Prediction of Blood Pressure and Cholesterol By Machine Learning Technique

Dip Das[1],Arnam Ghosh[2], Aishik Banerjee[3] , Mrs. Sulekha Das[4] ,
Dr. Avijit Kumar Chaudhuri[5]
[1]UG- Information Technology, Techno Engineering College Banipur
2UG- Computer Science and Engineering, Techno Engineering College Banipur
[3]UG- Computer Science and Engineering, Techno Engineering College Banipur
[4]Assistant Professor, Techno Engineering College Banipur
[5]Assistant Professor, Techno Engineering College Banipur
Author's Orcid ID: [1]0000-0002-7714-0533, [2]0000-0002-3232-8620, [3]0000-0003-2705-0537, [4]0000-0002-6641-3268, [5]0000-0002-5310-3180

1. ABSTRACT

Blood pressure (BP) is the pressure of circulating blood against the walls of blood vessels. Cholesterol is any of a class of certain organic particles called lipids [1]. High blood pressure along with high cholesterol are two of the main causes of heart disease. Heart disease, stroke, and other cardiovascular (blood vessel) diseases are among the primary cause of death.

Approximately 10 million people every year die having high blood pressure and it was found that 6 in 10 Indians have abnormal levels of blood cholesterol [2]. An increase in blood cholesterol signifies an increase in a man's blood pressure, which can easily lead to abnormal blood pressure,e.g. high blood pressure.

Generally, Prediction refers to the output of an algorithm after it has been trained on a dataset. Here Authors tried to predict whether a sample can have blood pressure and cholesterol or both present or not by analyzing some datasets. Generally, for making these types of predictions, machine learning algorithms are a very appropriate and most-used technique.

A Multiple Linear Regression was applied in this study. Data has been analyzed through Multiple Linear Regression Analysis (MRA). The proposed model is tested on a Primary Data Set, prepared by Techno Engineering College Banipur's students, and the relevant data of this Dataset has been collected from some districts in West Bengal, India. Through this research, approximately 76 percent of the data was predicted correctly.

Keywords: Blood pressure, Cholesterol, Predictions, Multiple Regression, Primary Dataset.

2. INTRODUCTION

Statistics is the most widely used aspect which is used in science, medicine, Blood Pressure, fisheries, etc. There are many methods in statistics but one of the most useful methods is regression. Regression is six types which are simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression, and discriminate analysis. In this project, the author used multiple linear regressions for the analysis of high blood pressure and cholesterol. In multiple linear regressions(MLR) the dependent variable is only one and the independent variable is two or more than two [3]. The objective of this project is to find a suitable model and the reason behind High Blood Pressure and how Blood Pressure increases for some reason.

Worldwide, elevated blood pressure (BP) is the cause of death, even exceeding deaths attributable to smoking and elevated cholesterol. Blood pressure is a measure of the force that your heart uses to pump blood around your body. Many events occur within the body as the heart pumps blood, known as the cardiac cycle, and so pumps blood is measured at different points throughout this cycle. Systolic blood pressure measures the maximum pressure in the arteries during the cardiac cycle, which occurs when the heart contract or beat to pump blood. Stress-related habits such as eating more, using tobacco, or drinking alcohol can increase blood pressure [4]. Certain chronic conditions. Kidney disease and diabetes are some of the conditions that can lead to high blood pressure. Blood pressure is measured in units of millimeters of mercury (mmHg). Blood pressure is measured as two numbers:

- Systolic blood pressure (the first and higher number) measures the
the pressure inside your arteries when the heart beats.
- Diastolic blood pressure (the second and lower number) measures the pressure inside the artery when the heart rests between beats[5].

A normal blood pressure level is considered to be between 90/60mmHg and 120/80mmHg.If
it is increased up to 140/90mmHg or higher then it is called High Blood Pressure. The overall prevalence of hypertension among adults in India is about 30% with an urban prevalence of 34% and a rural prevalence of 28%. Unfortunately, only 25% of rural and 38% of urban Indians with hypertension are being treated. In the blood having a High Cholesterol level can increase the risk of stroke & heart disease. High Blood Pressure and High Cholesterol are linked [6]. Lack of sleep, excess salt, excess body weight, smoking, alcohol, and air pollution are the risk factors for high Blood Pressure. Usually, lifestyle &genetic factors are the causes of High Blood Pressure. According to the WHO 2008 estimates, the prevalence of raised BP in Indians was 32.5% (33.2% in men and 31.7% in women). However, only about 25.6% of treated patients had their BP under control [7]. A systematic review of the prevalence of HTN in India, for studies published between 1969 and July 2011, reported a range between 13.9 to 46.3% and 4.5 to 58.8% in urban and rural areas of India, respectively. The regional variations (between urban and rural) reported in the prevalence of HTN are also seen in cardiovascular diseases. Age is one of the major factors that affect the blood pressure of an adult. Age comes along with growth, development, and death. Age brings about accumulation and depreciation and with this greater risk of medical instability, age brings about a change in blood pressure.

