Mini Project - I Report

on

Heart Disease Prediction

by

Rahul Shukla
2020400058

Ayush Singh
2020400059

Srikanth Iyengar
2020400062

Guide Name:
Prof. Rupali Sawant

Information Technology Department

Bharatiya Vidya Bhavan's

Sardar Patel Institute of Technology

Munshi Nagar, Andheri(W), Mumbai-400058

University of Mumbai

April 2022

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Rahul Shukla (2020400058)

Ayush Singh (2020400059)

Srikanth Iyengar (2020400062)

Date:

Acknowledgements

We feel great pleasure in presenting the stage one report of our mini project titled ‘Heart Disease Prediction’. We have channelized our best efforts towards a systematic approach to the project, keeping in mind the aim we need to achieve.

We are highly grateful to our project guide Rupali Sawant, Department of Information Technology, Sardar Patel Institute of Technology (SPIT) for constant encouragement, effort and guidance. She has always been involved in discussing our topic at each phase to make sure that our approach was designed and carried out in an appropriate manner and that our conclusions were appropriate, given our results.

Rahul Shukla

Ayush Singh

Srikanth Iyengar

Abstract

In living organisms, the heart is very important. Diagnosis and prediction of heart-related diseases require greater precision, perfection, and correctness because a minor error can result in fatigue or death. There are numerous death cases related to heart disease, and the number is growing exponentially. A prediction system for disease awareness is critical for dealing with the problem.

Machine learning is a branch of Artificial Intelligence (AI) that provides prestigious assistance in predicting any type of event by using training data from natural events. We calculate the accuracy of machine learning algorithms for predicting heart disease, including Random Forest, Decision Tree, Linear Regression, and Support Vector Machine (SVM), in this paper.

Table of Contents

1.INTRODUCTION

1.1. Problem Statement :

Heart disease can be effectively managed through a combination of lifestyle changes, medications, and, in some cases, surgery. Heart disease symptoms can be reduced and heart function improved with the right treatment. The predicted outcomes can be used to avoid and thus reduce the cost of surgical treatment and other costs.

The overall objective of our project will be to accurately predict and attribute the presence of heart disease using a limited number of tests. The attributes considered form the primary basis for tests and provide more or less accurate results. Many more input attributes can be used, but our goal is to predict the risk of heart disease with fewer and faster attributes. Decisions are frequently made on the basis of doctors' intuition and experience rather than the knowledge-rich data hidden in the data set and databases. This practice results in unintended biases, errors, and excessive medical costs, all of which have an impact on the quality of service provided to patients.

1.2. Literature Survey :

[1]Bhoyar S, Wagholikar N, Bakshi K, Chaudhari S (2021) Real-time heart disease prediction system using multilayer perceptron. In: Proceedings of the 2021 2nd international conference for emerging technology (INCET), pp. 1–4. IEEE. DOI: 10.1109/INCET51464.2021.9456389

Research Paper	Learning Paradigm	Method
Real-time Heart Disease Prediction System using Multilayer Perceptron	Supervised Learning	Decision Tree (82.5%) Multi layer perceptron(73%)

[2]Singh, Archana; Kumar, Rakesh (2020). [IEEE 2020 International Conference on Electrical and Electronics Engineering (ICE3) - Gorakhpur, India (2020.2.14-2020.2.15)] 2020 International Conference on Electrical and Electronics Engineering (ICE3) - Heart Disease Prediction Using Machine Learning Algorithms. , (), 452–457. doi:10.1109/ICE348803.2020.9122958

Research Paper	Learning Paradigm	Method
Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	Supervised Learning	Decision Trees (85% accuracy) Logistic Regression (82.9% accuracy) Support Vector Machine (86.1%accuracy) Random Forest Classification (86.1% accuracy)

[3]Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021). Heart Disease Prediction using Hybrid machine Learning Model. 2021 6th International Conference on Inventive Computation Technologies (ICICT). doi:10.1109/icict50816.2021.9358597

Research Paper	Learning Paradigm	Method
Heart Disease Prediction using Hybrid machine Learning Model	Supervised Learning	Decision Tree(79% accuracy) Random Forest (81%accuracy) Hybrid (Decision Tree + Random Forest) (88%accuracy)

1.3. Scope and Objectives:

The scope of this research is to develop an efficient heart disease prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set. Heart disease prediction system aims to exploit data mining and machine learning techniques on medical dataset to assist in the early prediction of heart diseases.

