Heart Disease Risk Level Predictor

A Report on Final Year Project Submitted
in
Partial Fulfillment of the Requirements for the Award of the Degree
of

Bachelor of Technology
in
Computer Science and Engineering
by

***** Shivansh Sagar (18105111016) *****

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DARBHANGA COLLEGE OF ENGINEERING

Under the guidance of:
Prof. Ajeet Kumar Gupta
Assistant Professor, CSE Dept.
Darbhanga College of Engineering.

JULY 2022

DARBHANGA COLLEGE OF ENGINEERING DARBHANGA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
Lalsahpur, Mabbi – 846005, Darbhanga

****

BONAFIDE CERTIFICATE

Certified that this Project titled "**Heart Disease Risk Level Predictor**" is the Bonafide work of **Shivansh Sagar (18105111016)** who carried out the work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of anyother project or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other student.

Prof. Ajeet Kumar Gupta
Project Co-Ordinator
Computer Science and Engineering
DCE Darbhanga

Prof. Ajeet Kumar Gupta
Project Guide
Computer Science and Engineering
DCE Darbhanga

Prof. Ajeet Kumar Gupta
Professor and HOD
Computer Science and Engineering
DCE Darbhanga

(External Name)

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 Engineering DARBHANGA COLLEGE OF ENGINEERING, Darbhanga** in partial fulfilment of requirements for the final year project. This is an authentic record of our work carried out during the period from **MAY 20, 2022 to JULY 31, 2022** under the guidance of **Prof. Ajeet Kumar Gupta**, **Head of Department, C.S.E., DARBHANGA COLLEGE OF ENGINEERING, Darbhanga**, and is original, free from plagiarism and not copied from any source without proper citation.Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the project report and giving their details in the references. Whenever we have quoted written materials from other sources, we have put them under quotation marks and given due credit to the sources by citing them and giving the required details in the references. We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

This work has not previously formed the basis for the award of any degree, diploma, fellowship or any other similar titles or recognition. In case, plagiarism is detected at any stage, we shall be solely responsible for this.

Name	Roll no.	Reg. No	Signature
SHIVANSH SAGAR	**18-CS-01**	**18105111016**	_____

PROF. AJEET KUMAR GUPTA
(Head of Department, CSE)
DCE, Darbhanga

Acknowledgment

I wish to place on record my deep sense of gratitude to our honorific Guide **Prof. Ajeet Kumar Gupta,** Assistant Professor, Computer Science and Engineering, DCE Darbhanga 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 would like to thank **Prof. Ajeet Kumar Gupta (HOD)** of **Computer Science and Engineering**, **DCE Darbhanga** who has motivated us to work harder and do our best. Last but not least, we would like to owe our sincere and incessant gratitude to the almighty God for the immense blessing on us.

I want to thank all my teachers for providing a solid background for my studies and research thereafter. They have been great sources of inspiration to me, and I thank them from the bottom of my heart.

Above all, I would like to thank all my friends whose direct and indirect support helped me complete our project in time. I wish to dedicate this work to my PARENT, for they are the pillars of support giving me confidence in whatever I do.

SHIVANSH SAGAR

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

LIST OF FIGURES

LIST OF TABLES

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 system software that manages computer hardware, software resources, and provides common services for computer programs.

- **Network (Wi-Fi Internet or cellular Network)**
  A network is a collection of computers, servers, mainframes, network devices, peripherals, or other devices connected to allow data sharing

- **Visual Studio Code (Create and design data flow and Context Diagram)**
  Visual Studio Code is a source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

- **GitHub (Versioning Control)**
  GitHub, Inc. is a provider of Internet hosting for software development and version control using Git. It offers the distributed version control and source code management functionality of Git, plus its own features.

- **Google Chrome (Used for hosting website and system testing)**
  Google chrome is a web browser application software for accessing the World Wide Web or a local website. When a user requests a web page from a particular website, the web browser retrieves the necessary content from a web server and then displays the page on the user's device.

- **Jupyter Notebook**
  The Jupyter Notebook is an open source web application that can be use to create and share documents that contain live code, equations, visualizations, and text. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

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)

We just downloaded a dataset from Kaggle. We have split the dataset into 90% (5980 instances) for training and 10% (664 instances) for tests. We used the dataset to train the multivariable polynomial regression model to calculate the heart disease risk percentage using the above features.

