Classification Analysis on Heart Attack Analysis on Patients Dataset

Saiprakash A[1], Ananth Kumar A[2], Deepak S[3], Praghash M[4], Anbarasan Balakrishnan[5], M.S.Sassirekha[6]
[1,2,3,4]*Department of Computer Applications Thiagarajar College of Engineering Madurai, India*
[5]*Director - Engineering, Capgemini Engineering, Shahul Hammed S Department of Computer Applications Thiagarajar College of Engineering Madurai, India*
[6]*Assistant Professor Department of Computer Applications Thiagarajar College of Engineering Madurai -15, India*

Abstract **- The Dataset contains the data about medicinal records of Patients who are diagnosed with either Heart Attack or Not. Heart Attack occurs when a blood clot blocks the blood flow to the part of the Heart. It is researched as, more persons are prone to Heart Attack during the ages between 40 – 50 years of Age. Let's do classification techniques on the dataset to create prediction models that can predict whether a patient would have a Heart attack using their medicinal records.**

Index Terms **– Classification, Logistic Regression, Decision Tree Classification, Random Forest Classification, Support Vector Machine, Normalization.**

I.INTRODUCTION

Given Problem is a classification problem since the solution is to find out whether a patient has more possibility of getting a Heart attack or not. This Project intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using machine learning. The machine learning model predicts the likelihood of patients getting a heart disease trained on dataset of other individuals. As the result, the probability of getting a heart disease based on current lifestyle and diet is calculated. Primary Goals of the analysis are: (1) To try to understand more about the Dataset and its attributes (2) To figure out relationships between the Attributes of the Dataset and Target of the Dataset. (3) To find out inferences regarding Heart Attack prediction from the Dataset. (4) To find out the Accuracy score for different Classification models

II. METHODOLOGY

A. Examining the Dataset

There are 14 columns and 303 rows in the Dataset. All the columns has integer values in it. No Columns possess null values. Target is our target variable and rest of them are independent variable. Column values of the given dataset are: age(Age of the atient), sex (Sex of the patient), cp (chest pain type), trestbps (resting blood pressure), chol(serum cholesterol in mg/dl), fbs(fasting blood sugar), restecg(resting electrocardiographic results), thalach(maximum heart rate achieved), exang(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest), slope(the slope of the peak exercise ST segment), ca (number of major vessels (0-3) colored by flourosopy), thal(thal: 0 = normal; 1 = fixed defect; 2 = reversable defect), target(0= less chance of heart attack 1= more chance of heart attack)

B. Data Cleaning

Dataset does not possess any Null values or redundant values. All the Dataset variables are needed for the classification so we cannot remove any rows or columns in the dataset.

C. Visualization

The Visualization techniques are used to visualize the dataset and find out any relationships between the attributes and understand more about the Dataset. Following are the Inferences from the Visualization techniques: (1) Dataset contains more Male candidates than Female Candidates. (2) Age Attribute of the people is widely distributed. (3) Target has more 1's than 0. That is, we have more records of patients who had Heart Attack. (4) Patients with

Cholesterol level 200 – 300mg/dl were highly considered.

D. Data Pre-Processing

All the values in the dataset are integer values. No string values are used to represent a classification of data. So there is no need for changing any values to integer values. (All the Categorical and numerical variables are integer values in the Dataset).

But the dataset has both Categorical and Numerical variables. And the range of Numerical variable values are random and cannot be compared with each other. The Numerical variables are "age", "trestbps", "chol", "thalach", "oldpeak".

So we normalize the dataset where all the values will be in the range from 0 to 1. All the Categorical variables are either 0 or 1. We scale the range of every variable to 0 to 1.

E. Training and testing split

Before splitting the Data for training and testing, we have to assign target variable to Y and Predictor Variables to X. We have to split the dataset into 80:20 ratio where 80% of the Dataset will be used for training the Models and 20% will be used for testing the models.

F. Performing Classifications

The Above Training dataset will be used for following algorithms to create Machine learning models: (1) Decision Tree classification (2) Random Forest Classification (3) Logistic regression (4) Support vector Machine.

After creating machine learning models with these classification techniques, 20% of the original dataset also known as the testing dataset will be used for finding out the accuracy of these models.

III. RESULTS

Classification Technique	Accuracy
Decision Tree Classification	77.27%
Random Forest Classification	80.16%
Logistic Regression	80.16%
Support Vector Machines	78.51%

IV. USING THE TEMPLATE

These are the discussions made after analysing the dataset:

- Male patients are more prone to get a Heart Attack than Female patients.
- Patients with Fasting sugar level less than 120mg/dl has more possibility of having Heart Attack.
- Patients below the age of 35 has lesser possibility of having a Heart attack.
- Patients between the age of 50 and 60, has higher possibility of getting a heart attack.
- Patients with Cholesterol level of 150 – 350 mg/dl are having the more possibility of having a Heart Attack
- Patients with “Thal” value 0 has the least possibility
- of having a heart attack.
- Patients with “Oldpeak” value 0 has the highest possibility of having a heart attack
- Chest Pain Attribute is not consistent with values of Target Variable

V. CONCLUSION

We have analysed and got to know some of the important medical attributes that may lead to Heart Attack possibility. We have used different Learning models, out of which, both logistic regression and Random forest classification seems to be the best and give more accurate predictions. All the Columns seems vital and removing any of them may reduce the model accuracy.

REFERENCES

[1] https://www.kaggle.com/anshumansharma002/heart-attack

[2] Prediction of Heart Attack Using Machine Learning By Akshit Bhardwaj, Ayush Kundra, Bhavya Gandhi, Sumit Kumar, Arvind Rehalia, Manoj Gupta, Department of Instrumentation & Control Engineering, Bharati Vidyapeeth’s College of Engineering Delhi-110063.