ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2107

Predicting Coronary Heart Disease using Various Regression Analysis

Varun Chavan[1]
Department of Artificial Intelligence and Data Science
Thadomal Shahani Engineering College
Mumbai, India

Niyati Doaj[2]
Department of Artificial Intelligence and Data Science
Thadomal Shahani Engineering College
Mumbai, India

Naveen Vaswani[3]
Department of Artificial Intelligence and Data Science
Thadomal Shahani Engineering College
Mumbai, India

Abstract:- Nowadays, cardiovascular diseases are the major concern for Human beings. Cardiovascular Heart Diseases (CHDs) are the major reason for mortality globally causing millions of deaths each year. The global death toll from Cardiovascular Heart Disease (CHD) rose from 12.1 million in 1990 to 20.5 million in 2021. The age-standardized death rate for CHD in India was 282 deaths/100000 was higher compared with global levels that is 233 deaths/100000. The leading factors contributing to cardiovascular diseases or fatalities encompass high cholesterol level, triglyceride, etc. These are the most common factors nowadays which result in heart attack in humans. The main aim of this particular project is to explore various risk factors associated with myocardial infarction, commonly known as Heart Attack. We have used tensorflow and Keras frameworks for the following project. In this project, we utilize neural network models to forecast the risk of coronary heart disease (CHD) using various features. The prediction process involves training the models on a portion of the data and assessing their effectiveness on the remaining test set. These models are designed to discern patterns and correlations between input features and the target variable, 'chd'. The choice of features, architecture, and training parameters influences the model's predictive performance. The research focuses on two distinct neural network models: the first, 'sbp_model,' predicts CHD using only systolic blood pressure ('sbp'), while the second, 'linear_model,' utilizes all available features after normalization. Both models are evaluated on their ability to predict CHD through mean absolute error, with training histories and loss curves analysed. We have taken into consideration all the important Regression Models.

Keywords:- *Cardiovascular Diseases, CHD, Regression Model.*

I. INTRODUCTION

Heart is one of the most vital organ of human body. The main role of the heart is to facilitate the circulation of blood throughout the body. Blood transports oxygen and vital nutrients to cells, concurrently eliminating carbon dioxide and other waste materials, thereby enabling other organs to expel them. Nowadays there are many myocardial infection happening which causes heart attack and which may also be dangerous for the life of an individual.To observe the health rate of an individual it is important to get known about the causes of heart attack. Here we take into consideration heart related issues and assume the rate of heart attack according to the individual. Cardiovascular diseases are one of the major concern nowadays(resulting in 30% of all deaths). With this paper we propose to work towards decreasing heart attack rate and aim to design a machine learning based classification model by analyzing different heart related issues addresses abd collected from the people outside. This model will basically tell you what needs to be done to decrease the rate of heart attach for the individual.

There are many such different attributes for heart attack. Our model will try to reduce it. We will be taking the same dataset for multiple regression solutions which will be helpful to achieve the desired output. This technique is very useful as it shows us in multiple data visualization ways. . In fields like financial, medical diagnosis, agricultural, ecommerce, manufacturing, customer service etc, feature adjustment is a popular practice.

We will be doing multiple regression and will be finding out which suits best for this particular cardiovascular disease dataset. In medical domain data visualization proves to be very useful.Different regression models will give different answers and best would be the most efficient out of all.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2107

II. LITERATURE SURVEY

It is known that predicting cardiovascular diseases with a cheaper and more reliable method is a big challenge for poor people. Based on previous studies, there were cases where they predicted cardiovascular disease using various regression techniques. One of them was [1] "Logistic Regression Technique for Predicting Cardiovascular Disease" on June 9, 2022. Their goal was to avoid the worsening effect of disease in the pharmaceutical industry. They selected only highly positive correlation properties. Logistic regression has an accuracy of 87.10% when the ratio of training to testing is 90:10. The results were better than previous studies.

[2] On April 28, 2023, a research paper on "Predicting heart disease using chi-square test and linear regression" was published. This article focused primarily on people living in the United States, which has a higher percentage of people with heart disease. They conducted both a chi-square test and linear regression analysis to anticipate heart disease, considering symptoms like chest pain and dizziness. The examination outcomes indicated that men were at a higher probability of having heart disease, exhibiting symptoms such as chest pain, dizziness, shortness of breath, fatigue, and nausea. It also showed that 90 percent of people with chest pain have heart disease. They tested the data using a logistic regression prediction model and found an accuracy of 85.12 percent.

