RANDOM FOREST CLASSIFIER FOR PREDICTING CORONARY ILLNESS

A MINI PROJECT REPORT

Submitted by

SHRI VISHNU P (811719104092)

SAKTHIVEL G (811719104077)

SOMASUNDARAM S (811719104095)

in partial fulfillment of the

requirement for the award of

the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

JUNE 2022

K.RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(AUTONOMOUS)

SAMAYAPURAM - 621112

BONAFIDE CERTIFICATE

Certified that this mini project report **"RANDOM FOREST CLASSIFIER FOR PREDICTING CORONARY ILLNESS"** is the bonafide work of **"SHRI VISHNU P (811719104092), SAKTHIVEL G (811719104077), SOMASUNDARAM S (811719104095)"** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE
HEAD OF THE DEPARTMENT
Mr.M.SIVAKUMAR M.E., (Ph.D)
Head of the Department,
Department of Computer Science and Engineering,
K.Ramakrishnan College of Technology (Autonomous),
Samayapuram, Trichy – 621112.

SIGNATURE
PROJECT SUPERVISOR
Mrs.M.MATHUMATHI M.E.,
Assistant professor,
Department of Computer Science and Engineering,
K.Ramakrishnan College of Technology (Autonomous),
Samayapuram, Trichy – 621112.

Submitted for the project viva voce examination held on **……………**

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION

We jointly declare that the mini project report on **“RANDOM FOREST CLASSIFIER FOR PREDICTING CORONARY ILLNESS "** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of Bachelor of Engineering. This mini project report is submitted onthe partial fulfilment of the requirement of the award of Degree of Bachelor of Engineering.

SIGNATURE

SHRI VISHNU P

SAKTHIVEL G

SOMASUNDARAM S

Place: Samayapuram
Date:

ACKNOWLEDGEMENT

It is with great pride that we express our gratitude and in-debt to our institution "**K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY**", for providing us with the opportunity to do this project.

We are glad to credit honorable chairman **Dr. K. RAMAKRISHNAN, B.E.,** for having provided for the facilities during our study in college.

We would like to express our sincere thanks to our beloved Executive Director **Dr.S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

We would like to thank **Dr. N. VASUDEVAN, M.E., Ph.D.,** Dean, who gave opportunity to frame the project the full satisfaction.

We whole heartily thanks to **Mr. M. SIVAKUMAR, M.E., (Ph.D.),** Head of theDepartment, **COMPUTER SCIENCE AND ENGINEERING** for providing his encourage pursuing this project.

We express my deep and sincere gratitude to my project guide **Mrs. M. MATHUMATHI M.E.,** Department of **COMPUTER SCIENCE AND ENGINEERING**, for her incalculable suggestions, creativity, assistance, and patience which motivated me to carry out this project.

We render my sincere thanks to Course Coordinator and other staff members for providing valuable information during the course. We wish to express my special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

Abstract

The process of discovering or mining information from a huge volume of data is known as data mining technology. Today, data mining has lots of applications in every aspect of human life. Among these, health care is a major application of data mining. The medical field has benefited more from data mining. Heart disease is the most dangerous and life-threatening chronic disease globally. Cardiovascular disease prediction is a critical challenge in the area of clinical data analysis. Variation in blood pressure, sugar, pulse rate, etc. can lead to cardiovascular diseases that include narrowed or blocked blood vessels. The objective of the work is to predict the occurrence of heart disease in a patient using a random forest algorithm. The dataset was accessed from the Kaggle site. The dataset contains 303 samples, and 14 attributes are taken for features of the dataset. The datasets are classified and processed using the machine learning algorithm Random Forest. We propose a narrative method that aims at finding significant features by applying machine learning techniques that results in improving the accuracy of the prediction of cardiovascular disease. Using the random forest algorithm, we obtained an accuracy of 86.9% for the prediction of heart disease, with a sensitivity value of 90.3% and a specificity value of 76.7%. From the receiver operating characteristics, we obtained the diagnosis rate for prediction of heart disease using random forest is 91.3%. The random forest algorithm has proven to be the most efficient algorithm for the classification of heart disease, and therefore it is used in the proposed system.

TABLE OF CONTENT

LIST OF FIGURES

Chapter – 1

Introduction

Data mining is also known as proficiency discovering from data. It attempts to withdraw hidden pattern and trends from huge data bases. Data mining also support automatic exploration of data. The main objective of data mining technique is to find the hidden data in the data base. It is also called as exploratory data analysis, data driven and deduction learning. It extracts meaningful information from database. When the database is very large i.e. in terabyte to petabytes manual analysis of data is not possible. So, we need automatic data analysis.

Data mining was introduced in 1990s.Various data mining technologies are as follows.

(i) Statistics: Regression analysis, cluster analysis, standard deviation etc. are the foundation of data mining.

(ii) Artificial Intelligence: It is the applying of human thoughts like processing

(iii) Machine Learning: It is the integration of statistics and AI technology. It is about learning by the software about data.

Chapter – 2

LITERATURE SURVEY

The proposed study gives a prediction method for classification of heart disease. The risk factor which can control and which cannot control was explained in this project. The prediction of heart disease has been done by random forest machine learning algorithm.

[2.1] Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: "Heart disease prediction system". In 2011 Computing in Cardiology (pp. 557-560). IEEE.

Author proposed a user-friendly heart disease prediction system (HDPS). Authors have taken 13 clinical features for classifying heart disease using artificial neural network. Prediction accuracy obtained by the system is approximately 80%. HDPS system include clinical data section, ROC curve section, estimation display section.

In this paper, the authors have developed a heart disease prediction system that can assist medical professionals in predicting heart disease status based on the clinical data of patients. Our approach includes three steps. Firstly, the authors have selected 13 important clinical features, i.e., age, sex, chest pain type, trestbps, cholesterol, fasting blood sugar, resting ecg, max heart rate, exercise induced angina, old peak, slope, number of vessels colored, and thal. Secondly, the authors have developed an artificial neural network algorithm for classifying heart disease based on these clinical features. The accuracy of prediction is near 80%. Finally, we developed a user-friendly heart disease prediction system (HDPS). The authors' approaches are effective in predicting the heart disease of a patient. The HDPS system developed in this study is a novel approach that can be used in the classification of heart disease.

[2.2] Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. "Diabetes disease prediction using data mining."In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-5. IEEE, 2017.

