Heart Disease Risk Level Predictor

A thesis/dissertation submitted to the Mahatma Gandhi Central University in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

BY

NISHA GUPTA & NISHANT RAJ

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

MAHATMA GANDHI CENTRAL UNIVERSITY,

MOTIHARI BIHAR-845401, INDIA

MARCH-2023

Heart Disease Risk Level Predictor

A project submitted to the Mahatma Gandhi Central University

in partial fulfilment of the requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

BY

NISHA GUPTA
(MGCU2019CSIT3011)
NISHANT RAJ
(MGCU2019CSIT3012)

Under the Supervision of

Mr. Shubham Kumar

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY,

MAHATMA GANDHI CENTRAL UNIVERSITY,

MOTIHARI, BIHAR-845401, INDIA

MARCH-2023

कंप्यूटर विज्ञान और सूचना प्रौद्योगिकी विभाग
Department of Computer Science and Information Technology
महात्मा गाँधी केन्द्रीय विश्वविद्यालय
MAHATMA GANDHI CENTRAL UNIVERSITY
बिहार/Bihar-845401

DECLARATION

We hereby declare that the work being presented in this report entitled with "**Heart Disease Risk Level Predictor**" Submitted to department of **Computer Science and Information Technology, Mahatma Gandhi Central University Motihari, Bihar-845401, India,** in partial fulfillment of the requirements for the award of the degree of **Batchelor of Technology** in **Computer Science & Engineering**, is a record of bonafide work carried out by me under the supervision of **"Mr. Shubham Kumar"**.

The matter embodied in the dissertation has not been submitted in part or full to any University or Institution for the award of any other degree or diploma.

Nisha Gupta(MGCU2019CSIT3011)

Nishant Raj(MGCU2019CSIT3012)

Department of Computer Science and
Information Technology,Mahatma
Gandhi
Central University, MotihariBihar-
845401, India

कंप्यूटर विज्ञान और सूचना प्रौद्योगिकी विभाग
Department of Computer Science and Information Technology
महात्मा गाँधी केन्द्रीय विश्वविद्यालय
MAHATMA GANDHI CENTRAL UNIVERSITY
बिहार/Bihar-845401

CERTIFICATE

This is to certify that this dissertation entitled **"Heart Disease Risk Level Predictor"** submitted by **Nisha Gupta & Nishant raj**, to the Department of Computer Science & Information Technology, Mahatma Gandhi Central University, Motihari, Bihar-845401, India, for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** , is a project work carried out by him under the supervision of **Mr. Shubham Kumar**.

Shubham Kumar
13/03/23

Head of the Department
Prof. Vikash Pareek
Department of Computer Science and Information Technology,
Mahatma Gandhi Central University,
Motihari, Bihar-845401, India

Supervisor
Mr. Shubham Kumar
Department of Computer Science and Information Technology,
Mahatma Gandhi Central University,
Motihari, Bihar-845401, India

Acknowledgment

I wish to place on record my deep sense of gratitude to our honorific Guide **Mr. Shubham Kumar,** Assistant Professor, Deptartment of CS & IT, MGCUB for his supervision, valuable guidance and moral support leading to the successful completion of the work. Without his continuous encouragement and involvement, this project would not have been a reality.

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to My faculties for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & Friends for their kind co-operation and encouragement, which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons for giving me such attention and time.

My thanks and appreciations go to my colleague in developing the project and people who have willingly helped me out with their abilities.

NISHA GUPTA

NISHANT RAJ

Abstract

Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. Taking various aspects of a dataset collected for heart disease risk level predictor, and the methodology followed for building a predictive model, results with high levels of accuracy are generated, and these observations can be used to predict heart disease risk.

In our project we use different algorithms to detect risk of heart disease such as Linear Regression and Multivariable Polynomial Regression. And it gives us the best accuracy of 75.8%. And we created a website by using html, CSS and bootstrap for taking the input of patient details and used the flask module for deploying the machine learning model and processing that data.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	DECLARATION	2
	CERTIFICATE	3
	ACKNOWLEDGMENT	4
	ABSTRACT	5
	LIST OF FIGURES	7
	LIST OF TABLES	8
	LIST OF ABBREVIATIONS	9
Chapter-1.	Introduction	10
Chapter-2.	Software and Hardware Requirements	11
2.1	Introduction	11
2.2	Hardware requirements	11
2.3	Software requirements	11
Chapter-3.	Proposed Methodology	12
3.1	Dataset and Features	12
3.2	Methods	13-15
Chapter-4.	Implementations and Results	16-18
Chapter-5.	Libraries and Module for Algorithm Development Using Python	19
Chapter-6.	Conclusion	20
References		21

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
1.	Linear regression	13
3.	Multivariable Polynomial Regression	14
4.	R Squared and Coefficient of Determination Theory	15
5.	Compute Coefficient of Determination	15
6.	Home Page	16
7.	Patient detail page	16
8.	Patient Result page	17
9.	Linear regression model	18