In this paper, a machine learning method is applied to detect whether a sample suffered from high blood pressure, Cholesterol, or both or not. First, SVM Algorithm has been applied after the author recursively checked by adding one by one feature as Feature Selection Techniques. After that Multiple Linear Regression Model has been developed, which trained on 80%, 66%, and 50% of the data Simultaneously. To deliver the result more accurately 10 Fold Cross Validation was also applied here.

3. LITERATURE REVIEW

Machine learning is a simple way that helps to predict the data easily. Multiple Linear Regression is an applied math technique, the target of Multiple linear regression analysis is to use the independent variables whose worth square measure acknowledged to predict the worth of the only dependent value. The main aim of this research is prediction High Blood Pressure and Cholesterol.

R. E. Ogunsakin had done research(2012) presented the relationship between Age & Blood Cholesterol on systolic Blood Pressure. The data used for this paper were obtained from Ekiti State University Teaching Hospital. They compared their system with the Multiple Linear Regression algorithm and found that their proposed method performed better. Meanwhile, they showed details about the variation of the maximum and the minimum values of the accuracy with the classification set sizes and the number of neighbors.

Christopher JL Murray had done research(2003) declaring that Cardiovascular disease is a major contributor to the global burden of disease. Interventions, Regions, Intervention effects, and Sensitivity Analysis is used in this paper. This dataset includes Total cholesterol concentrations over 3·8 mmol/L accounting for about 18% of strokes and 55% of cases of ischaemic heart disease.

LJ. Appel had done research(2008) that applied risk factors for elevated BP including excess salt intake, low potassium intake, excess weight, high alcohol consumption, and suboptimal dietary pattern. The sensitivity, specificity, and accuracy of their algorithm were shown to be 96%, 9,4%, and 94% respectively.

Those and many others had done their research prediction using Multiple Linear Regression.

We have gathered some specific ideas about machine learning and Multiple Linear Regression. So we were very interested in doing a project based on it. So we collected some real-life data on the "Heart Disease Data Set" and try to predict Blood Pressure and Cholesterol using independent fields like Age, taking medicine for the lever, Smoking, Exercising at a gym, and taking junk food, etc.

4. METHODOLOGY

DATA: In this paper Primary dataset has been used. It's a big challenge for finding the right people and collects all the relevant information like his/her daily life activities, health conditions, food habits, professional work, etc. Nearly 135 samples and 45 attributes are collected from the author's end for this dataset.

DATA FILTERING: Though it's a primary dataset, there are so many difficulties that have to be faced, like some fields are blank and some of the fields have been filled by garbage value. So to successfully perform the research works first step is to clean the data. Here authors filled the field by putting the mean value of this attribute's value.

Finds Most Relevant Attributes: After cleaning the data, the processing part of the dataset has been completed. After that author finds the most relevant and mostly dependent attributes for having high blood pressure and cholesterol by applying SVM Algorithm to the 44 attributes(the remaining 1 is Predicting Attribute). Then author recursively checked by adding them one by one attributes and finally it was found that from the 44 attributes only 13 attributes are the most relevant attribute, which has a direct impact on having high blood pressure and cholesterol.

4.1 **Table 1: The attributes along with datatypes**:

Attributes	Datatypes	Mean Value of these Attribute
'Do you take medicine for Lever'	Integer	4.7462
'Age'	Float	40.9328
'Do you like to take salt in your food?'	Integer	1.7238
'Mode of Transport	Integer	3.3805
'Stops smoking recently?'	Integer (0/1)	1.1044
'Do you frequently take alcohol?'	Integer(0/1/2/3)	1.3283
'Do you take medicine for Kidney?'	Integer	4.8283
'Are you trying to give up taking junk food'	Integer	4.5895