Authors have proposed a Diabetes disease prediction system that gives diabetes malady analysis. Two algorithms were applied namely Bayesian and K-NN for prediction of diabetes. The primary target of this examination is to assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, we propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

[2.3] Rajesh , T Maneesha, Shaik Hafeez, Hari Krishna“Prediction of Heart Disease Using Machine Learning Algorithms“May 2018International Journal of Engineering & Technology 7(2):363-366DOI: 10.14419/ijet. v7i2.32.15714

Author has proposed a model for predicting heart disease by taking samples of 300 patient record using Naïve Bayes and decision trees. Data was taken from UCI repository site Author used id3 algorithm for constructing decision tree. The author also used algorithms for prediction. The Naive Bayes algorithm is analysed on a dataset based on risk factors. The author also used decision trees and a combination of algorithms for the prediction of heart disease based on the above attributes. The results show that when the dataset is small, the naive Bayes algorithm gives the accurate results, and when the dataset is large, the decision trees give the accurate results. For small data set decision tree does not give accurate result but Naïve bayes gives more accurate result if the input data is cleaned.

[2.4] J. Krishnan Santana; S. Geetha "Prediction of Heart Disease Using Machine Learning Algorithms". 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)Publisher: IEEE

Author have proposed a data mining model to predict weather a patient has heart disease or not. Two types of data mining algorithm decision tree and naïve bayes were used for forecasting. These two algorithms were applied on the same data set. Decision tree show an accuracy of 91% and naïve bayes algorithm show an accuracy of 87%. So, in the paper decision tree gives better accuracy for predicting heart disease.

[2.5] Rajdhan Apurb, Agarwal Avi, Sai Milan, Ravi Dundigalla, Ghuli Poonam.” Heart Disease Prediction using Machine Learning” INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY

Authors have proposed a data mining model for prediction of heart disease. Dataset was taken from UCI machine learning repository site. Four data mining algorithms such as Naïve bayes, random forest, Linear regression, Decision tree were applied by the authors to predict the heart disease. Among these algorithms random forest gives good accuracy of 90.16% compared to other algorithms.

[2.6] Singh, A., & Kumar, R. (2020). “Heart Disease Prediction Using Machine Learning Algorithms”. 2020 International Conference on Electrical and Electronics Engineering (ICE3). doi:10.1109/ice348803.2020.9122958

Authors have used knn, decision tree, linear regression, support vector machine algorithms for prediction of heart disease and compared their accuracy. All the datasets for prediction are accesses from UCI repository site. For implementation of the algorithm’s python software is used. All the algorithms are processed in jupyter notebook. From the experimental result authors have obtained best accuracy of 87% by using k-nearest neighbor algorithm followed by support vector machine 83%, decision tree 79%and linear regression of 78% accuracy among all these algorithms for prediction of heart disease.

[2.7] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques". IEEE Access, 1–1. doi:10.1109/access.2019.2923707

Authors have proposed an application for prediction of heart disease for juveniles using multilayer perceptron algorithms. Authors used Cleveland dataset accessed from UCI library the dataset containing 76 parameters such as chest pain, CT scan, ECG etc. The data set was processed in python code using PyCharm tool. From the experimental result authors obtained precision, recall, support value for positive classes were 0.92,0.9,93and for negative classes 0.91,0.89,0.72 respectively.

[2.8] Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). "Heart disease prediction using data mining techniques". In 2017 International Conference on Intelligent Computing and Control(I2C2) (pp. 1-8). IEEE

Authors have proposed a model for prediction of cardiovascular disease using machine learning algorithm hybrid random forest with linear mode. Authors obtained 88.7% accuracy for prediction of CVD using hybrid random forest with linear model. The dataset was collected from UCI repository site. Authors have chosen Cleveland dataset for this proposed study.

CHAPTER 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

The patient provides the input details for this system. The cardiac illness is then assessed using machine learning algorithms based on the user inputs. The generated findings are now compared to those of current models in the same domain and found to be superior. Patterns are discovered using NN, DT, Support Vector Machines SVM, and Naive Bayes on data from heart disease patients obtained at the UCI laboratory. With these algorithms, the performance and accuracy of the outcomes are compared. In comparison to other current methods, the suggested hybrid method produces results of 87 percent for F-measure.

3.1.1 DISADVANTAGES

1. Prediction of cardiovascular disease results is not accurate.
2. Data mining techniques does not help to provide effective decision making.
3. Cannot handle enormous datasets for patient records

3.2 PROPOSED SYSTEM

In this paper, python and pandas operations to classify heart disease using data from the UCI repository after reviewing the findings from existing approaches. It offers a simple visual representation of the dataset, working environment, and predictive analytics development. The machine learning process begins with data pre-processing, then moves on to feature selection based on data cleansing, categorization, and evaluation of modelling performance. To improve the accuracy of the outcome, a random forest technique is applied.

3.2.1 ADVANTAGES

1. Increased accuracy for effective heart disease diagnosis.
2. Handles roughest(enormous) amount of data using random forest algorithm and feature selection.
3. Reduce the time complexity of doctors.
4. Cost effective for patients.

CHAPTER 4

SYSTEM SPECIFICATIONS

4.1.1. SOFTWARE REQUIREMENTS

- Operating system: > Windows XP/7
- Coding Language: Python
- Platform: Kaggle Notebook
- Database: UCI Machine Learning Repository

4.1.2. HARDWARE REQUIREMENTS

- System Processor: Intel core 2
- Hard Disk: 40 GB
- Monitor: 15 VGA Colour
- Mouse: Logitech

CHAPTER 5

SYSTEM DESIGN

5.1 Architecture Diagram

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in away that supports reasoning about the structures and behaviors of the system. System architecture can comprise system components, the externally visible properties of those components, the relationships (e.g. the behavior) between them.

