Heart Disease Detection using Machine learning

Shubham Gupta M190718CS
Tejaswinee Langhe M190737CS
Dept. of Computer Science and Engineering
National Institution of Technology, Calicut.

Abstract- We have chosen a problem of detecting the chance of heart disease based on various factors and lifestyle choices. The main motive of this project is to use machine learning model prediction techniques and methods for detecting heart diseases. Initially, we are making this project using KNN algorithm (k-Nearest Neighbours), but we extended it on others algorithm too for more accuracy like Logistic Regression, Support Vector Machines, etc

Keywords-Machine learning, Data analysis, Bioinformatics, KNN, Logistic Regression, Support Vector Machine etc.

I INTRODUCTION

With the rampant increase in the heart stroke rates at juvenile ages, we need to put a system in place to be able to detect the symptoms of a heart stroke at an early stage and thus prevent it. It is impractical for a common man to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of a heart disease.[1] Thus we propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc and also some life style choices. The machine learning algorithms have been proven to be the most reliable and hence we are using it in the proposed system.

II PROBLEM STATEMENT

Problem statement is “Given a set of clinical parameters about a patient, the system will predict whether the patient has heart disease or not”.

We are using “UCI Heart Disease Dataset” [2] as our dataset input. This is a well-known data set used by various researcher in machine learning to predict and analyse data in the field of cardiology. UCI Heart Disease Dataset has 303 instances and 14 attributes with no missing values. This dataset will be our input for algorithm to function and present us the vulnerability of heart disease with accuracy value.

III LITERATURE SURVEY

For detection of chances of heart disease various data minding algorithms are used. These data mining algorithms include Naïve Bayes, K-means, Support Vector Machine, Simple Logistic Regression, Random Forest & Artificial Neural Network (ANN) etc. By using several cardiovascular system parameters such as age, blood pressure, ECG results, sex, and blood sugar, it is possible to measure the possibility of getting affected by heart disease. Few of the data mining algorithm are explained as follows to get the insights of the techniques of detection of heart diseases.

a) K-means Algorithm

K-means creates k groups from a set of given objects so that the members of a group are more similar. Other than specifying the number of clusters, k-means also “learns” the clusters on its own. That’s why k-means can be called as semi-supervised learning method. K-means is especially effective over large datasets.[3]

b) Artificial Neural Network (ANN)

An artificial neural network (ANN) is a computational model based on the structure and functions of biological neural networks. Information which flows through the network affects the structure of the artificial neural network because a neural network changes or learns in a sense-based on input and output, for that particular stage and consequently, for each stage. ANN’s are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modelled or patterns are found. ANN's have layers that are interconnected. Artificial neural networks are fairly simple mathematical models to enhance existing data analysis technologies.[3]

However, Machine learning is proven more accurate for detection of the vulnerability as machine learning can look at patterns and learn from them to adapt behaviour for future incidents. So, our proposed system makes use of Machine Learning for detecting the vulnerability of heart disease.

IV PROPOSED METHOD

Our model has 3 phases.

1-Exploratory Data Analysis (EDA) -

The goal here is to find out more about the data and become a subject matter export on the dataset we are working with.

1. What question(s) are you trying to solve?
2. What kind of data do we have and how do we treat different types? Etc.
3. How can you add, change or remove features to get more out of your data. Etc
4. What's missing from the data and how do you deal with it?
5. Where are the outliers and why should you care about them?

2-Data Pre-processing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Main aim of Data Pre-processing is to remove unnecessary values, so that it can be given as input to your machine learning algorithm. For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values.

3- Training dataset and Applying Machine Learning Algorithm

This is the main part of the project. Here dates set is divided into 3 parts, Training, testing, and validation. We will train our model using Training data set. Ratio of training to testing dataset is 80:20. Using that we test our model using testing dataset and validating. Initially, we are making this project using KNN algorithm (k-Nearest Neighbours), but we extended it on others algorithms for more accuracy. We added Logistic Regression and Support Vector Machines algorithm to get more accuracy. More the accuracy more is the accuracy of the results we can get. After that we make model for general input.

V WORK PLANS

We are planned to complete this project in four phases.

1) Data Analysis:

Based on the dataset, we are deciding the kind of clusters we are going to need for prediction, their limits and how the overall differentiation of dataset is going to be.

2) Data Pre-processing:

This includes bringing the data in the form which is required by the machine learning algorithm.

3) Building an actual system:

This includes building and implementation of machine learning algorithm to detecting chances of heart diseases based on training of dataset.

4) Testing and Documentation:

This includes testing the system in different scenarios and cases and completing the final report

VI RESULTS

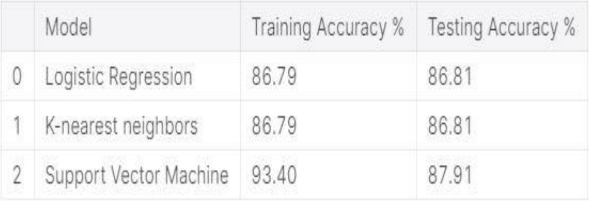
As we have implemented three algorithms on the given dataset to get more accuracy -

(i) Logistic Regression,

(ii) K-Nearest-Neighbours

(iii) Support Vector Machine

The behaviour of each algorithm on the given dataset is as mentioned below.