LIST OF TABLES

TABLE NO.	TITLE	PAGE NO.
1.	Dataset and Features	12

LIST OF ABBREVIATIONS

TC: Total cholesterol

HDL: High-density lipoprotein

SBP: Systolic blood pressure

Diab: Diabetic type 1

INTRODUCTION

In this fast-moving world the risk of heart disease is increasing proportionally as people want to live a very luxurious life, so they work like a machine in order to earn a lot of money and live a comfortable life. The rate of heart attacks for people under 40 is increasing and various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. Heart disease is very fatal, and it should not be taken lightly. So, a risk predictor can be used to predict the magnitude of future cardiovascular disease

Our aim is to develop a model to predict whether patients have a chance of heart disease by giving some features of users. This is important in medical fields. If such a prediction is accurate enough, then a patient with heart disease can be diagnosed early, which will reduce the death rate caused by heart failure or can get the treatment on time.

By applying our machine learning tool into medical prediction, we will save human resources because we do not need complicated diagnosis processes in hospital (though it is a very long way to go.) The input to our algorithm is 8 features with number values and binary values. We use algorithms such as Linear Regression and multivariable polynomial regression to output the risk percentage which indicates the chances of having heart disease.

Software & Hardware Requirements

Software Requirements

- **Operating System (Any OS with clients to access the internet)**
  An operating system (OS) is a type of system software that controls how computer hardware and software resources are used and offers basic services to other software applications.

- **Network (Wi-Fi Internet or cellular Network)**
  A network is made up of several linked devices, such as computers, servers, mainframes, network devices, peripherals, or other gadgets, that enable data exchange.

- **Visual Studio Code (Create and design data flow and Context Diagram)**
  Microsoft created the source-code tool Visual Studio Code for Windows, Linux, and macOS. Debugging support, grammar colouring, clever code completion, snippets, code refactoring, and integrated Git are among the features.

- **GitHub (Versioning Control)**
  The company GitHub, Inc. offers web hosting services for Git version management and software development. It provides its own features in addition to Git's global version control and source code administration capabilities.

- **Google Chrome (Used for hosting website and system testing)**
  Google Chrome is an online browser programme that can be used to view a local or global website. The web browser obtains the required information from a web server and shows the page on the user's device in response to a user request for a web page from a specific website..

- **Jupyter Notebook**
  An open source online tool called the Jupyter Notebook can be used to make and share papers with active code, equations, visualisations, and text. Users can create and organise workflows in data science, scientific computing, computational journalism, and machine learning using the interface's flexibility. A modular structure encourages modifications to increase and improve utility.

Hardware Requirements

- **Processor:** Intel or high
- **Ram:** 1024MB
- **Space on disk:** minimum 100mb
- **For running the application:**
  - **Device:** Any device that can access the internet
  - **Minimum space to execute:** 20MB