In 2022, [3] published a research article, "Accuracy Analysis of Heart Disease Prediction Using Logistic Regression Compared to Linear Regression Algorithm". Its main purpose was to use logistic regression to detect heart disease. Logistic regression method with 30 samples and linear regression model with 30 samples are used to detect heart diseases. The test results were: The accuracy of logistic regression is 88.68 percent higher than the accuracy of linear regression model is 78.56 percent. Thus, it was concluded that logistic regression gives better results in terms of accuracy compared to linear regression model.

III. METHODOLOY

- *Identify Coronary Heart Disease (CHD) Dataset CHD Dataset consists of records of 462 white males aged between 15 to 64, out of which 160 of them have Coronary Heart Disease or Myocardial Infarction (MI).*

The dataset, Coronary Heart Disease (CHD) is obtained from the Coronary Risk Factor Study conducted in South Africa by Rousseauw et al. in 1983.

- *Pre-Processing*

Dataset had been normalized before being used for Coronary Heart Disease prediction using linear regression model. All the attributes were scaled down to a range between 0 to 1. Here we take into consideration minimum value and maximum value of feature represented as min(x) and max(x) respectively.

Normalization is done using

$$x_{\text{normalise}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- *LRM*

Linear Regression Model (LRM) is a statistical Method employed to examine the association between A dependent and one or more independent variable. It Anticipated the value of dependent variable by Considering values of independent variable. Simple linear regression and Multiple linear regression Are distinct from each other.Simple linear regression Takes into consideration one independent variable Whereas multiple regression takes many independent Variables. Relationship between variables is taken into Consideration and best-fit line equation is determined. This equation can accurately predict the values.

- *Simple Linear Regression Formula :*

$Y=\beta 0 +\beta 1X$

(Y is dependent variable and X in independent)

Ridge Regression is an extension of linear regression that adds a regularization term which is used to overcome the overfitting the model on training data and the problem of multicollinearity.

- *General Formula:-*

$Y= X0 + B1X1 + B2X2 + \ldots + BnXn$

Where,

(Here Y is predicted value and X is independent variable)

X0 is value of dependent variable when independent value equals zero

In **polynomial regression** the nth degree polynomial equation is used to express the connection between the variables. This method permits the modelling of a more flexible curve compared to linear regression, accommodating nonlinear relationships between the variables.

In the the general formula of polynomial regression

Dependent variable(y), independent variable(x) ,

Coefficients and error term is considered.

- *General Form:-*

$y = \beta 0 + \beta 1x + \beta 2x^2 + \beta 3x^3 + \ldots + \beta nx^n + \varepsilon$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2107

➢ *Accuracy Prediction:*

It evaluates the model's capacity to precisely predict the values of the dependent variable by Based on the independent variable(s). There are various methods to calculate the accuracy,

Mean Absolute Error (MAE) quantifies the typical disparities between the observed value and the forecasted value..

- *Formula to Calculate Accuracy:-*

MAE = (1/n) Σ(i=1 to n) |yi – ŷi|

Where,

n is the number of observations yi represents actual value

ŷi represents predicted value

RMSE (Root Mean Square Error) is used to calculate the difference between predicted and observed value then is divided by the number of observations and the square root of the result is taken into consideration. It is basically use to take a look at the precision of the model.This metric provides insight into the degree to which the forecasts generated by the model align with the actual observed values.

- *Formula to Calculate RMSE:-*

RMSE = sqrt(1/n $\sum^{n}$i=1 (yi – ŷi)^2) n = number of observations.

Yi=actual value of the dependent variable.

Ŷi=predicted value of the dependent variable.

➢ *Use of TensorFlow and Keras:*

TensorFlow is an open-source machine learning framework created to simplify the development, training, and deployment of machine learning models. Keras, on the other hand, serves as TensorFlow's high-level API, offering a user-friendly interface for constructing and training machine learning models.