‘Any family history of coronary heart disease?’	Integer(1/2)	1.6641
‘Exercising at a gym/free hand regularly?’	Integer(1/2/3)	2.1119
‘How time do you socialize through Social Media(e.g.facebook, Twitter)?’	Integer	3.1818
‘Do you think you need to reduce some weight?’	Integer(1/2)	1.4328
‘Are you trying to give up taking meat?’	Integer	4.7014

5. RESEARCH METHOD:

Multiple linear regression is the method of statistics in regression that is used to analyze the relationship between a single response variable (dependent variable) with two or more controlled variables (independent variables). This method was selected for this research because there was more than one controlled variable. In this research, the response variable is **Have high blood pressure/high cholesterol/both(Y)** while the state (X1), ‘Do you take medicine for Lever’,(X2) ‘Age’, (X3) ‘Do you like to take salt additionally in the food you eat?’, (X4) ‘Mode of Transport’, (X5) ‘Stops smoking recently?’, (X6) ‘Do you frequently take alcohol?’, (X7) ‘Do you take medicine for Kidney?’, (X8) ‘Are you trying to give up taking junk food’, (X9) ‘Any family history of coronary heart disease?’, (X10) ‘Exercising at a gym/free hand regularly?’, (X11) ‘How time you socialize through Social Media(e.g. Facebook, Twitter)?’, (X12) ‘Do you think you need to reduce some weight?’ and (X13) ‘Are you trying to give up taking meat?’ are controlled variables.

5.1. Relevant equation & formula regarding multiple linear regression model:

Y = a+$b1X1$ +$b2X2$+ ....+$bnXn$

Where,

$b_i=\sum(x_i-x)(y_i-y)/\sum(x_i-x)2$[coefficient of the control variable]

$a = y_i - (b_i\, x_i + b_2x_2+$ .... $+ bnXn)$ [constant variable]

x_i = controlled variables (independent variables) x = controlled variables mean value

y = dependent variable mean value b1=Coefficient of the 1st control variable, b2=Coefficient of the 2nd control variable, b3=Coefficient of the 3rd control variable, b4=Coefficient of the 4th control variable b5=Coefficient of the 5th control variable b6=Coefficient of the6thcontrol variable b7=Coefficient of the7thcontrol variable b8=Coefficient of the 8thcontrol variable b9=Coefficient of the 9thcontrol variable b10=Coefficient of the 10th control variable b11=Coefficient of the 11thcontrol variable b12=Coefficient of the 12th control variable and
b13=Coefficient of the 13th control variable

x1=controlled variable (Do you take medicine for Lever)

x2=controlled variable (‘Age’)

x3=controlled variable (Do you like to take salt additionally in the food you eat?)

x4=controlled variable (Mode of Transport)

x5=controlled variable (Stops smoking recently?)

x6=controlled variable (Do you frequently take alcohol?)

x7=controlled variable (Do you take medicine for Kidney?)

x8=controlled variable (Are you trying to give up taking junk food)

x9=controlled variable (Any family history of coronary heart disease?)

x10=controlled variable (Exercising at a gym/free hand regularly?)

x11=controlled variable (How time you socialize through Social Media(e.g.facebook, twitter)?)

x12=controlled variable (Do you think you need to reduce some weight?) x13=controlled variable (Are you trying to give up taking meat?' are controlled variables)

5.2. Confusion-Matrix :

After finding the accuracy of the difference between actual data and calculated data we did the Confusion Matrix. In this confusion matrix it can be seen that[2] we find the **TP** – which stands for '**TRUE POSITIVE'** means the accuracy of classified positive data, **TN** – which stands for '**TRUE NEGATIVE'** means the accuracy of classified negative data, **FP** – which stands for '**FALSE POSITIVE'**, means which remark that actual value is negative but predicted data is positive, **FN**– which stands for '**FALSE NEGATIVE'**means that actual data and the predicted data both are negative and append the TP, TN, FP, FN values in a 2*2 matrix(mat1). After that, we find the accuracy, sensitivity, precision, recall, and specificity. This matrix contains all the raw information about the predictions done by a classification model on a given data set.

5.3. Cross-Validation :

After finding the accuracy of the training-testing partition of 50%, 66%, and 80% data, the author did cross-validation. In this cross-validation process first, the author divides the whole list into 10 sub-list and then finds the accuracy of 10 sub-list elements, also finds the Confusion Matrix of each Sub-list and finds the accuracy, sensitivity, precision, recall, and specificity.

ACCURACY: It's the ratio of the correctly labeled subjects to the whole pool of subjects. Accuracy is intuitional.

