Heart Disease Detection Project Report

Anantha Padmanaban
Rajalakshmi Eduverse

(Dated: July 10, 2023)

Under guidance from **Mr. Rahul Ramesh**, Data Scientist, Volvo Trucks India.

Abstract—This project object is to detect whether patients have heart disease or not by given a number of features from patients data. The motivation of this project is to save human resources in medical centers and improve accuracy of diagnosis. In this project we use different methods to detect heart disease such as Logistic Regression,SVM, DecisionTreeClassifier, GradientBoostingClassifier, AdaBoostClassifier ,XGBClassifier, Random Forest and KNeighborsClassifier. And among all these algorithms KNeighborsClassifier gives us the best accuracy of 0.875.

***Keywords*— prediction, machine learning, heart, chest pain, gender**

I. DEFINITION

A Project Overview

The work proposed during this paper focus mainly on various data processing practices that are employed in heart condition prediction. Human heart is that the principal a part of the physical body. Basically, it regulates blood flow throughout our body. Any irregularity in the heart can cause distress in other parts of body. Any kind of disturbance to normal functioning of the heart are often classified as a heart condition. In today's times, heart condition is one among the first reasons for occurrence of most deaths. Heart disease may occur thanks to unhealthy lifestyle, smoking, alcohol and high intake of fat which can cause hypertension [2]. According to the world Health Organization quite 10 million die thanks to heart diseases every single year round the world. A healthy lifestyle and earliest detection are only ways to stop the heart diseases.

The main challenge in today's healthcare is provision of highest quality services and effective accurate diagnosis. Even if heart diseases are found as the prime source of death within the world in recent years, they are also those which will be controlled and managed effectively. The whole accuracy in management of a disease lies on the right time of detection of that disease

The proposed work makes an effort to detect these heart diseases at early stage to avoid disastrous consequences. Records of huge set of medical data created by doctors are available for analysing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the massive amount of knowledge available. Mostly the medical database consists of discrete information. Hence, deciding using discrete data becomes complex and hard task. Machine Learning (ML) which is subfield of knowledge mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning are often used for diagnosis, detection and prediction of varied diseases. The main goal of this paper is to provide a tool for the doctors to detect the heart disease as early stage. This successively will help to supply effective treatment to patients and avoid severe consequences. ML plays a really important role to detect the hidden discrete patterns and thereby analyse the given data. After analysis of knowledge ML techniques help in heart condition prediction and early diagnosis. This paper presents performance analysis of varied

ML techniques like xgboost, Decision Tree, Logistic Regression and Random Forest for predicting heart condition at an early stage.

B Related Work

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine Learning algorithms has motivated this work. This paper contains a brief literature survey. An efficient Cardiovascular disease prediction has been made by using various algorithms some of them include Logistic Regression, KNN, Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register the defined objectives.The model incorporating IHDPS had the ability to calculate the decision boundary using the previous and new model of machine learning and deep learning. It facilitated the important and the most basic factors/knowledge such as family history connected with any heart disease. But the accuracy that was obtained in such IHDPS model was far more less than the new upcoming model such as detecting coronary heart diseases using the artificial neural networks and other algorithms of machine and deep learning. The risk factors of coronary heart disease or atherosclerosis is identified by McPherson et al., using the inbuilt implementation algorithm using uses some techniques of Neural Network and were just accurately able to predict whether the test patient is suffering from the given disease or not.

Diagnosis and prediction of heart disease and Blood Pressure along with other attributes using the aid of neural networks was introduced by R. Subramanian. A deep Neural Network was Built incorporating the given attributes related to the disease which were able to produce a output which was carried out by the output perceptron and almost included 120 hidden layers which is the basic and most relevant technique of ensuring a accurate result of having heart disease if we use the model for Test Dataset. The supervised network has been advised for diagnosis of heart diseases. When the testing of the model was done by a doctor using an unfamiliar data, the model used and trained from the previous learned data and predicted the result thereby calculating the accuracy of the given model.