➢ *Graph Representation of Simple Linear Regression Model*

- *Simple Liner Regression:*

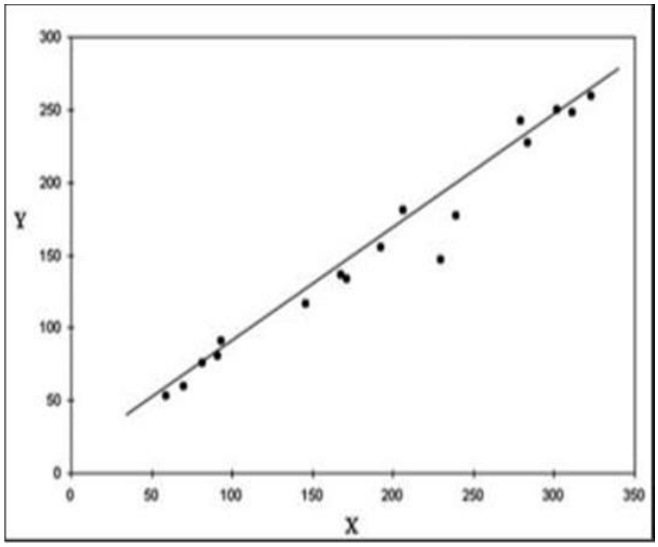

Fig 1 Simple Liner Regression

In this graph, X is the independent variable and Y is the Dependent variable

For this paper, we have taken X as sbp (Sytolic Blood Pressure) and Y as CHD(Coronary Heart Disease)

- *Ridge Regression:*

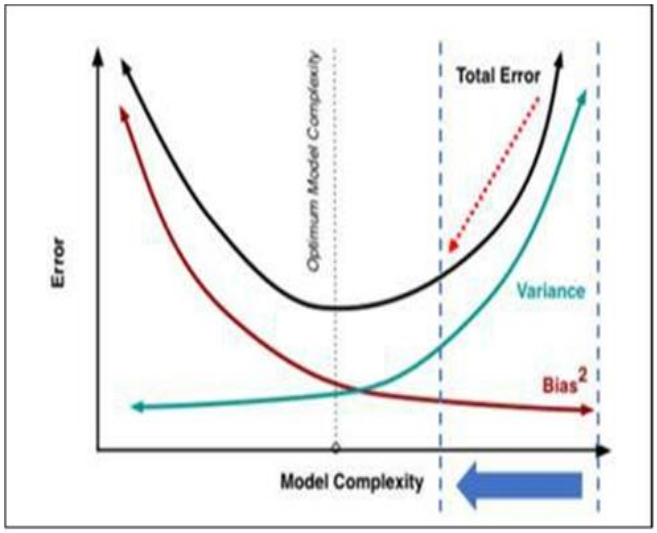

Fig 2 Ridge Regression

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2107

- *Polynomial Regression:*

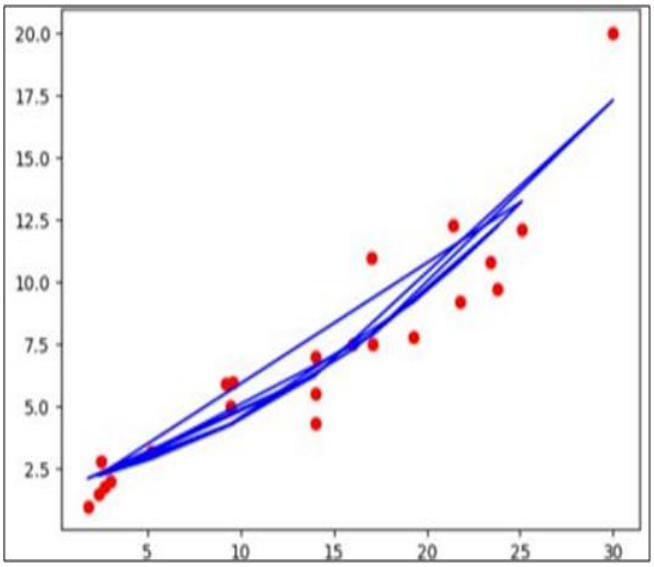

Fig 3 Polynomial Regression

Logistic Regression: Logistic regression is a statistical technique employed to model the likelihood of a discrete outcome, often binary, based on one or more input variables.

- *Pair Plot which Contains all Variables:*

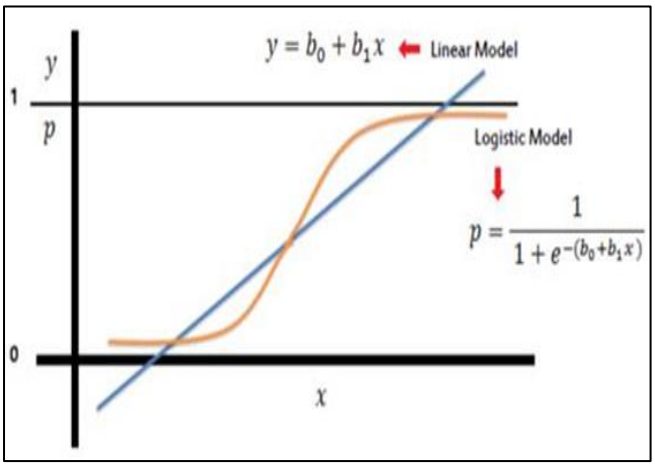

Fig 4 Pair Plot which Contains all Variables

IV. DATASET ANALYSIS

Here we will be looking at dataset with 10 different variables.This dataset is about Coronary Heart disease and all the 10 variables undertaken are the major reason a person can get a heart attack.This data is obtained from the Coronary Risk Factor Study conducted in South Africa by Rousseauw et al. in 1983. The goal is to get to know how one factor can lead to a heart disease.There are 462 observations in total with 160 individuals experiencing an Myocardial Infarction(MI) and 302 without an Myocardial infarction. Data cleaning is done onto the dataset to avoid obstacles while implementing. Also Standardization and Normalization is also applied for smooth implementation.