The main objectives of this project are as follows:-

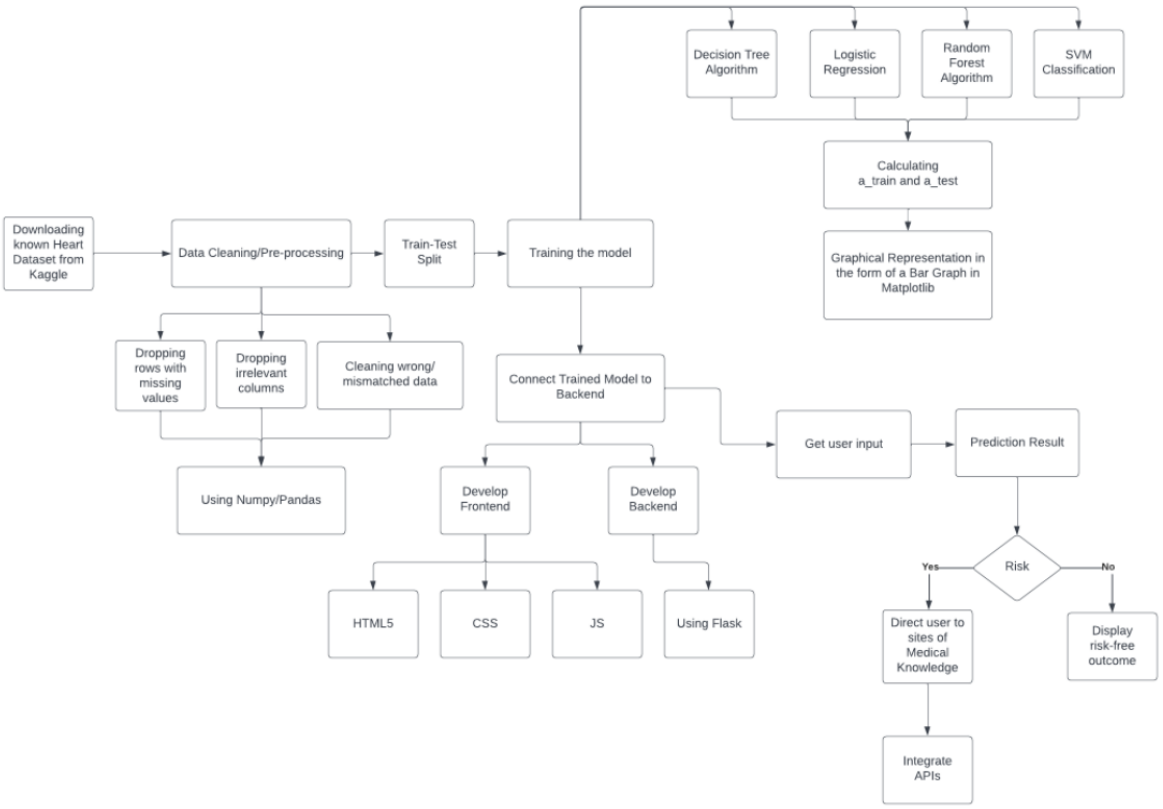
- To analyze feature selection methods and understand their working principle.
- To determine significant risk factors based on medical dataset which may lead to heart disease.
- To develop machine learning model to predict future possibility of heart disease

1.4. Constraints:

Input Parameters	Minimum Constraint	Maximum Constraint
Blood Pressure	70	160
Serum Cholesterol	0	370
Maximum Heart Rate	220	10

2. PROPOSED SYSTEM

2.1. Architecture Diagram :



2.2. Algorithms used :

The algorithms used for the implementation of the project are :

Decision Tree Algorithm:

Input: X_train, Y_train, X_test, Y_test

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contain possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and call the final node as a leaf node.

