijeat.org/wp-content/uploads/papers/v8i3S/C11410283S19.pdf

Thank You