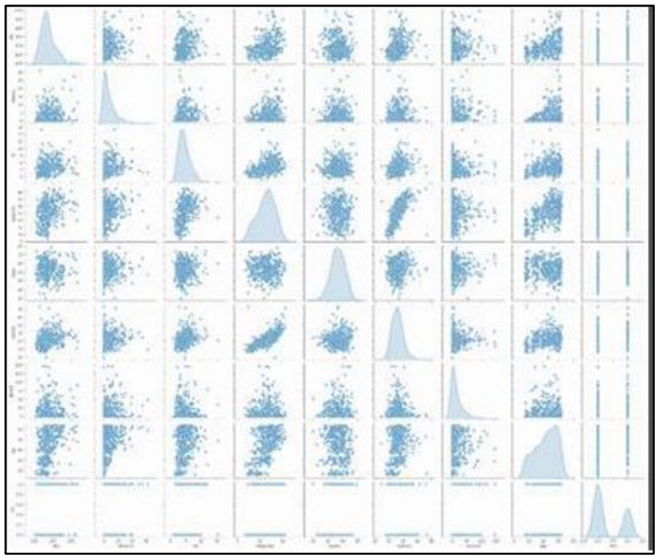
Fig 5 Dataset Analysis

V. RESULTS

➢ *Mean Absolute Error and Accuracy Table:*

Table 1 Mean Absolute Error and Accuracy

Model	Linear regression	Logistic regression	Ridge regression	Polynomial regression
Mean Absolute error(MAE)	0.348445	0.315217	0.375763	0.366607
Accuracy	71.84%	41.52%	72.42%	73.33%

➢ *The Variables in Dataset are as Follows:-*

Table 2 The Variables in Dataset

Variables	Description
	Systolic blood pressure
sbp	
	Cumulative tobacco
tobacco	
	Low density lipoprotein Cholestrol
Idl	Level
	Severe Overweight
adiposity	
	Family History of Heart Disease
famhist	
	Type-A Behavior
typea	
	Excessive fat accumulation
obesity	
	Current alcohol consumption
alcohol	
	Age at onset
age	
	Response, coronary heart disease
chd	

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2107

The main aim of the research paper was to predict the Cardiovascular Disease in white males aged to 15 to 64 in South Africa in 1983.The Objective was to explore the risk factors for the same. As we started the research paper by extracting a dataset from Kaggle. Then we cleaned and pre-processed the data. We applied Simple Linear Regression on the dataset for predicting the Cardiovascular Disease. We found the accuracy for the same by calculating Root Mean Square Error (RMSE) and it was 68.84 %. Then we applied Logistic Regression and found the accuracy to be 40.35% . Further we applied Ridge regression and found the accuracy to be 62.42 %. Finally ,we applied Polynomial Regression and it turned out to be the best result with the accuracy of 73.34%.

Let us see some graphical representation done by us in our code which tells us the predictions made by different models. All the graphical representations will show graph of Predicted CHD vs Actual CHD

Below are some graphical representations of different models proposed by us for this dataset