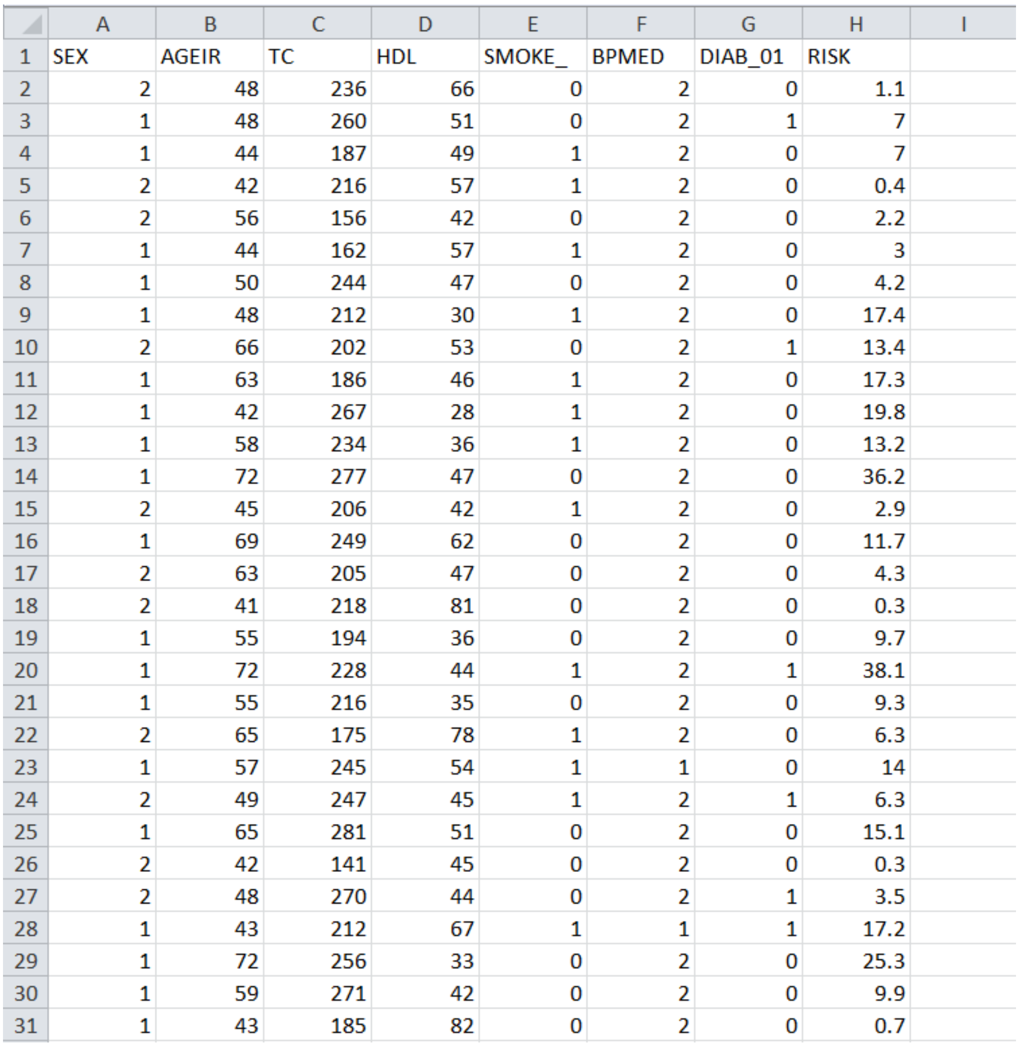
Proposed Methodology

DATASET AND FEATURES

The data set for this model was taken from Kaggle (data repository) and it has 6644 instances.

- gender: gender (1=male; 2=female)
- age: age (in years)
- tc: Total cholesterol (in mg/dL)
- hdl: High-density Lipoprotein (in mg/dL)
- sbp: Systolic Blood Pressure (in mm)
- smoke: smoke (1=yes; 0=no)
- blood pressure medication: Blood Pressure Medication (1=no; 2=yes)
- diab: diabetic type 1(1=yes; 0=no)

From Kaggle, we downloaded a dataset. 90% of the dataset (5980 examples) are used for training, and 10% (664 instances) are used for testing. In order to determine the heart disease risk proportion using the aforementioned characteristics, we trained the multivariable polynomial regression model using the dataset.



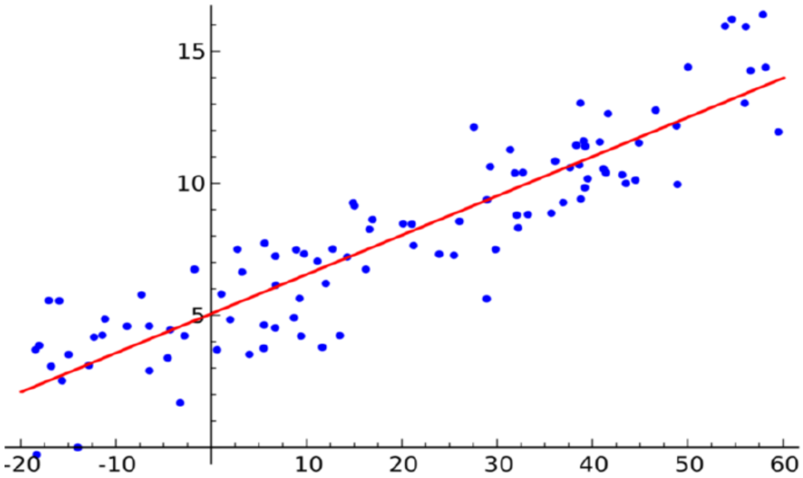
	A	B	C	D	E	F	G	H	I
1	SEX	AGEIR	TC	HDL	SMOKE_	BPMED	DIAB_01	RISK	
2	2	48	236	66	0	2	0	1.1	
3	1	48	260	51	0	2	1	7	
4	1	44	187	49	1	2	0	7	
5	2	42	216	57	1	2	0	0.4	
6	2	56	156	42	0	2	0	2.2	
7	1	44	162	57	1	2	0	3	
8	1	50	244	47	0	2	0	4.2	
9	1	48	212	30	1	2	0	17.4	
10	2	66	202	53	0	2	1	13.4	
11	1	63	186	46	1	2	0	17.3	
12	1	42	267	28	1	2	0	19.8	
13	1	58	234	36	1	2	0	13.2	
14	1	72	277	47	0	2	0	36.2	
15	2	45	206	42	1	2	0	2.9	
16	1	69	249	62	0	2	0	11.7	
17	2	63	205	47	0	2	0	4.3	
18	2	41	218	81	0	2	0	0.3	
19	1	55	194	36	0	2	0	9.7	
20	1	72	228	44	1	2	1	38.1	
21	1	55	216	35	0	2	0	9.3	
22	2	65	175	78	1	2	0	6.3	
23	1	57	245	54	1	1	0	14	
24	2	49	247	45	1	2	1	6.3	
25	1	65	281	51	0	2	0	15.1	
26	2	42	141	45	0	2	0	0.3	
27	2	48	270	44	0	2	1	3.5	
28	1	43	212	67	1	1	1	17.2	
29	1	72	256	33	0	2	0	25.3	
30	1	59	271	42	0	2	0	9.9	
31	1	43	185	82	0	2	0	0.7	

METHODS

During this project, we have tried 2 algorithms for experiment, and they are Linear Regression and Multivariable Polynomial Regression.