Output: acc_train, acc_test

Random Forest Algorithm:

Input: X_train, Y_train, X_test, Y_test

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Output: acc_train, acc_test

Random Forest Algorithm:

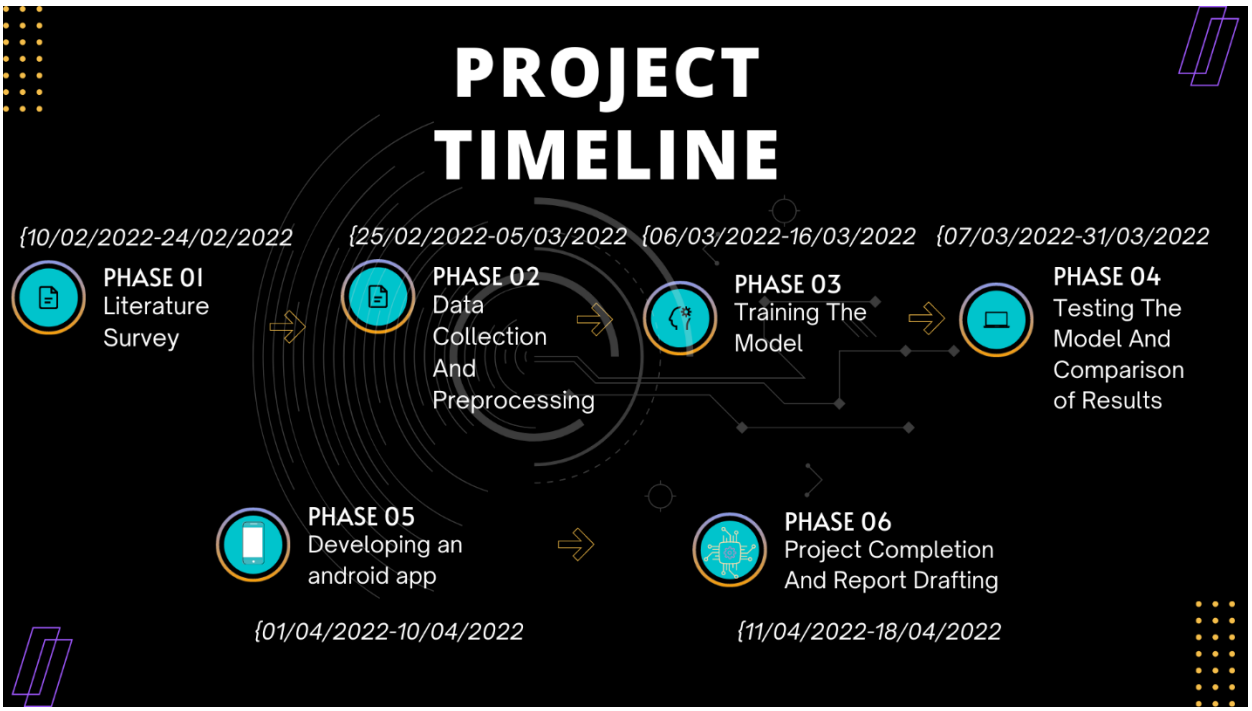
Input: X_train, Y_train, X_test, Y_test

Step-1: Identifying all the separating planes

Step-2: Find the hyper plane to segregate the classes.

Step-3: Predicting the results through the classes separated by the hyper planes.