C. Problem Statement

The goal of this project is to build a robust and accurate predictive model that utilizes supervised learning techniques to classify whether an individual is at risk of developing heart disease based on parameters influencing heart rate. The model will be trained on a dataset containing various physiological and lifestyle parameters such as age, gender, blood pressure, cholesterol levels, smoking habits, and exercise patterns, among others. The aim is to identify patterns and relationships between these parameters and the presence or absence of heart disease.

The developed model should be able to take as input the relevant parameters of an individual and accurately classify whether they are likely to have heart disease. This predictive capability can potentially aid medical professionals in making early interventions and providing appropriate treatment to individuals at high risk of developing heart disease.

The success of this project will be measured based on the model's accuracy, precision, recall, and F1-score in classifying heart disease cases. The model's performance will be evaluated using suitable evaluation metrics, and the best performing model will be selected for deployment in real-world scenarios to assist healthcare practitioners in making informed decisions.

The proposed project aims to contribute to the field of cardiovascular health by leveraging supervised learning techniques to develop an efficient and reliable heart disease prediction system, ultimately leading to improved patient outcomes and reduced mortality rates associated with heart-related conditions.

D Evaluation Metrics

At least one evaluation metric is necessary to quantify the performance of the benchmarks

and solution model. For this project, it will be used the accuracy, which is the number of correct predictions made as a ratio of all predictions made.

Accuracy = Number of correct predictions / Total number of predictions made

This metric only works well if there are a similar number of samples belonging to each class. For this reason, we will divide the range of retweets and likes count in a way that respects this distribution.

II. ANALYSIS

A. Data Source

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and in accordance with other medical conditions [2]. Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to Cardiovascular diseases. We take a data source which is comprised of medical history of304 different patient of different age groups. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 304 patients that helps us detecting if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. This dataset contains 302 rows and 14 columns, where each row corresponds to a single record. All attributes are listed in ‘Table 1’

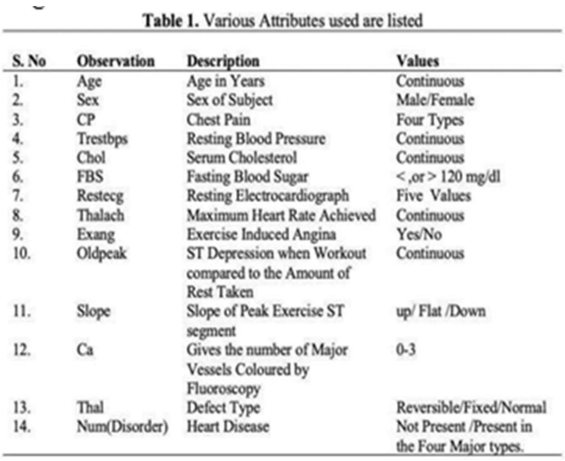
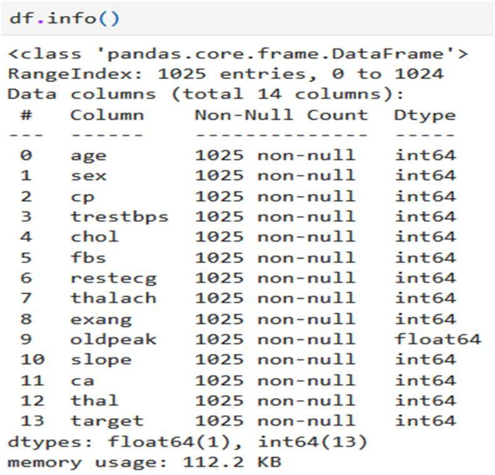


Table 1. Various Attributes used are listed

S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal
14.	Num(Disorder)	Heart Disease	Not Present /Present in the Four Major types.