Linear Regression

- Regression is a method of modeling a target value based on independent predictors.
- This method is mostly used for forecasting and finding out the cause-and-effect relationship between variables.
- Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



- Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variables.

$$y = m*x + c$$

$$m = \frac{\overline{x} \cdot \overline{y} - \overline{xy}}{(\overline{x})^2 - \overline{x^2}}$$

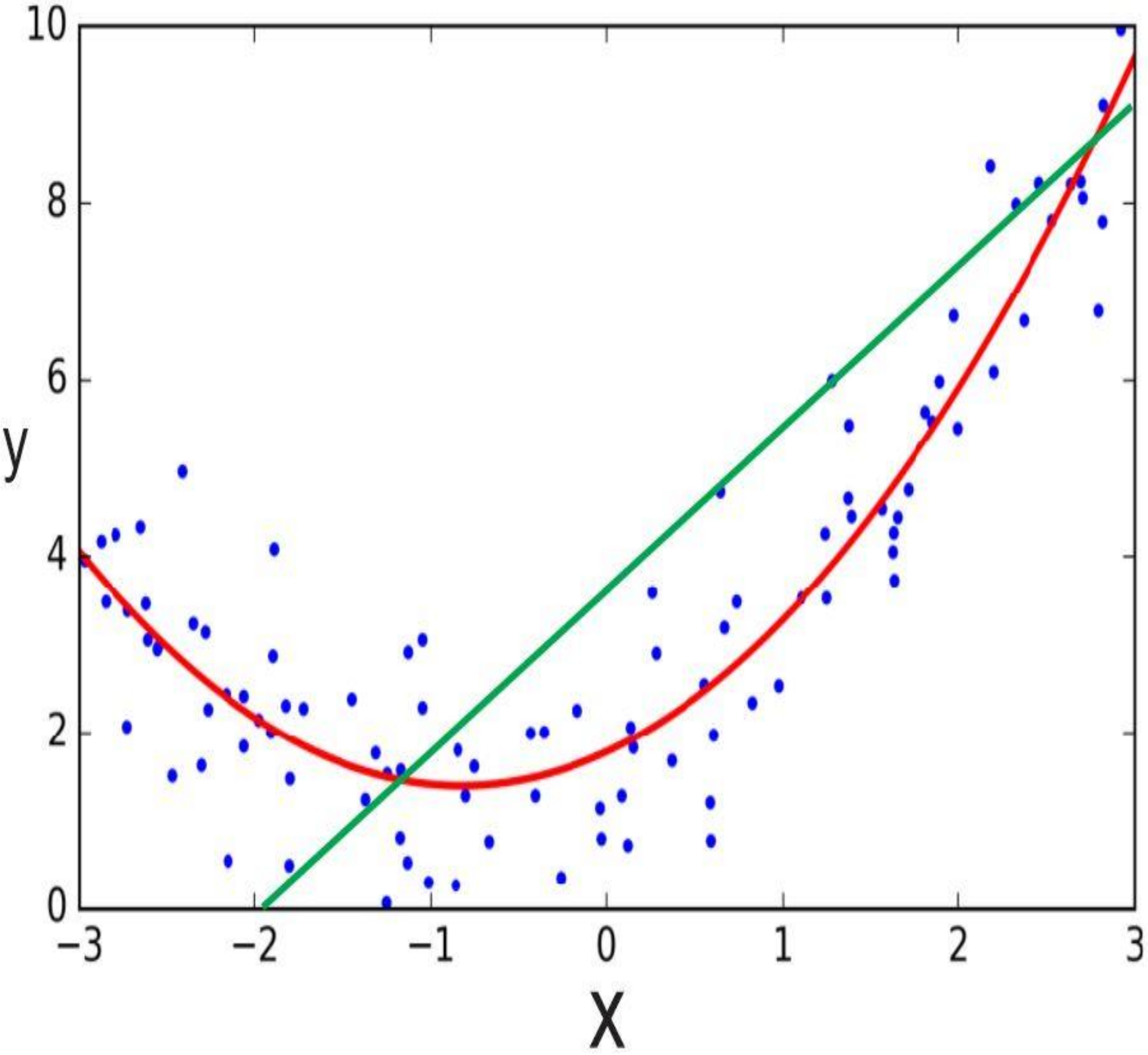
$$b = \overline{y} - m\overline{x}$$

Multivariable Polynomial Regression

- Multivariate Multiple Regression is the method of modeling multiple responses, or dependent variables, with a single set of predictor variables.