3. PROJECT PLAN



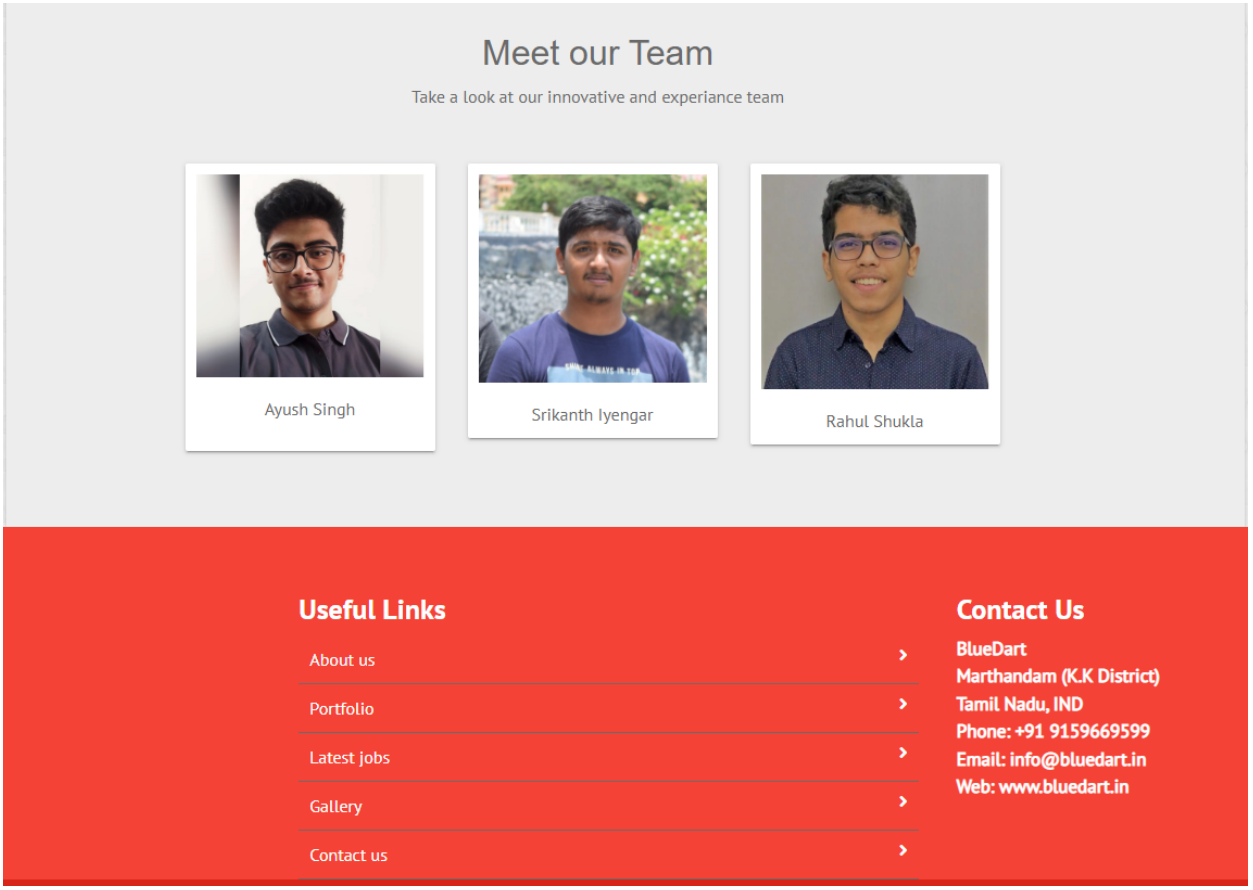
1. Detail the tasks completed module wise for the entire project :