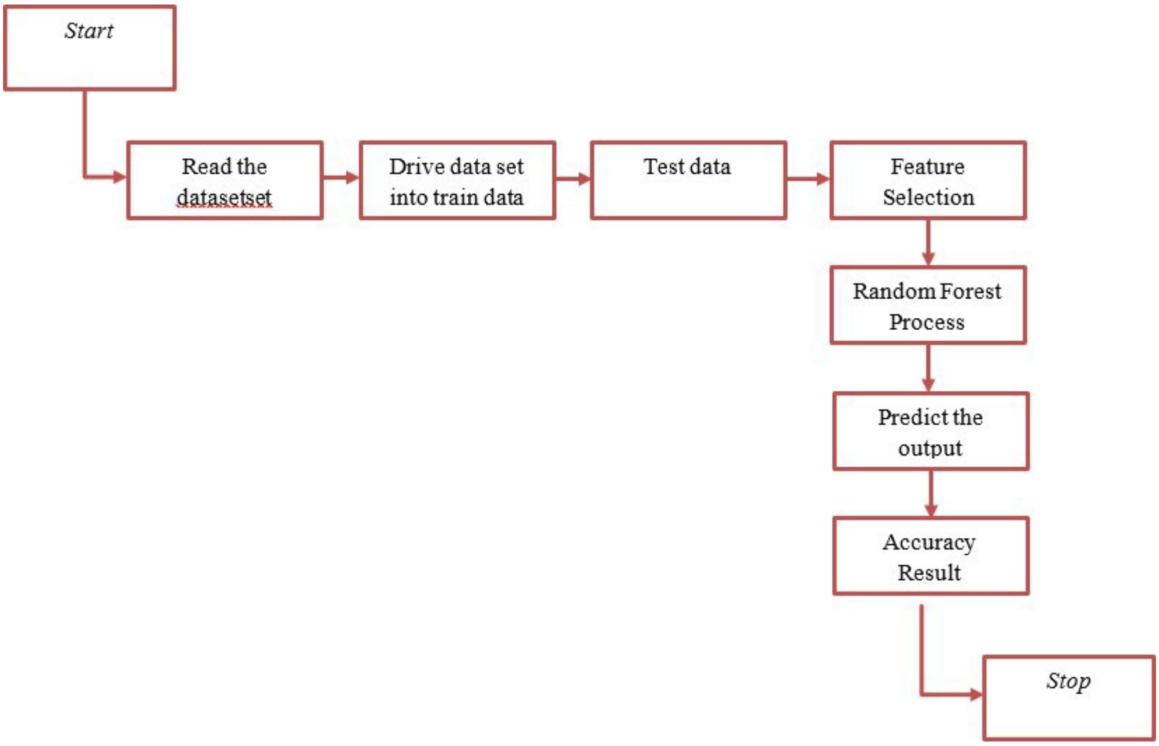


Fig 5.1 System Architecture

Decision trees are the building blocks of a random forest algorithm. A decision tree is a decision support technique that forms a tree-like structure. An overview of decision trees will help us understand how random forest algorithms work.

A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further.

The nodes in the decision tree represent attributes that are used for predicting the outcome. Decision nodes provide a link to the leaves. The following diagram shows the three types of nodes in a decision tree.

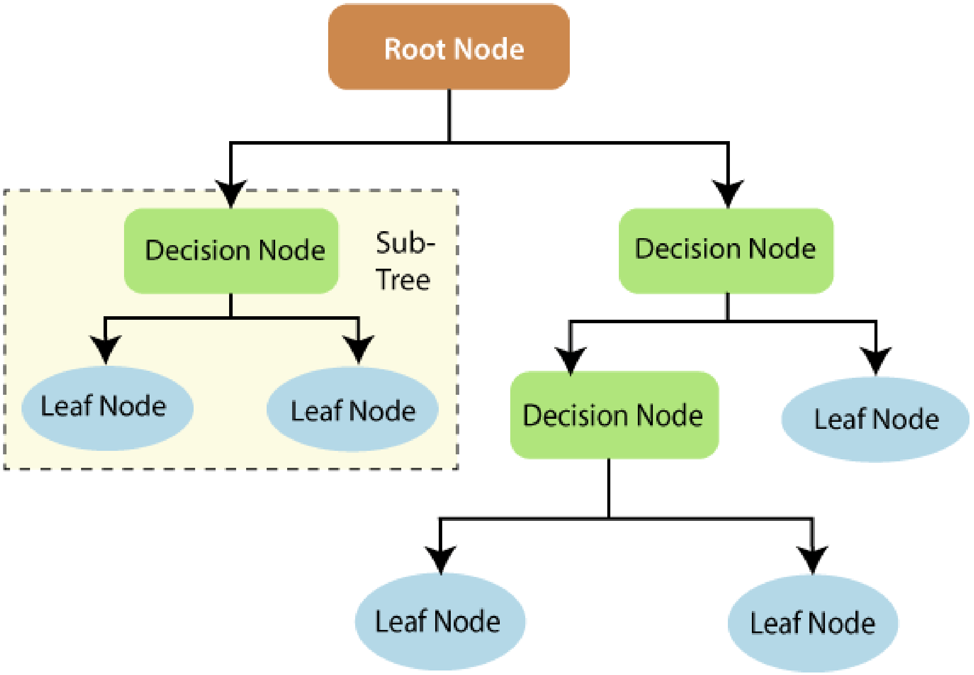


Fig 5.2 Decision tree Architecture

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

The predictions from each tree must have very low correlations.

CHAPTER 6

SOFTWARE DESCRIPTION

6.1 PYTHON

Python is an interpreter, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large- scale projects. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object- oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library. Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

Python is Interactive − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs. Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects. Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

6.2 Kaggle

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Goldbloom and Jeremy Howard. Nicholas Gruen was founding chair succeeded by Max Levchin. Equity was raised in 2011 valuing the company at $25.2 million. On 8 March 2017, Google announced that they were acquiring Kaggle.

Kaggle has run hundreds of machine learning competitions since the company was founded. Competitions have ranged from improving gesture recognition for Microsoft Kinect[9] to making a football AI for Manchester City to improving the search for the Higgs boson at CERN.

Competitions have resulted in many successful projects including furthering the state of the art in HIV research, chess ratings and traffic forecasting. Geoffrey Hinton and George Dahl used deep neural networks to win a competition hosted by Merck. And Vlad Mnih (one of Hinton's students) used deep neural networks to win a competition hosted by Adzuna. This resulted in the technique being taken up by others in the Kaggle community. Tianqi Chen from the University of Washington also used Kaggle to show the power of XGBoost, which has since taken over from Random Forest as one of the main methods used to win Kaggle competitions.

6.3 UCI machine learning repository

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets. As an indication of the impact of the archive, it has been cited over 1000 times, making it one of the top 100 most cited "papers" in all of computer science. The current version of the web site was designed in 2007 by Arthur Asuncion and David Newman, and this project is in collaboration with Rexa.info at the University of Massachusetts Amherst. Funding support from the National Science Foundation is gratefully acknowledged.