- *Logistic Regression:*

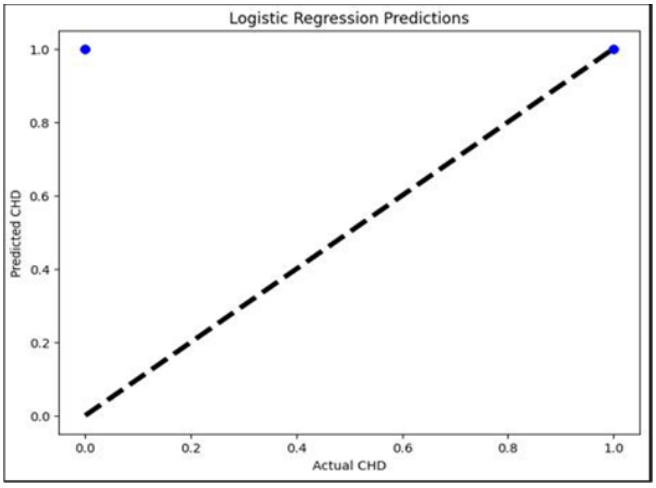

Fig 6 Logistic Regression

- *Ridge Regression:*

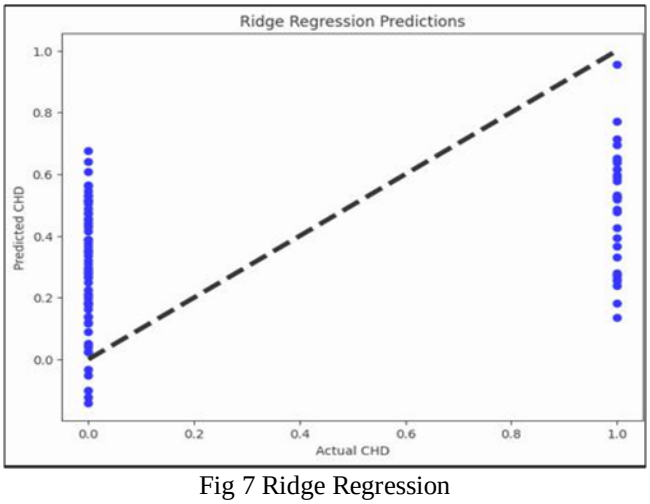

Fig 7 Ridge Regression

- *Polynomial Regression*

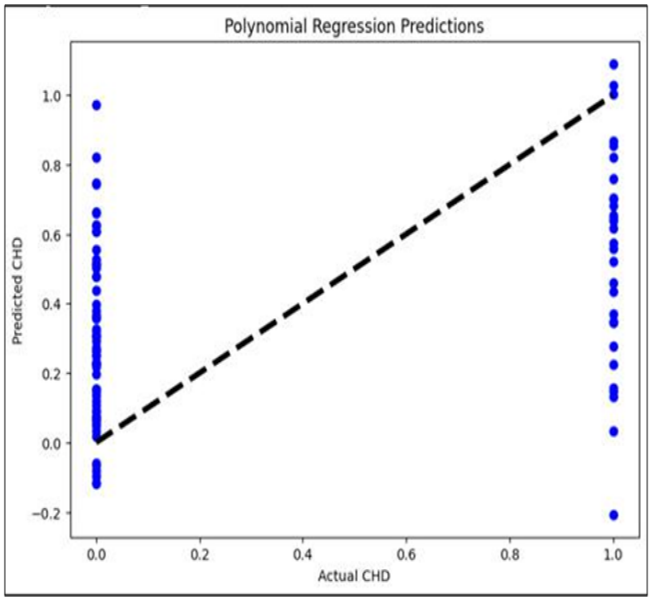

Fig 8 Polynomial Regression

VI. CONCLUSION

We have determined that various factors significantly influence cardiovascular disease. Additionally, we have constructed and tested several models including linear regression, ridge regression, polynomial regression, and logistic regression using machine learning algorithms. These models were trained and evaluated using datasets, encompassing both training and testing data to assess their predictive performance. The evaluation of model accuracy was conducted using Root Mean Square Error (RMSE), revealing variability in prediction accuracies among different regression models for Cardiovascular disease. The best accuracy predicted is by Polynomial Regression model which is 73.34 % compared to other regression models. So therefore, we can say that Polynomial Regression suits this study on the particular dataset preferred in this research paper. We can say, its better we use one or more algorithms to calculate because more algorithms means more results and then the best result can be considered.

REFERENCES

[1]. Arvind Chandrasekaran and Dinesh Kalla (2023) “Heart Disease Prediction using Chi-Square Test and Linear Regression”.
[2]. Ambrish G, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, Kiran Mensinkal (2022) “Logistic regression technique for prediction of cardiovascular disease”.
[3]. Tania Ciu and Raymond Oetama (2020) “Logistic Regression Prediction Model for Cardiovascular Disease”.
[4]. Cangao (Steven) Chu (2015) “Predicting Cardiovascular Disease based on Regression Analysis and Classification of Nhane Survey Statistics.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR2107

[5]. Cardiovascular Diseases Dataset (kaggle.com)
[6]. https://www.ibm.com/topics/logistic-regression
[7]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8898839/
[8]. www.google.com
[9]. Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath (2020) "Heart Disease Prediction using machine learning algorithms".
[10]. Mafia Rasheed, Muhammad Adnan Khan, Nouh Sabri, Ghassan Issa (2022) " Heart Disease Prediction Using Machine Learning Method".