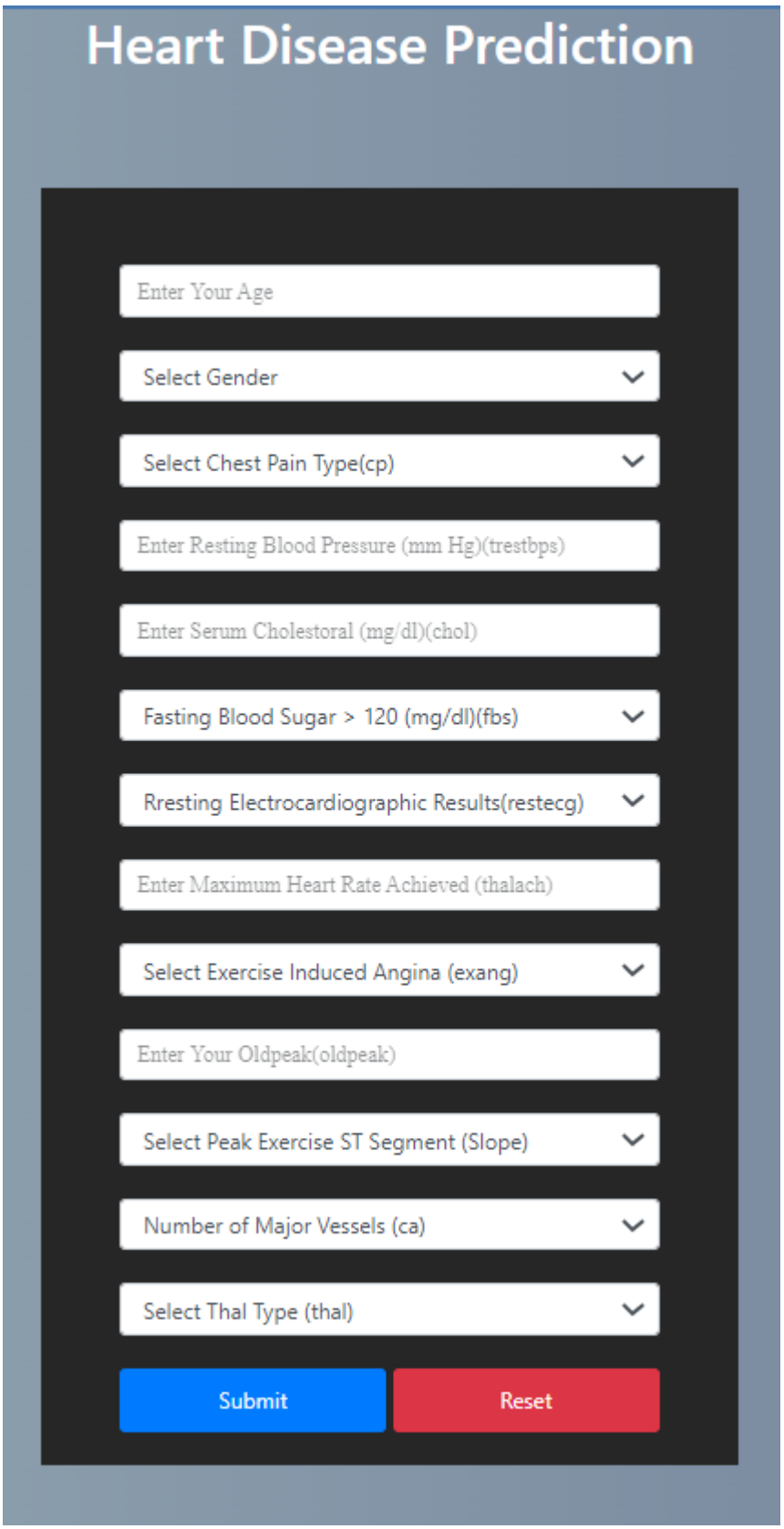
Module 1 : Selection of the topic and understanding the algorithm required for the project. Starting the implementing of the algorithm and getting the basic UI and structure ready for the project.