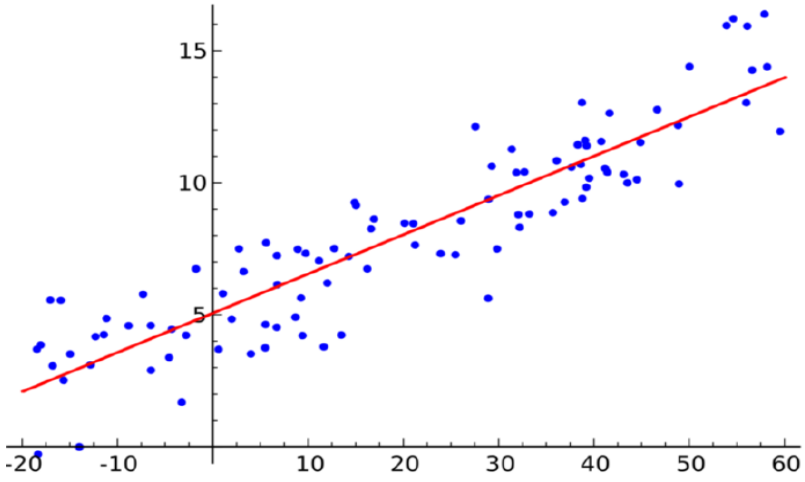
	A	B	C	D	E	F	G	H	I
1	SEX	AGEIR	TC	HDL	SMOKE_	BPMED	DIAB_01	RISK	
2	2	48	236	66	0	2	0	1.1	
3	1	48	260	51	0	2	1	7	
4	1	44	187	49	1	2	0	7	
5	2	42	216	57	1	2	0	0.4	
6	2	56	156	42	0	2	0	2.2	
7	1	44	162	57	1	2	0	3	
8	1	50	244	47	0	2	0	4.2	
9	1	48	212	30	1	2	0	17.4	
10	2	66	202	53	0	2	1	13.4	
11	1	63	186	46	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.