B. Data Exploration and Visualization

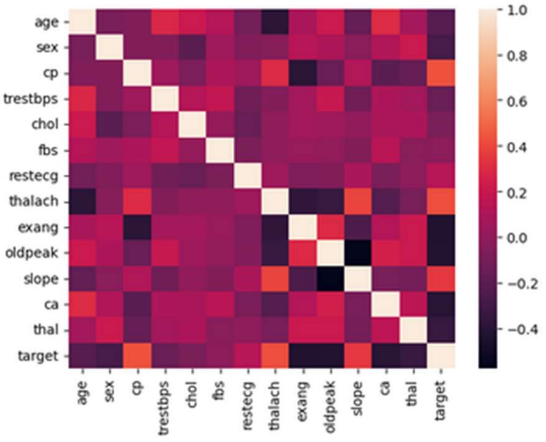
This data consist of fourteen columns all of which are non null values and int64 except oldpeak which is of float64

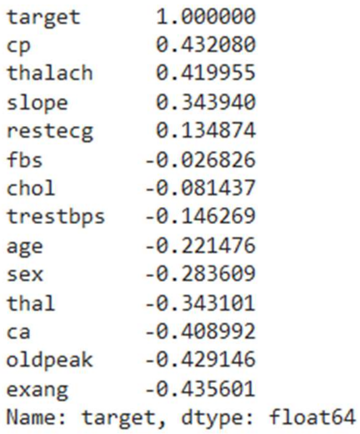


```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1025 non-null   int64  
 1   sex       1025 non-null   int64  
 2   cp        1025 non-null   int64  
 3   trestbps  1025 non-null   int64  
 4   chol      1025 non-null   int64  
 5   fbs       1025 non-null   int64  
 6   restecg   1025 non-null   int64  
 7   thalach   1025 non-null   int64  
 8   exang     1025 non-null   int64  
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64  
 11  ca        1025 non-null   int64  
 12  thal      1025 non-null   int64  
 13  target    1025 non-null   int64  
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

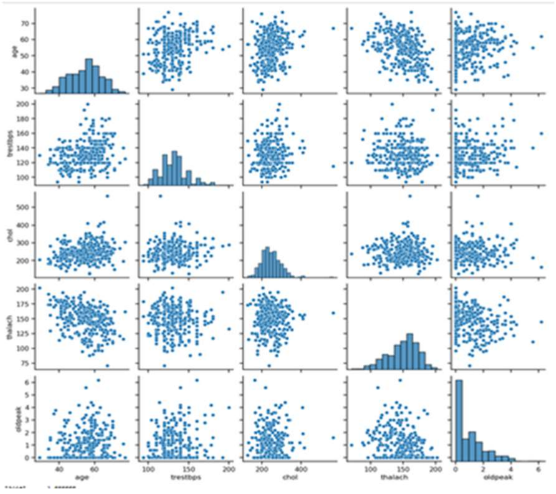
Correlation heatmap show the interrelationship between the features with dark color being not related and towards light color highly related.



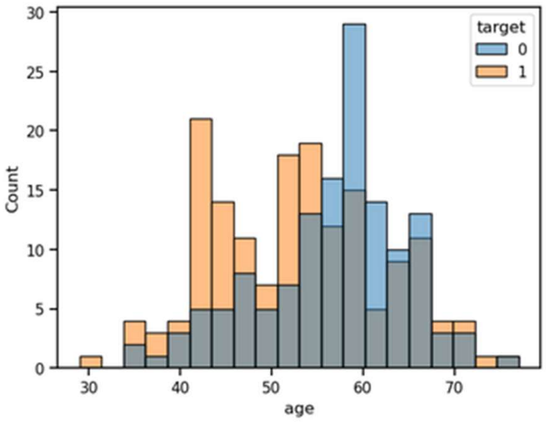