Module 2 : Completion of the implementation of algorithm and developing the UI of the website. Integrating the algorithm in the web app and making the running model ready for Phase III implementation. Completing 70% of the work in Module 2 and finishing the remaining work in Module III.

Module 3 : Improving the frontend of the website and sorting out the errors present. Getting the model ready for the phase III implementation and completing all the documents.

4. IMPLEMENTATION




Heart Disease Prediction
Enter Your Age
Select Gender
Select Chest Pain Type(cp)
Enter Resting Blood Pressure (mm Hg)(trestbps)
Enter Serum Cholestoral (mg/dl)(chol)
Fasting Blood Sugar > 120 (mg/dl)(fbs)
Rresting Electrocardiographic Results(restecg)
Enter Maximum Heart Rate Achieved (thalach)
Select Exercise Induced Angina (exang)
Enter Your Oldpeak(oldpeak)
Select Peak Exercise ST Segment (Slope)
Number of Major Vessels (ca)
Select Thal Type (thal)
Submit
Reset

1. Tech Stack Used :

We have used python libraries such as :

Scikit-learn:
Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

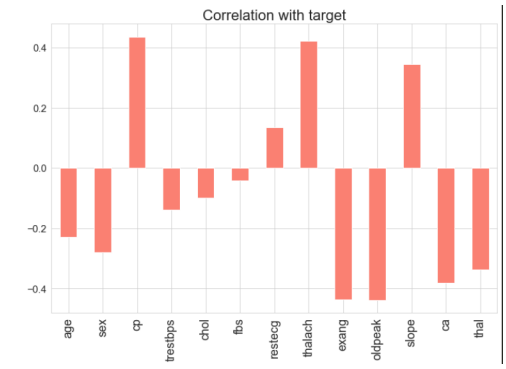
Flask:
Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries.[2] It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Matplotlib:
Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.[3] SciPy makes use of Matplotlib.

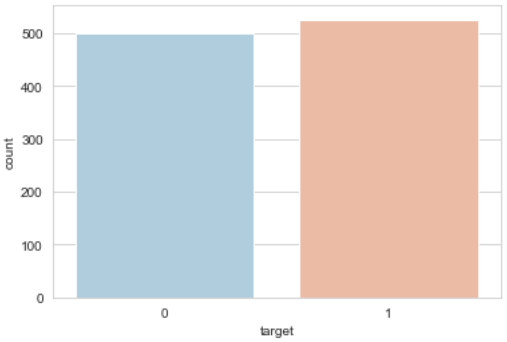
Seaborn:
Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