$$y = m*x + c$$

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

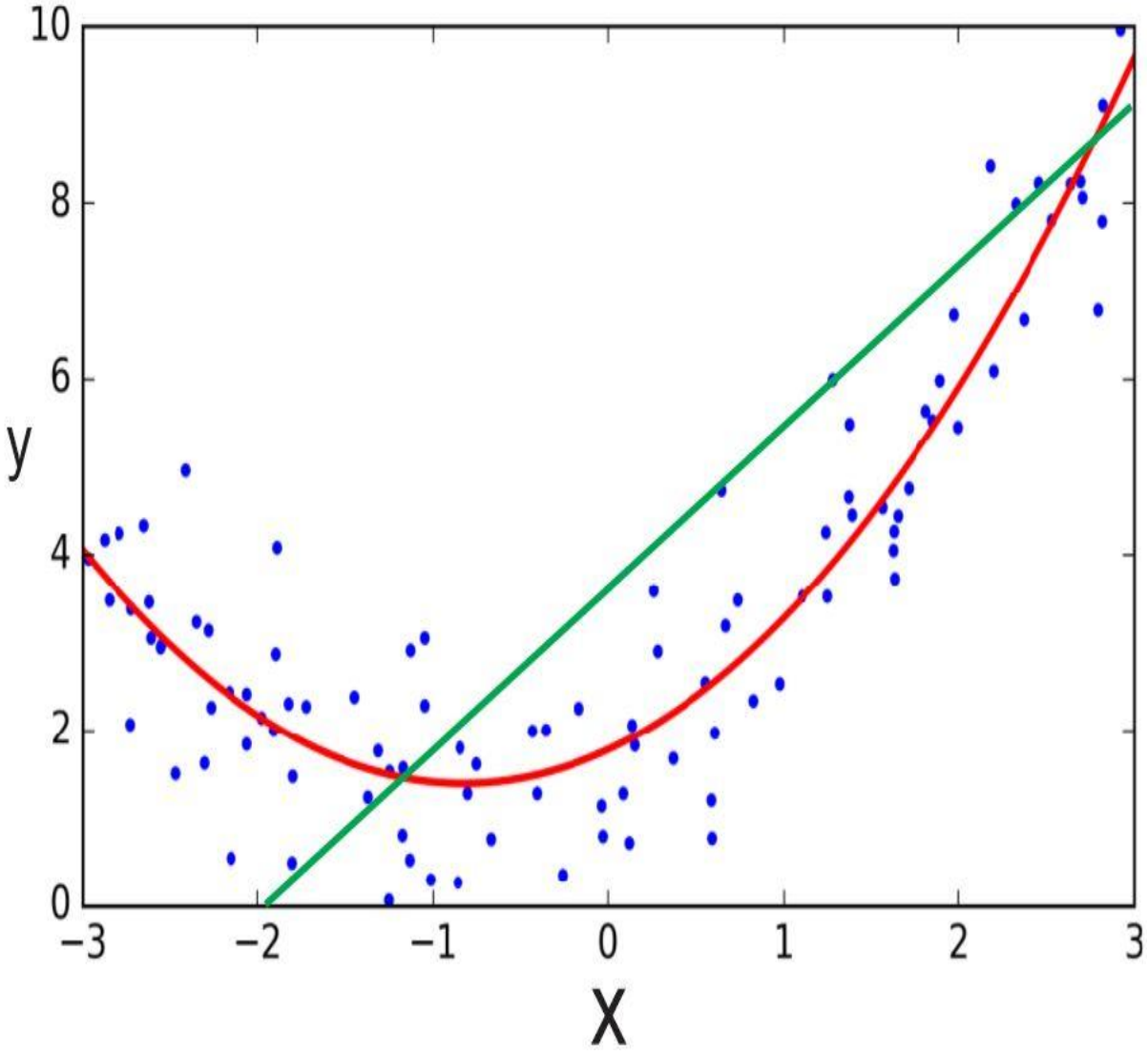
$$b = \overline{y} - m\overline{x}$$

Multivariable Polynomial Regression

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

- Like many other things in machine learning, polynomial regression as a notion comes from statistics. Statisticians use it to conduct analysis when there is a non-linear relationship between the value of xx and the corresponding conditional mean of yy.

- Imagine you want to predict how many likes your new social media post will have at any given point after the publication. There is no linear correlation between the number of likes and the time that passes. Your new post will probably get many likes in the first 24 hours after publication, and then its popularity will decrease.



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.

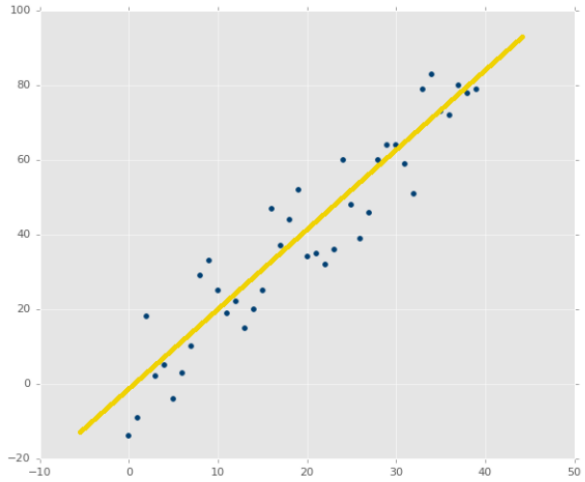


Figure 1

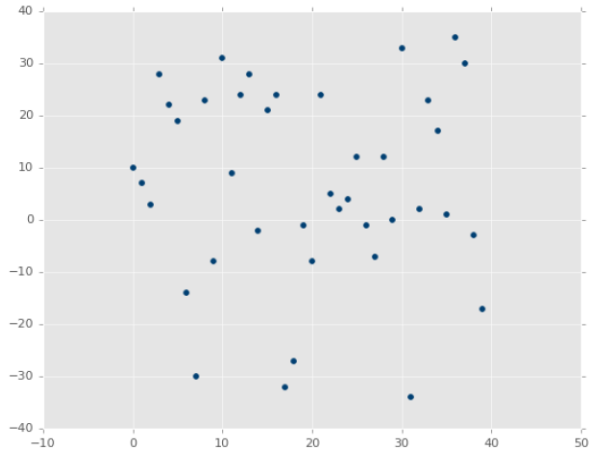
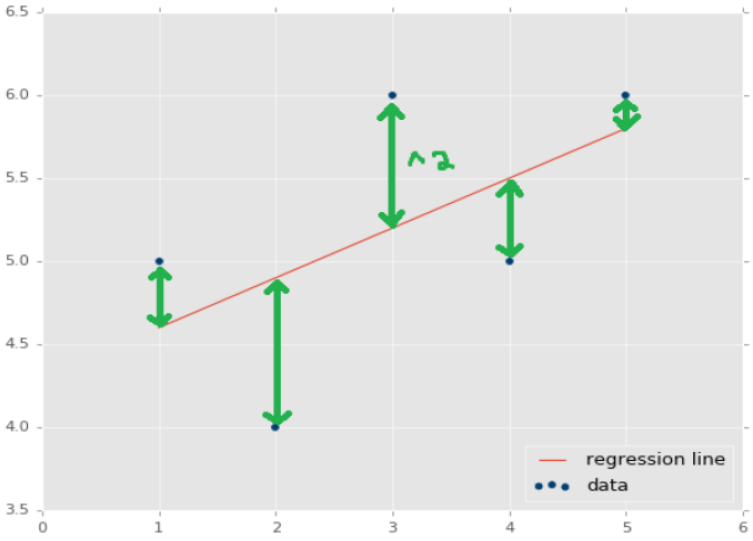