	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	86.79	86.81
1	K-nearest neighbors	86.79	86.81
2	Support Vector Machine	93.40	87.91

Fig1.Model Accuracy -Table

As we have seen accuracy of Logistic Regression and K-Nearest-Neighbours is almost same. Model will give same accuracy but result of both algorithm on same input may give different results. Support Vector Machine have much accurate result as 93.40 on training and 87.91 in testing.

So, the final result is model is working with accuracy 87.91% and predicting whether is person is suffering Heart Disease or not.

VII CONCLUSIONS

Model have lots of conclusions based on Data Analysis (Step 1- *Exploratory Data Analysis)* and Machine learning implementation (Step -3 *Training dataset and Applying Machine Learning Algorithm)* as follows-

(i) Chest Pain: People with having chest Pain equal to 1 2 3 have higher chance of heart disease than Chest Pain 0.

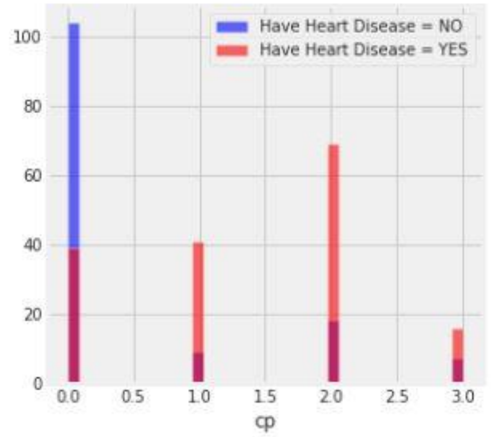


Fig2. -chest Pain vs number of people

(ii) Resting Electrocardiography Results: People with value non-normal heart beat and having mild symptoms are having high chance to have heart disease than normal heart beat people.

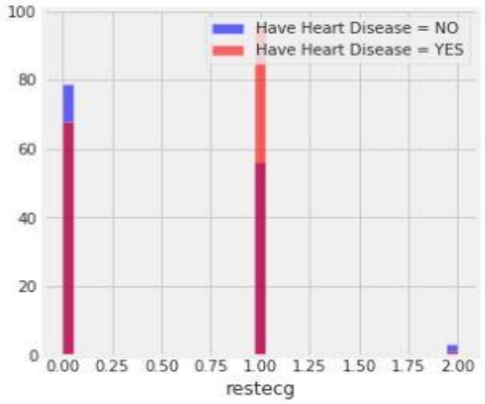


Fig3. -restecg vs number of people

(iii) Exercise induced angina: People not having exercise induced angina have high chance heart disease more than people with exercise induced angina.

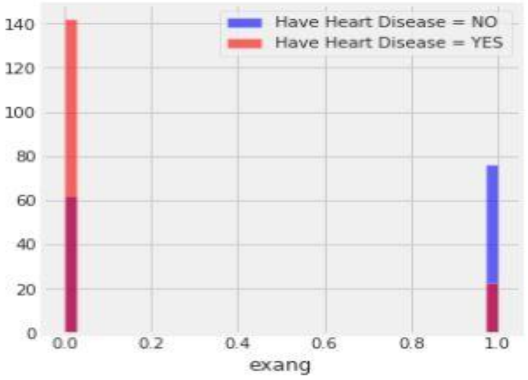


Fig4- exang vs number of people

(iv) The slope of the peak exercise ST segment: People with slope value Down-sloping means unhealthy heart are having high chance of heart disease than people with slope value Up-sloping means better heart rate with exercise.

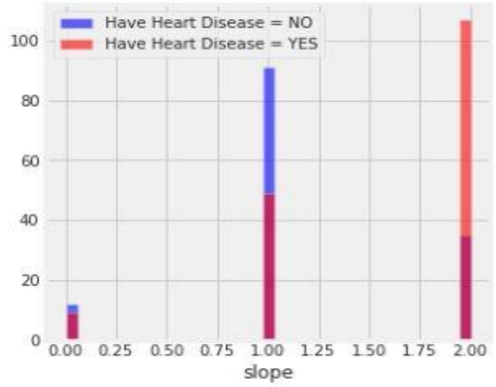


Fig5- slope vs number of people

(v) Thali-um stress result: People with Thallium value equal to fixed defect have high chance of heart disease than rest.

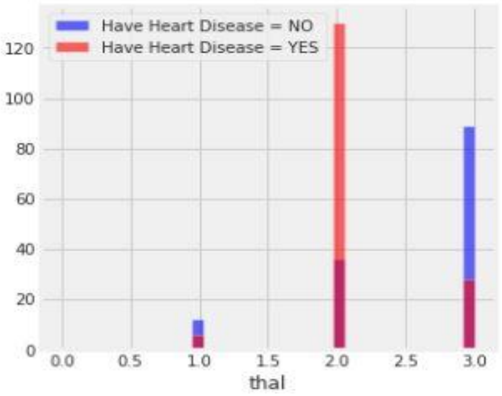


Fig6. -thal vs number of people

(vi) Support Vector Machine Algorithm is much better of this dataset as compared to other algorithms which we tested (KNN and Logistic Regression).

VIII REFERENCES

[1] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, 2018, pp. 1275-1278.

[2] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. *https://archive.ics.uci.edu/ml/datasets/Heart+Disease*

[3] Hazra, Animesh & Mandal, Subrata & Gupta, Amit & Mukherjee, Arkomita & Mukherjee, Asmita. (2017). Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology. 10. 2137-2159

[4] Cardiovascular disease Wikipedia *https://en.wikipedia.org/wiki/Cardiovascular_disease*