CHAPTER 7

Methodology

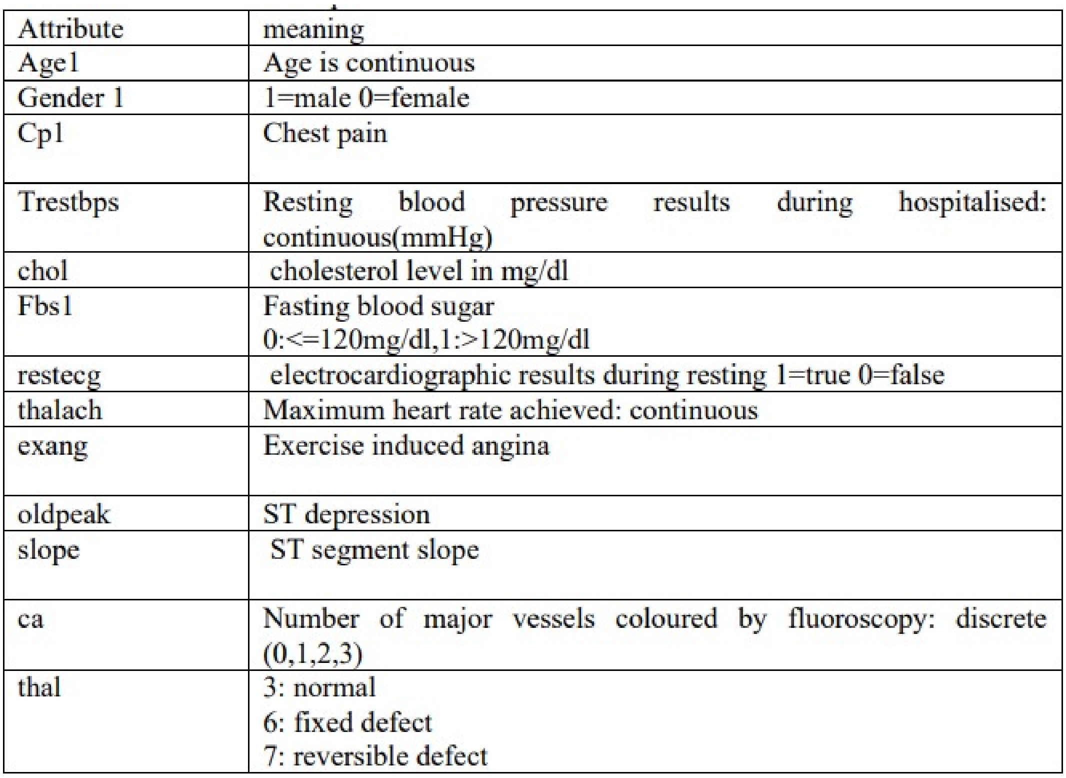
For the proposed study dataset was taken from Kaggle site. Then it was downloaded in excel file using comma separated format. Data has processed by python programming using Jupiter notebook. The data set contains 303 sample instances as shown in table3. The dataset contains 14 clinical features as shown in Fig 7.1. Different types of python libraries such as pandas, Sklearn, NumPy, matplotlib are used for processing the algorithms. Using explorative data analysis technique data was analysed in jupyter notebook.10-fold cross validation technique is used for spitting the data set into training and testing data. Then using random forest algorithm dataset was processed. description of the algorithms: Machine learning is the ability of computer to learn automatically from the experience.

Machine can learn by three ways.

1.Supervised learning

2.Unsupervised Learning

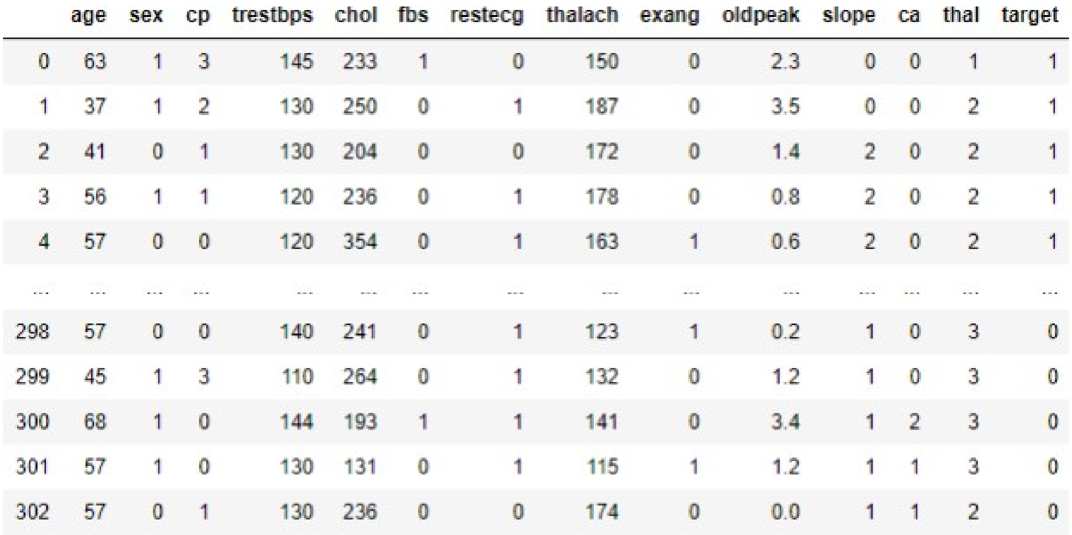
3.Reinforcement
learning



Attribute	meaning
Age1	Age is continuous
Gender 1	1=male 0=female
Cp1	Chest pain
Trestbps	Resting blood pressure results during hospitalised: continuous(mmHg)
chol	cholesterol level in mg/dl
Fbs1	Fasting blood sugar 0:<=120mg/dl,1:>120mg/dl
restecg	electrocardiographic results during resting 1=true 0=false
thalach	Maximum heart rate achieved: continuous
exang	Exercise induced angina
oldpeak	ST depression
slope	ST segment slope
ca	Number of major vessels coloured by fluoroscopy: discrete (0,1,2,3)
thal	3: normal 6: fixed defect 7: reversible defect

Fig 7.1. Abbreviation for data set

In supervised learning label data is given to the machine for prediction. K-NN, Naïve Bayes, Support vector machine, Decision tree, Random Forest algorithms are supervised machine learning algorithms. In unsupervised learning algorithms label data is not given to the machine for prediction. Clustering, c-means are the examples of unsupervised learning In reinforcement learning machine learn by itself without any guidance. It learns from the environment and there is a reward for every action. Q-learning is one of the examples of Reinforcement learning. Random forest is a supervised machine learning algorithm that constructs several decision trees. The final decision is made based on the majority of decision tree. Decision tree suffer from low bias and high variance. Random forest converts high variance to low variance. The present work predicts suffering rate of a patient from heart disease using random forest algorithm. Total 303 data samples (Fig 7.2) of 14 clinical features (Fig 7.1) have taken for prediction of heart disease.80% of the dataset has taken for training and 20% has taken for testing phase.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

Fig 7.2 Overview of Data Set

Chapter - 8

8.1 SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a wayto check the functionality of components, Sub-assemblies, assemblies and\or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test type addresses a specific testing requirement.