Figure 2

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.

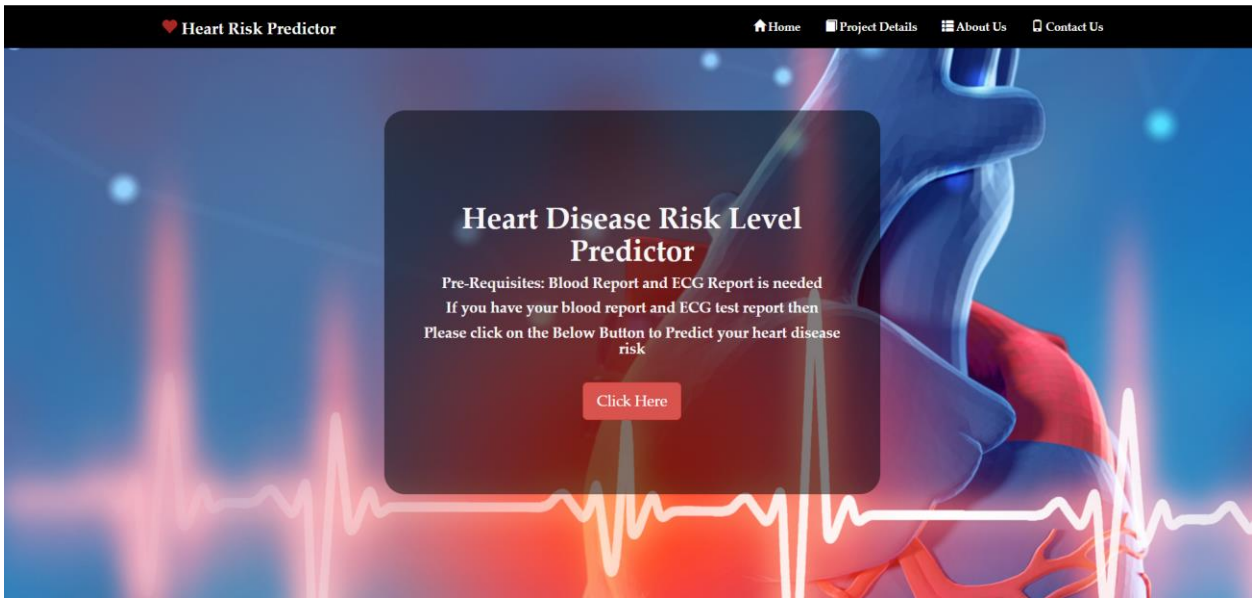
$$r^2 = 1 - \frac{SE\hat{y}}{SE_{\bar{y}}}$$