- As with many other concepts in machine learning, polynomial regression is a statistical concept. When there is a non-linear relationship between the value of xx and the associated conditional mean of yy, statisticians use it to perform analysis..

- Suppose you want to forecast the number of likes your new social media post will receive at various times after it is posted. The quantity of likes and the passage of time are not linearly correlated. After being published, your new article will probably receive a lot of likes for the first 24 hours before losing some of its fame.



R Squared and Coefficient of Determination Theory

- The coefficient of determination is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event.
- In the second image, there is a best fit line, though even the best fitting line is still going to be useless.
- And we'd like to know that before we spend precious computational power on it.
- The standard way to check for errors is by using squared errors. You will hear this method either called R squared or the coefficient of determination.

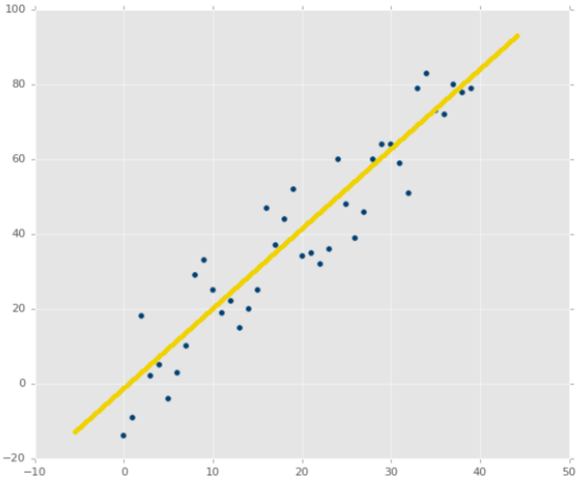


Figure 1

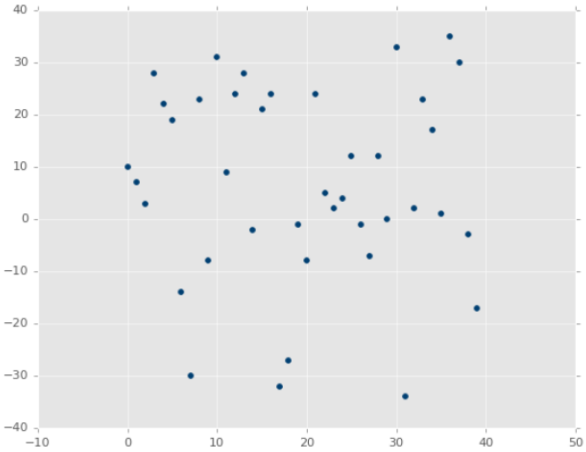
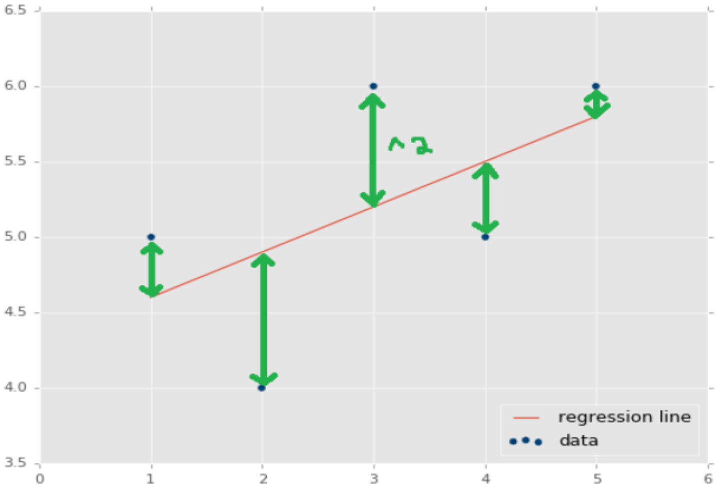


Figure 2

How to Compute Coefficient of Determination

The distance between the regression line's y values, and the data's y values is the error, then we square that. The line's squared error is either a mean or a sum of this, we'll simply sum it.