TESTING STEPS

- Unit Testing
- Integration Testing
- Functional testing
- System testing
- White Box testing
- Black Box testing
- Output Testing
- User Acceptance Testing

8.2 TYPES OF TESTS

8.2.1. Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

8.2.2 Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components. Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

8.2.3 Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : Classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

8.2.4 System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre- driven process links and integration points.

8.2.4.1 White box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

8.2.4.2 Black box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

8.3 Output Testing

After performing the next step is output of testing of the proposed system since no system could be useful if it does not produce the required output in the specific format. The output generated or displayed by the system under consideration is tested asking the users about the format required by then. Here, the output is considered into two ways: one is on the screen and the others print format. The output format on the screen is found to be correct as the format designed according to the user needs. For the hard copy also; the outcome comes as specified by the user. Hence output testing doesn't result in any connection in the system.

8.4 User Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

8.5 Integration testing

Integration tests are designed to test integrated software components to determine it they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shows by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

CHAPTER 9

IMPLEMENTATION

9.1 Algorithm

9.1.1 Random Forest

Random forest algorithm and decision trees algorithm we have extracted the accurate percentage of detection of fraud from the given dataset by studying its behavior. A confusion matrix is basically a summary of prediction results or a table which is used to describe the performance of the classifier on a set of test data where true values are known. It provides visualization of an algorithm's performance and allows easy identification of classes. Thus, resulting in the computing of most performance measures by giving insights not only the errors being made by the classification model but also tells the type of errors being made. Trained Data and Testing Data is represented in a confusion matrix which portrays:

TP: True Positive which denotes the real data where customers are subjected to fraud and are used for training and were accurately predicted.

TN: True Negative denotes the data which was not predicted and doesn't match with the data which was subjected to the fraud.

FP: False Positive is predicted but there is no possibility of the data to be subjected to the fraud.

FN: False Negative is not predicted but there is an actual possibility of the data who is subjected to fraud.

Accuracy and Recall can be calculated as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Recall = TP / (TP + FN)

9.2 Evaluation model

Model Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and over fitted models. To avoid over fitting, evaluation methods such as hold out and cross-validations are used to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation i.e. dividing the number of correct predictions by the number of total predictions. Mean Squared error is also low.

Random forest is the supervised learning. It is a kind of ensemble learning with majority voting techniques where mixture of expert's decision is taken into account. In ensemble learning, where predictions by applying different models.

Multiple decision trees are constructed with optimal data points as next node. Especially this algorithm work better in improper scaling and missing data values. This algorithm uses the bagging concept where bootstrap samples are randomly selected and decision tree is constructed by training the data. Similarly many trees are constructed parallel and get the decision from different processing training tree. The randomness is limited to a features subset and need to pick from the subset not from the entire set. This is obviously make the work faster and avoidspruning tree. At each training pass the search is limited within the square root oftotal number of sensitive feature count. Similarly build the trees until the error rate is stops in decreasing.

Working of Random Forest Algorithm

Step 1 – In given data set, we can select the random samples

Step 2 – Decision tree will be farming for each sample. It will show prediction result

Step 3 – Voting will be started

Step 4 – Final result is most voted prediction

CHAPTER 10

CONCLUSION

In this project, we proposed a machine learning-based strategy for predicting heart illness, and the findings demonstrated a high accuracy threshold for offering a superior estimation result. We offer a new Random Forest classification to handle the problem of rate prediction without equipment, as well as a way to estimate heart rate and condition. we obtained the Sensitivity value 90.3%. specificity value 76.7, and accuracy value of 91.3 for prediction. ML Techniques are used to extract information from the above input.

APPENDIX 1

Source Code

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns #for plotting
from sklearn.ensemble import RandomForestClassifier #for the model
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import export_graphviz #plot tree
from sklearn.metrics import roc_curve, auc #for model evaluation
from sklearn.metrics import classification_report #for model
evaluation from sklearn.metrics import confusion_matrix #for model
evaluation from sklearn.model_selection import train_test_split #for
data splitting import eli5 #for purmutation importance
from eli5.sklearn import PermutationImportance
import shap #for SHAP values
from pdpbox import pdp, info_plots #for partial plots
np.random.seed(123) #ensure reproducibility
```

```
pd.options.mode.chained_assignment = None #hide any pandas warnings

dt = pd.read_csv("../input/heart.csv")

dt.head(10)

dt.columns = ['age', 'sex', 'chest_pain_type', 'resting_blood_pressure',
'cholesterol', 'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate_achieved',

       'exercise_induced_angina', 'st_depression', 'st_slope', 'num_major_vessels',
'thalassemia', 'target']

dt['sex'][dt['sex'] == 0] = 'female'

dt['sex'][dt['sex'] == 1] = 'male'
dt['chest_pain_type'][dt['chest_pain_type'] == 1] = 'typical angina'

dt['chest_pain_type'][dt['chest_pain_type'] == 2] = 'atypical angina'

dt['chest_pain_type'][dt['chest_pain_type'] == 3] = 'non-anginal pain'

dt['chest_pain_type'][dt['chest_pain_type'] == 4] = 'asymptomatic'

dt['fasting_blood_sugar'][dt['fasting_blood_sugar'] == 0] = 'lower than
120mg/ml'
dt['fasting_blood_sugar'][dt['fasting_blood_sugar'] == 1] = 'greater than
120mg/ml'
dt['rest_ecg'][dt['rest_ecg'] == 0] = 'normal'

dt['rest_ecg'][dt['rest_ecg'] == 1] = 'ST-T wave abnormality'
```

```
dt['rest_ecg'][dt['rest_ecg'] == 2] = 'left ventricular hypertrophy'
dt['exercise_induced_angina'][dt['exercise_induced_angina'] == 0] = 'no'
dt['exercise_induced_angina'][dt['exercise_induced_angina'] == 1] = 'yes'
dt['st_slope'][dt['st_slope'] == 1] = 'upsloping'
dt['st_slope'][dt['st_slope'] == 2] = 'flat'
dt['st_slope'][dt['st_slope'] == 3] = 'downsloping'
dt['thalassemia'][dt['thalassemia'] == 1] = 'normal'
dt['thalassemia'][dt['thalassemia'] == 2] = 'fixed defect'
dt['thalassemia'][dt['thalassemia'] == 3] = 'reversable defect'
dt.dtypes
dt['sex'] = dt['sex'].astype('object')
dt['chest_pain_type'] = dt['chest_pain_type'].astype('object')
dt['fasting_blood_sugar'] = dt['fasting_blood_sugar'].astype('object')
dt['rest_ecg'] = dt['rest_ecg'].astype('object')
dt['exercise_induced_angina'] = dt['exercise_induced_angina'].astype('object')
dt['st_slope'] = dt['st_slope'].astype('object')
dt['thalassemia'] = dt['thalassemia'].astype('object')
dt.dtypes
```

```
dt = pd.get_dummies(dt, drop_first=True)

dt.head()