```
target      1.000000
cp          0.432080
thalach     0.419955
slope       0.343940
restecg     0.134874
fbs        -0.026826
chol       -0.081437
trestbps   -0.146269
age        -0.221476
sex        -0.283609
thal       -0.343101
ca         -0.408992
oldpeak    -0.429146
exang      -0.435601
Name: target, dtype: float64
```

in this correlation we find that cp, thalach, slope and restecg are highly correlated to target feature



In scatter matrix we don’t find much information regarding interrelation between features, but we can see outliers in the values which have to be removed.



heart disease patients age is around 35 to 70 years with more number male around 40 to 60 years

III. Data Cleaning

A. Duplicates

When removing duplicate values, found 723 duplicates. data points reduced to 302 value counts.

B. Null Values

When checking for null values. No null data found.

C. Outliers

There were 15 outliers in the dataset which were removed and cleaned, total data count came to 287

IV. Feature Engineering

A. Feature encoding

Convert categorical variable into dummy/indicator variables.Each variable is converted in as many 0/1 variables as there are different values. Columns in the output are each named after a value; if the input is a DataFrame, the name of the original variable is prepended to the value. The columns changed from initial 14 to current 21 columns

B. Feature Scaling

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

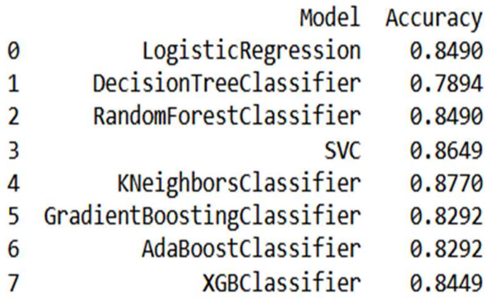
For instance many elements used in the objective function of a learning algorithm ((such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

C. Balancing the Dataset

The dataset is balanced using randomoversampler technique that is oversampling the minority class The total value count came 316

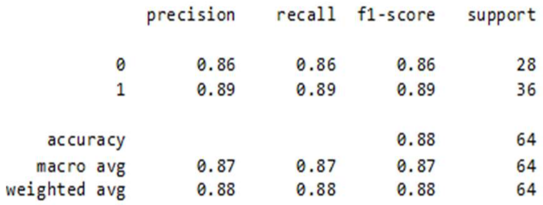
V. Methodology

we use gridsearchcv to find best model and also hyperparameter tuning to find the best hyperparameter for the individual model



```
                        Model  Accuracy
0          LogisticRegression    0.8490
1      DecisionTreeClassifier    0.7894
2      RandomForestClassifier    0.8490
3                         SVC    0.8649
4        KNeighborsClassifier    0.8770
5  GradientBoostingClassifier    0.8292
6          AdaBoostClassifier    0.8292
7               XGBClassifier    0.8449
```

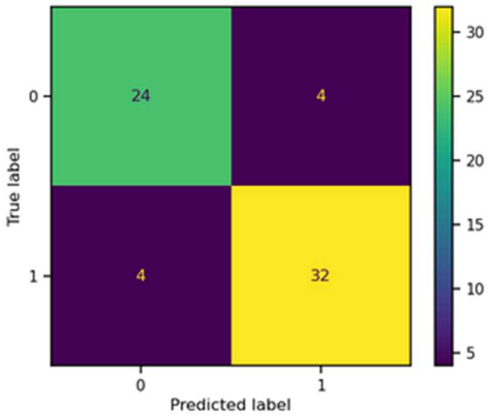
we used accuracy as evaluation metric, out all the eight models KNeighbors Classifier seems to have higher accuracy so we fitted the model. Predicted the model and evaluated the model with the accuracy of 0.875



```
              precision    recall  f1-score   support

           0       0.86      0.86      0.86        28
           1       0.89      0.89      0.89        36

    accuracy                           0.88        64
   macro avg       0.87      0.87      0.87        64
weighted avg       0.88      0.88      0.88        64
```



VI. Conclusion

We use some libraries provided by Python to implement this project. After the experiments, the algorithm of Neighbors-based classification is a type of *instance-based learning* or *non-generalizing learning*: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. Though we get a good result of 87.5% accuracy, that is not enough because it cannot guarantee that no wrong diagnosis happens. To improve accuracy, we hope to require more dataset because 300 instances of dataset are not sufficient to do an excellent job. In the future, to predict disease we want to try different diseases such as lung cancer by using image detection . In this way, the dataset becomes complicated and we can apply convolutional neural network to make accuracy predictions.

VII. References

1. Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shou, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2,

No. 3, June 2012

2. scikit-learn, keras, pandas and matplotlib
3. https://towardsdatascience.com/exploratory-data-analysis-on-heart-disease-uci-data-set-ae129e47b323
4. https://www.kaggle.com/code/microvision/heart-disease-classification/notebook
5. https://www.kaggle.com/code/namanmanchanda/heart-attack-eda-prediction-90-accuracy