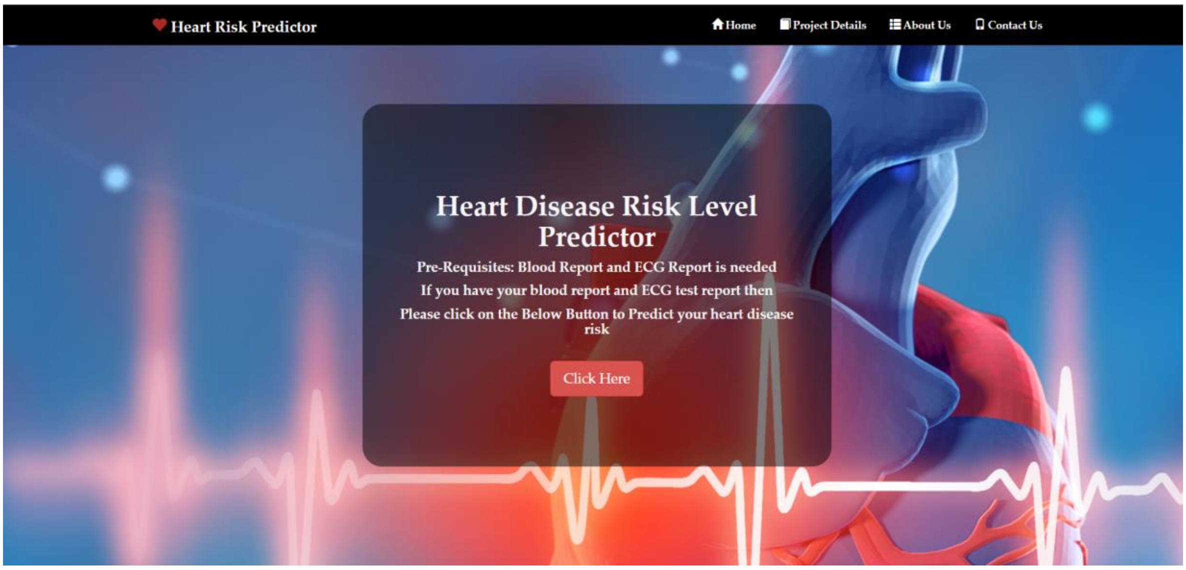
$$r^2 = 1 - \frac{SE\widehat{y}}{SE\overline{y}}$$



Implementations and Results

We created a website by using HTML, CSS and Bootstrap for taking the input from the user and displaying the calculated result.

- **Home page:**
  This is the first page of the website which contains the navigation bar and footer along with the (click here) button which will navigate the user to the patient detail page which contains the form.



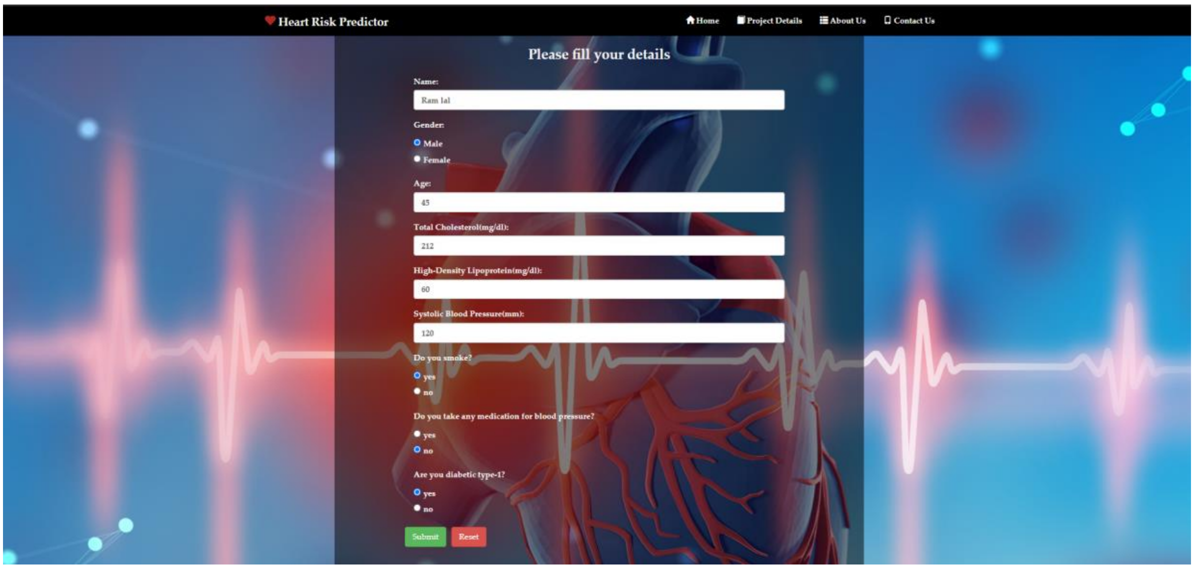
Heart Disease Risk Level Predictor

Pre-Requisites: Blood Report and ECG Report is needed

If you have your blood report and ECG test report then

Please click on the Below Button to Predict your heart disease risk

- **Patient detail page:**
  This page contains the form which is required to be filled by the user to calculate the heart risk. It contains all the features (gender, age, tc, hdl, sbp, smoke, blood pressure medication, diab) which are required by the machine learning model to predict the result.