X_train, X_test, y_train, y_test = train_test_split(dt.drop('target', 1), dt['target'], test_size = .2, random_state=10) #split the data

model = RandomForestClassifier(max_depth=5)

model.fit(X_train, y_train)

estimator = model.estimators_[1]

feature_names = [i for i in X_train.columns]

y_train_str = y_train.astype('str')

y_train_str[y_train_str == '0'] = 'no disease'

y_train_str[y_train_str == '1'] = 'disease'

y_train_str = y_train_str.values

export_graphviz(estimator, out_file='tree.dot',

                feature_names = feature_names,

                class_names =  y_train_str,

                rounded = True, proportion = True,

                label='root',

                precision = 2, filled = True)
```

```
from subprocess import call

call(['dot', '-Tpng', 'tree.dot', '-o', 'tree.png', '-Gdpi=600'])

from IPython.display import Image

Image(filename = 'tree.png')

perm = PermutationImportance(model, random_state=1).fit(X_test, y_test)


eli5.show_weights(perm, feature_names = X_test.columns.tolist())

base_features = dt.columns.values.tolist()

base_features.remove('target')

feat_name = 'num_major_vessels'

pdp_dist = pdp.pdp_isolate(model=model, dataset=X_test, model_features=base_features, feature=feat_name)

pdp.pdp_plot(pdp_dist, feat_name)

plt.show()

feat_name = 'age'

pdp_dist = pdp.pdp_isolate(model=model, dataset=X_test, model_features=base_features, feature=feat_name)
```

```
pdp.pdp_plot(pdp_dist, feat_name)
plt.show()
feat_name = 'st_depression'

pdp_dist = pdp.pdp_isolate(model=model, dataset=X_test,
model_features=base_features, feature=feat_name)
pdp.pdp_plot(pdp_dist, feat_name)
plt.show()
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_test)
shap.summary_plot(shap_values[1], X_test, plot_type="bar")
def heart_disease_risk_factors(model, patient):
    explainer = shap.TreeExplainer(model)
    shap_values = explainer.shap_values(patient)
    shap.initjs()
    return shap.force_plot(explainer.expected_value[1], shap_values[1], patient)
data_for_prediction = X_test.iloc[1,:].astype(float)
heart_disease_risk_factors(model, data_for_prediction)
data_for_prediction = X_test.iloc[3,:].astype(float)
```

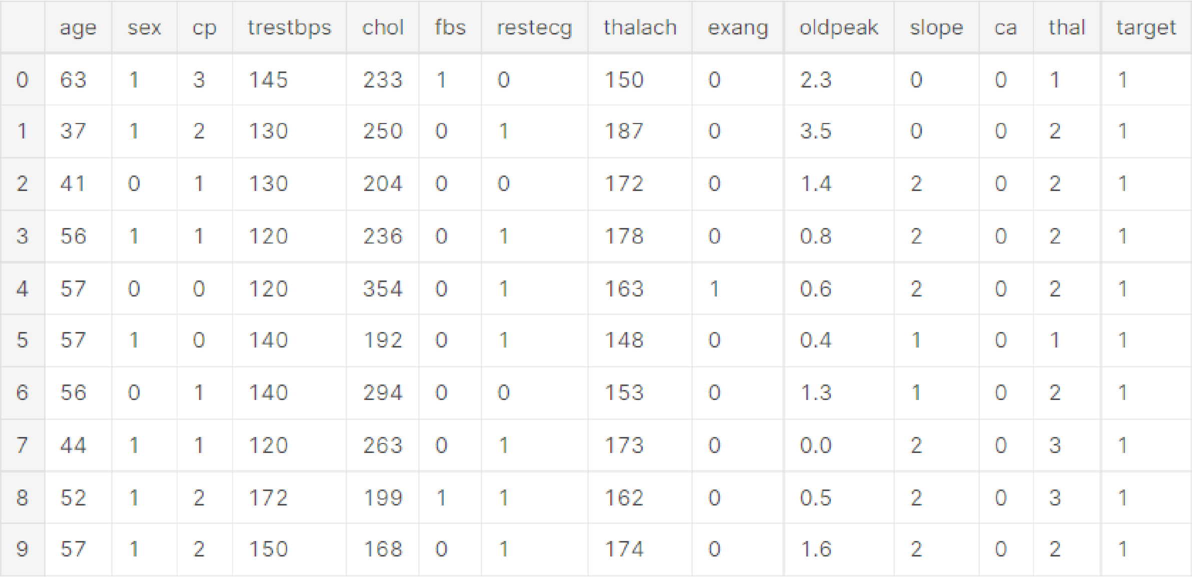
```
heart_disease_risk_factors(model, data_for_prediction)

shap_values = explainer.shap_values(X_train.iloc[:50])

shap.force_plot(explainer.expected_value[1], shap_values[1],
X_test.iloc[:50])
```

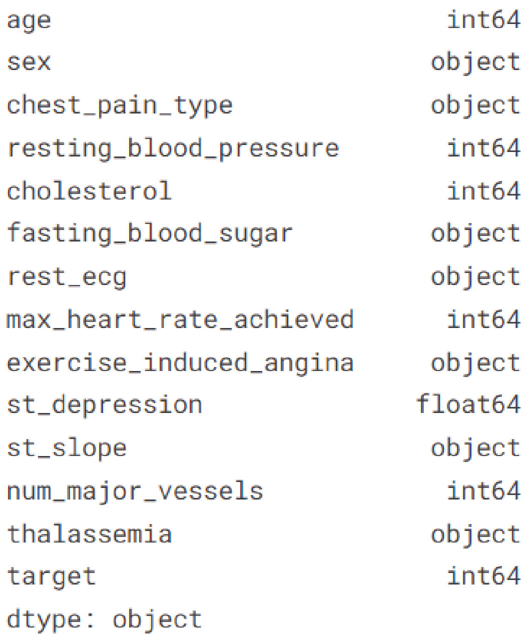
APPENDIX 2

Visualization



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Fig A2.1 Data set