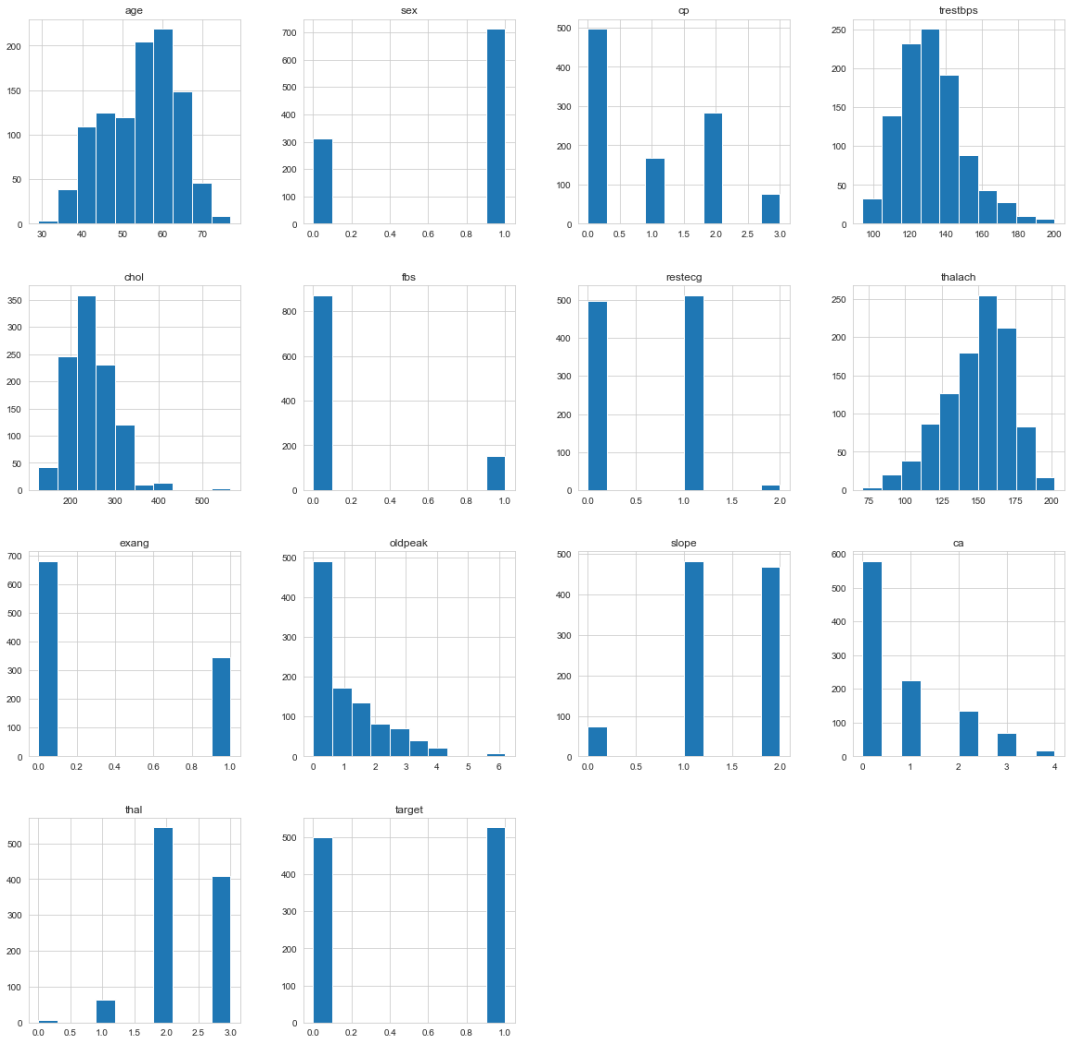
2. Results & Observations: Graphs/Comparative Tables/ Observations from results:



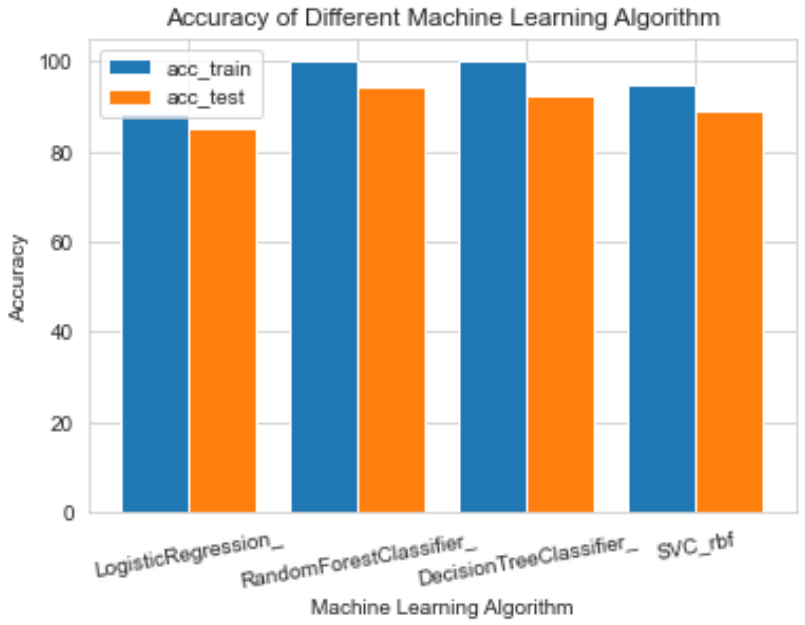
Correlation of every attribute with target