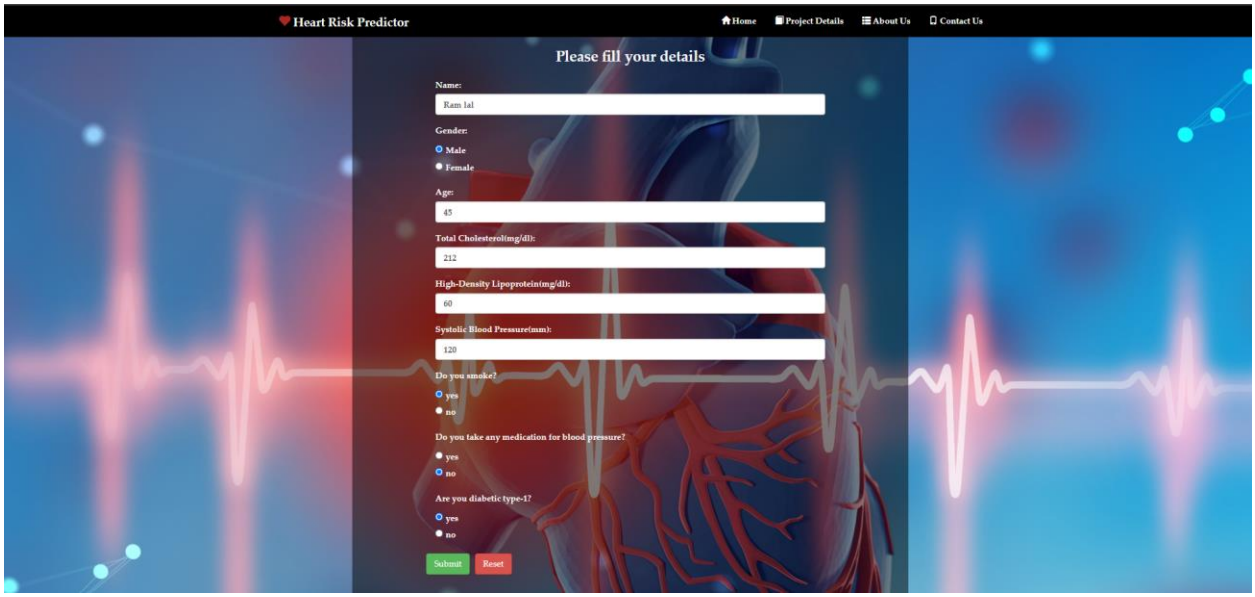
Implementations and Results

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.



- **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.



- **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.

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

Libraries imported for implementing the project:

Flask:
Flask is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python easier. It gives developers flexibility and is a more accessible framework for new developers since you can build a web application quickly using only a single Python file.

Matplotlib:
Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open-source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications.

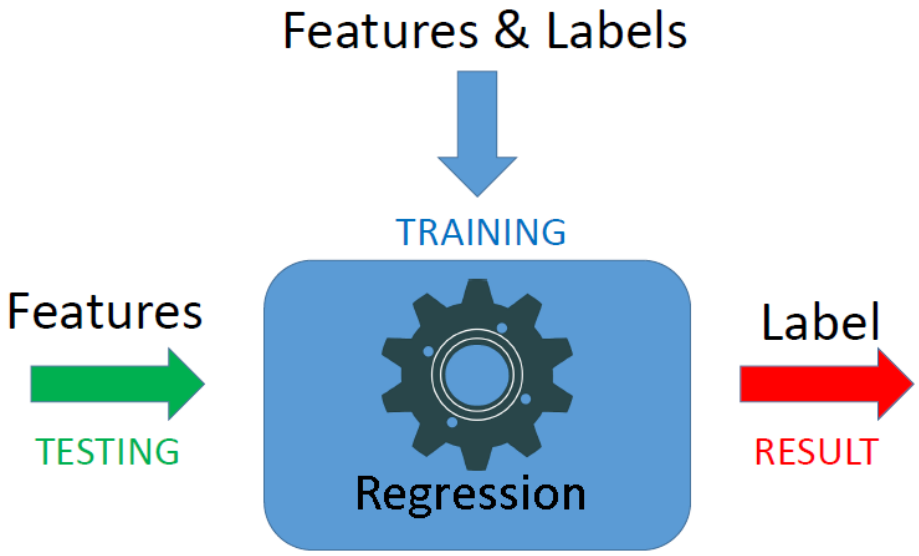
NumPy:
NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

Pandas:
Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks.

Sklearn:
Scikit-learn is probably the most useful library for machine learning in Python.The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction.

Tornado:
Tornado is a Python web framework and asynchronous network library. Tornado uses non-blocking network-io. Due to this, it can handle thousands of active server connections. It is a savior for applications where long polling and a large number of active connections are maintained.

Psutil:
Psutil is a Python cross-platform library used to access system details and process utilities. It is used to keep track of various resources utilization in the system. Usage of resources like CPU, memory, disks, network, sensors can be monitored.



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