Please fill your details

Name: Ram lal

Gender:
● Male
○ Female

Age: 45

Total Cholesterol(mg/dl): 212

High-Density Lipoprotein(mg/dl): 60

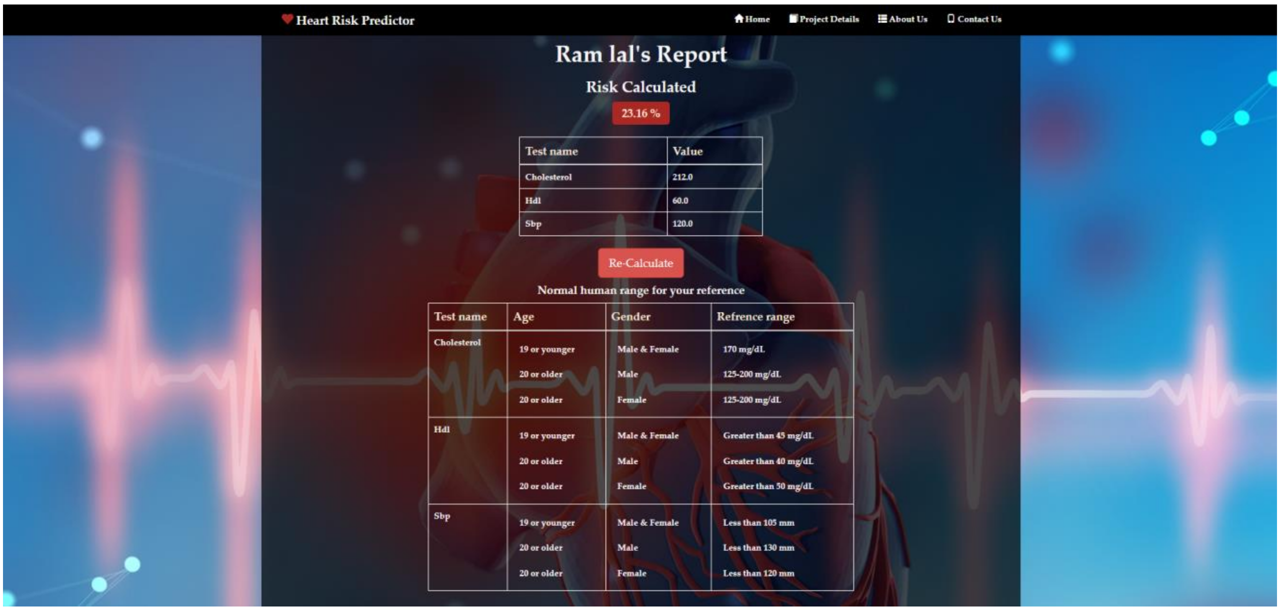
Systolic Blood Pressure(mm): 120

Do you smoke?
● yes
○ no

Do you take any medication for blood pressure?
○ yes
● no

Are you diabetic type-1?
● yes
○ no

- **Patient Result page:**
  This page will display the calculated result along with some reference data which can help the user to compare his/her data with the given normal range.



Ram lal's Report

Risk Calculated

23.16 %

Test name	Value
Cholesterol	212.0
Hdl	60.0
Sbp	120.0

Normal human range for your reference

Test name	Age	Gender	Refrence range
Cholesterol	19 or younger	Male & Female	170 mg/dL
	20 or older	Male	125-200 mg/dL
	20 or older	Female	125-200 mg/dL
Hdl	19 or younger	Male & Female	Greater than 45 mg/dL
	20 or older	Male	Greater than 40 mg/dL
	20 or older	Female	Greater than 50 mg/dL
Sbp	19 or younger	Male & Female	Less than 105 mm
	20 or older	Male	Less than 130 mm
	20 or older	Female	Less than 120 mm

We imported the module flask (web framework) for deploying the machine learning model and processing that data.

Libraries imported for implementing the project:

Flask:
With the help of helpful tools and features, Flask is a compact and lightweight Python web platform that facilitates the development of web apps. Since you can rapidly create a web application using only one Python file, it offers developers freedom and is a more approachable platform for new developers.

Matplotlib:
For Python and its numerical expansion NumPy, Matplotlib is a cross-platform data display and graphical plotting tool. As a result, it presents a strong open-source substitute for MATLAB. The APIs (Application Programming Interfaces) for Matplotlib allow programmers to incorporate charts into GUI apps..

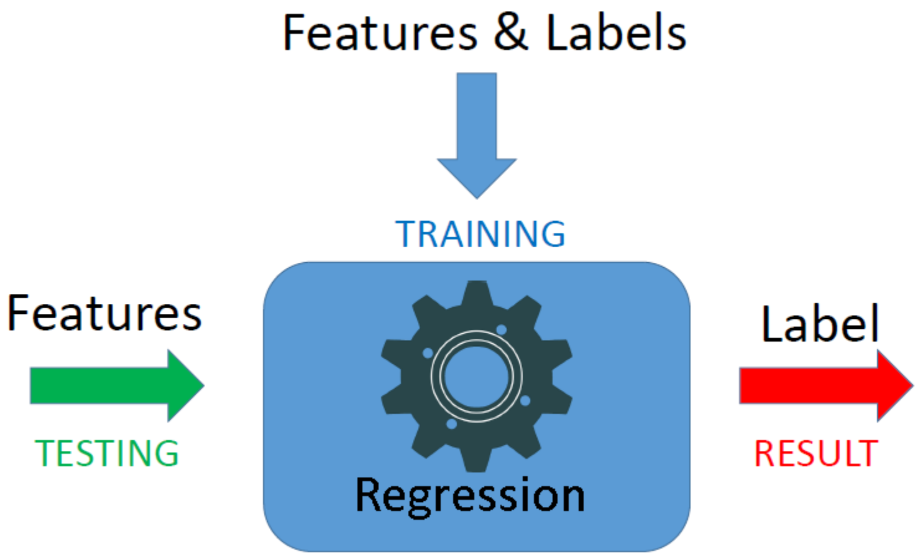
NumPy:
Numerous mathematical procedures can be carried out on collections using NumPy. It provides a vast collection of high-level mathematical functions that work on these arrays and matrices, as well as strong data structures that ensure efficient computations with arrays and matrices.