Frequency of targets in the dataset



Correlation matrix of all the attribute



Accuracy of all the algorithms used

3. Module-wise Implementation Screenshots:

a. Data Cleaning and Preprocessing

i) Dropping unnecessary "fbs" column

```
columns = ['sex', 'cp', 'restecg', 'exang', 'slope', 'ca', 'thal', 'fbs']
columns.remove('fbs')
```

ii) Implementing ONE-HOT encoding

```
dataset = pd.get_dummies(dataframe, columns = columns)
```

0.1s

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
standardScaler = StandardScaler()
columns_to_scale = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
dataset[columns_to_scale] = standardScaler.fit_transform(dataset[columns_to_scale])
```

0.1s

b. Training ML Model

i) Decision Tree Algorithm

```
decision_tree = DecisionTreeClassifier(random_state=0)
acc_train, acc_test = get_accuracy(decision_tree, X_train, Y_train, X_test, Y_test)
print("Accuracy Training", acc_train)
print("Accuracy Testing", acc_test)
```

ii) Random Forest Algorithm

```
randomforest_classifier= RandomForestClassifier(n_estimators=100, random_state=0)
acc_train, acc_test = get_accuracy(randomforest_classifier, X_train, Y_train, X_test, Y_test)
print("Accuracy Training", acc_train)
print("Accuracy Testing", acc_test)
```

iii) SVM Classification

```
from sklearn.svm import SVC
svm = SVC(kernel='rbf', probability=True)
acc_train, acc_test = get_accuracy(svm,X_train,Y_train,X_test,Y_test,'rbf')
print("Accuracy Training", acc_train)
print("Accuracy Testing", acc_test)
```

iv) Logistic Regression

```
from sklearn.datasets import load_iris
from sklearn.linear_model import LogisticRegression


logistic_regression = LogisticRegression(random_state=0)
acc_train, acc_test = get_accuracy(logistic_regression, X_train, Y_train, X_test, Y_test)
print("Accuracy Training", acc_train)
print("Accuracy Testing", acc_test)
```

5. CONCLUSIONS AND FURTHER WORK

The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time.

This project discusses different ML Algorithms such as SVM, Logistic Regression, Random Forest and Decision Tree and determines the test-train accuracy for each algorithm.

We finally conclude that for the given dataset SVM Algorithm gives the highest accuracy.

Certain parameters which did not affect "target" were dropped from the original dataset.

We finally link the model to the backend of the website and ask for user prediction

REFERENCES

[1]Bhoyar S, Wagholikar N, Bakshi K, Chaudhari S (2021) Real-time heart disease prediction system using multilayer perceptron. In: Proceedings of the 2021 2nd international conference for emerging technology (INCET), pp. 1–4. IEEE. DOI: 10.1109/INCET51464.2021.9456389
[2]Singh, Archana; Kumar, Rakesh (2020). [IEEE 2020 International Conference on Electrical and Electronics Engineering (ICE3) - Gorakhpur, India (2020.2.14-2020.2.15)] 2020 International Conference on Electrical and Electronics Engineering (ICE3) - Heart Disease Prediction Using Machine Learning Algorithms. , (), 452–457. doi:10.1109/ICE348803.2020.9122958
[3]Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021). Heart Disease Prediction using Hybrid machine Learning Model. 2021 6th International Conference on Inventive Computation Technologies (ICICT). doi:10.1109/icict50816.2021.9358597
[4]https://www.statista.com/statistics/248622/rates-of-leading-causes-of-death-in-the-us/
[5] https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
[6] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
[7] https://archive.ics.uci.edu/ml/datasets/heart+disease
[8]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html