```
age                          int64
sex                         object
chest_pain_type             object
resting_blood_pressure       int64
cholesterol                  int64
fasting_blood_sugar         object
rest_ecg                    object
max_heart_rate_achieved      int64
exercise_induced_angina     object
st_depression              float64
st_slope                    object
num_major_vessels            int64
thalassemia                 object
target                       int64
dtype: object
```

Fig A2.2 The type of data set

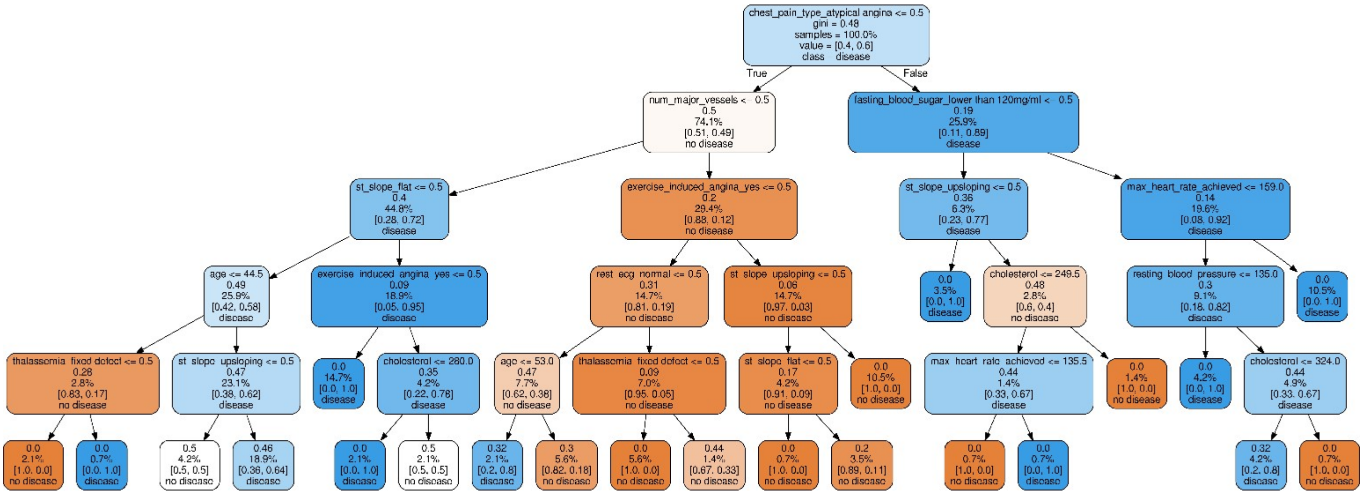

chest_pain_type_atypical angina <= 0.5 gini = 0.48 samples = 100.0% value = [0.4, 0.6] class disease
True
False
num_major_vessels <= 0.5 0.5 74.1% [0.51, 0.49] no disease
fasting_blood_sugar_lower than 120mg/ml <= 0.5 0.19 25.9% [0.11, 0.89] disease
st_slope_flat <= 0.5 0.4 44.8% [0.28, 0.72] disease
exercise_induced_angina_yes <= 0.5 0.2 29.4% [0.88, 0.12] no disease
st_slope_upsloping <= 0.5 0.36 6.3% [0.23, 0.77] disease
max_heart_rate_achieved <= 159.0 0.14 19.6% [0.08, 0.92] disease
age <= 44.5 0.49 25.9% [0.42, 0.58] disease
exercise_induced_angina_yes <= 0.5 0.09 18.9% [0.05, 0.95] disease
rest_ecg_normal <= 0.5 0.31 14.7% [0.81, 0.19] no disease
st_slope_upsloping <= 0.5 0.06 14.7% [0.97, 0.03] no disease
0.0 3.5% [0.0, 1.0] disease
cholesterol <= 249.5 0.48 2.8% [0.6, 0.4] no disease
resting_blood_pressure <= 135.0 0.3 9.1% [0.18, 0.82] disease
0.0 10.5% [0.0, 1.0] disease
thalassemia_fixed defect <= 0.5 0.28 2.8% [0.83, 0.17] no disease
st_slope_upsloping <= 0.5 0.47 23.1% [0.38, 0.62] disease
0.0 14.7% [0.0, 1.0] disease
cholesterol <= 280.0 0.35 4.2% [0.22, 0.78] disease
age <= 53.0 0.47 7.7% [0.62, 0.38] no disease
thalassemia_fixed defect <= 0.5 0.09 7.0% [0.95, 0.05] no disease
st_slope_flat <= 0.5 0.17 4.2% [0.91, 0.09] no disease
0.0 10.5% [1.0, 0.0] no disease
max_heart_rate_achieved <= 135.5 0.44 1.4% [0.33, 0.67] disease
0.0 1.4% [1.0, 0.0] no disease
0.0 4.2% [0.0, 1.0] disease
cholesterol <= 324.0 0.44 4.9% [0.33, 0.67] disease
0.0 2.1% [1.0, 0.0] no disease
0.0 0.7% [0.0, 1.0] disease
0.5 4.2% [0.5, 0.5] no disease
0.46 18.9% [0.36, 0.64] disease
0.0 2.1% [0.0, 1.0] disease
0.5 2.1% [0.5, 0.5] no disease
0.32 2.1% [0.2, 0.8] disease
0.3 5.6% [0.82, 0.18] no disease