Pandas:
The most frequently used open-source Python tool for data science, data analysis, and machine learning activities is called Pandas.

Sklearn:
The most practical Python tool for machine learning is undoubtedly scikit-learn. Numerous effective methods for machine learning and statistical modelling, such as classification, regression, clustering, and dimensionality reduction, are available in the sklearn package.

Tornado:
Python's Tornado is a network tool and web platform. Non-blocking network-io is used by Tornado. As a result, it is capable of managing thousands of live server links. It is a lifesaver for apps that require prolonged polling and a high volume of kept links.

Psutil:
Psutil is a cross-platform Python module used to obtain system information and utility processes. It is used to monitor how the system's different resources are being used. Monitoring the use of resources like the CPU, RAM, discs, network, and devices is possible.



Libraries and Module for Algorithm Development Using Python

- asttokens==2.0.5
- backcall==0.2.0
- click==8.0.4
- colorama==0.4.4
- debugpy==1.5.1
- decorator==5.1.1
- entrypoints==0.4
- executing==0.8.3
- Flask==2.0.3
- ipykernel==6.9.2
- ipython==8.1.1
- itsdangerous==2.1.1
- jedi==0.18.1
- Jinja2==3.0.3
- joblib==1.1.0
- jupyter-client==7.1.2
- jupyter-core==4.9.2
- MarkupSafe==2.1.1
- matplotlib-inline==0.1.3
- nest-asyncio==1.5.4
- numpy==1.22.3
- pandas==1.4.1
- parso==0.8.3
- pickleshare==0.7.5
- prompt-toolkit==3.0.28
- psutil==5.9.0
- pure-eval==0.2.2
- Pygments==2.11.2
- python-dateutil==2.8.2
- pytz==2022.1
- pywin32==303
- pyzmq==22.3.0
- scikit-learn==1.0.2
- scipy==1.8.0
- six==1.16.0
- sklearn==0.0
- stack-data==0.2.0
- threadpoolctl==3.1.0
- tornado==6.1
- traitlets==5.1.1
- wcwidth==0.2.5
- Werkzeug==2.0.3

Conclusion

In this project we successfully deployed a website which can be used to predict heart disease risk level by taking patient detail as input.

We used some libraries provided by Python and html, CSS and bootstrap to implement this project. After the experiments, the algorithm of Multivariable Polynomial Regression gives us the best test accuracy, which is 75.8%. The reason why it outperforms others is that it is not limited to the property of the dataset.

Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Though we get a good result of 75.8% 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 other algorithms to make accurate predictions.

References

[1] Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-48.

[2] Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10
(2012): 44-48.

[3] Uyar, Kaan, and Ahmet İlhan. "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks." Procedia computer science 120 (2017): 588-593.

[4] Kim, Jae Kwon, and Sanggil Kang. "Neural network-based coronary heart disease risk prediction using feature correlation analysis." Journal of healthcare engineering 2017 (2017).

[5] Baccouche, Asma, et al. "Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico." Information 11.4 (2020): 207.

[6] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[7] https://www.kaggle.com/ronitf/heart-disease-uci

[8] https://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

[9] https://www.kaggle.com/jprakashds/confusion-matrix-in-python-binaryclass

[10] scikit-learn, keras, pandas and matplotlib

CERTIFICATE OF APPROVAL

This is to certified that the thesis entitled **Heart Disease Risk Level Predictor** submitted by **Nisha Gupta Enrollment No. MGCU2019CSIT3011 & Nishant Raj, Enrollment No. MGCU2019CSIT3012** to Mahatma Gandhi Central University Motihari, Bihar for the award of the degree of Master of Technology in (Computer Science and Engineering) has been accepted by the Internal assessment committee and that the student has successfully defended the thesis in the Viva-voce examination held today.

13/03/23

(Supervisors) **(External Examiner)** **(Head of Department)**

Nisha Gupta

Nishant Raj

Heart Disease Risk Level Predictor

B.Tech.
(CSE)

2019-23

Department of Computer Science and Information Technology MAHATMA GANDHI CENTRAL UNIVERSITY,
MOTIHARI,
BIHAR-845401,
India