0.0 5.6% [1.0, 0.0] no disease
0.44 1.4% [0.67, 0.33] no disease
0.0 0.7% [1.0, 0.0] no disease
0.2 3.5% [0.89, 0.11] no disease
0.0 0.7% [1.0, 0.0] no disease
0.0 0.7% [0.0, 1.0] disease
0.32 4.2% [0.2, 0.8] disease
0.0 0.7% [1.0, 0.0] no disease

Fig A2.3 A decision tree of a model

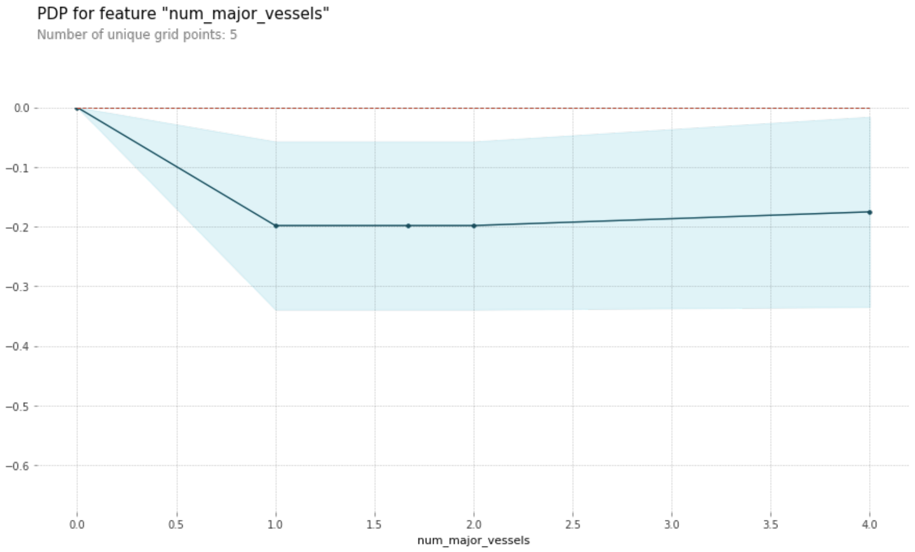


Fig A2.4 PDP for Number of major vessels

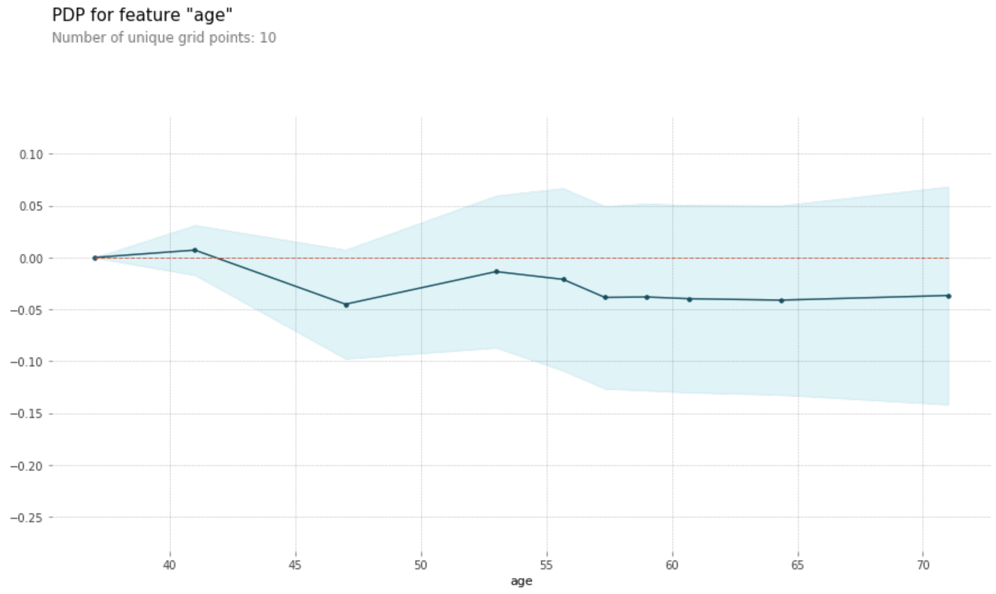


Fig A2.5 PDP for Age

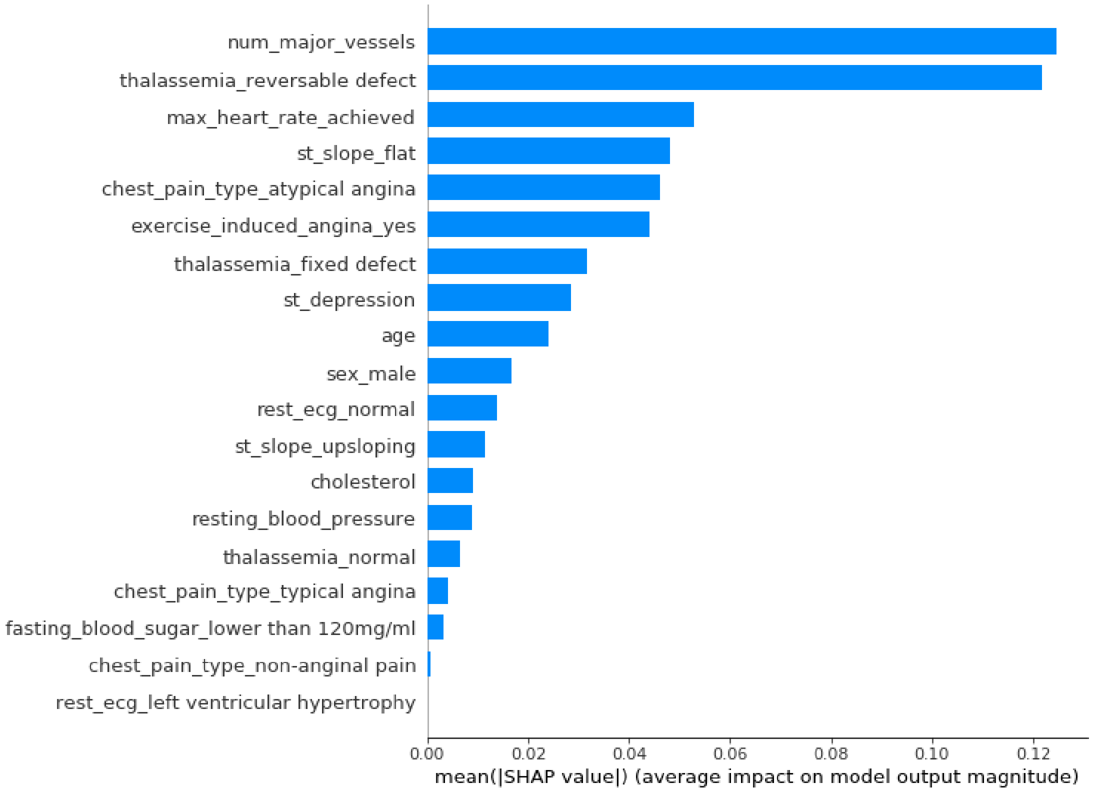


Fig A2.6 Output of a model

REFERENCES

[1] Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: “Heart disease prediction system”. In 2011 Computing in Cardiology (pp. 557-560). IEEE.

[2] Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. "Diabetes disease prediction using data mining."In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-5. IEEE, 2017.

[3] Rajesh , T Maneesha, Shaik Hafeez, Hari Krishna“Prediction of Heart Disease Using Machine Learning Algorithms“May 2018International Journal of Engineering & Technology 7(2):363-366DOI: 10.14419/ijet. v7i2.32.15714 North-Holland/American Elsevier) p 517

[4] J. Krishnan Santana; S. Geetha “Prediction of Heart Disease Using Machine Learning Algorithms”. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)Publisher: IEEE

[5] Rajdhan Apurb, Agarwal Avi, Sai Milan, Ravi Dundigalla, Ghuli Poonam.” Heart Disease Prediction using Machine Learning” INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY

[6] Singh, A., & Kumar, R. (2020). “Heart Disease Prediction Using Machine Learning Algorithms”. 2020 International Conference on Electrical and Electronics Engineering (ICE3). doi:10.1109/ice348803.2020.9122958

[7] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques”. IEEE Access, 1–1. doi:10.1109/access.2019.2923707

[8] Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). “Heart disease prediction using data mining techniques”. In 2017 International Conference on Intelligent Computing and Control(I2C2) (pp. 1-8). IEEE

[9] Al Essa, Ali Radhi, and Christian Bach. "Data Miningand Warehousing." American Society for EngineeringEducation (ASEE Zone 1) Journal (2014). [10] National Health Council, ‘Heart HealthScreenings’,2017. [Online]Available: http://www.heart.org/HEARTORG/Conditions/HeartHealthScreenings_UCM_428687_Article.jsp#. WnsOAeeYPIV