PRECISION: Precision is the ratio of the correctly positive labeled data to all positive labeled data.

SENSITIVITY: Sensitivity means out of the total positive, what percentage are predicted positive?

SPECIFICITY: Specificity is calculated as the number of true negative predictions divided by the total number of negative predictions.

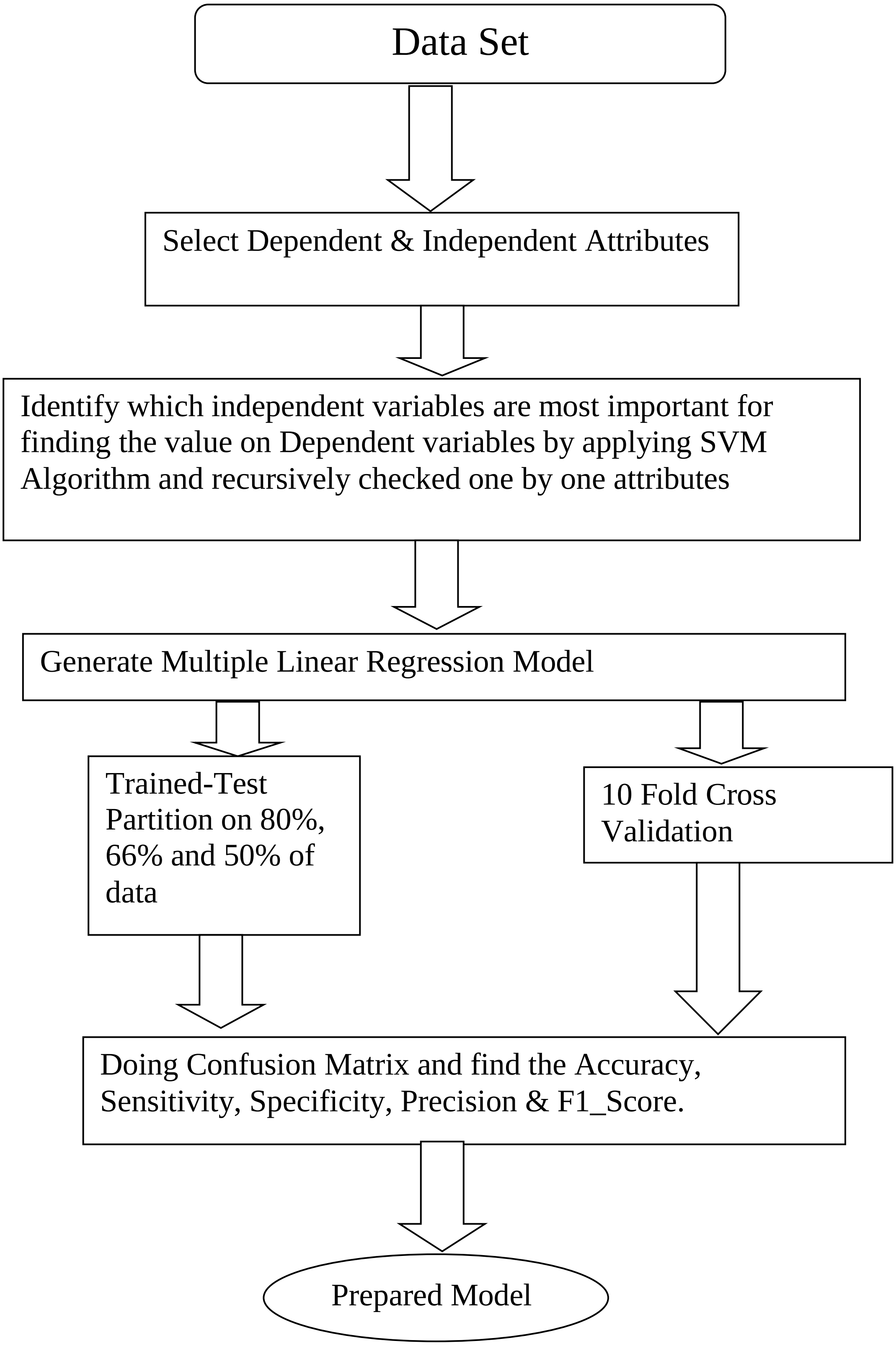
- **ACCURACY= (TP+TN/ TP+TN+FP+FN)*100**
- **PRECISION =(TP/FP+TP)*100**
- **SENSITIVITY=(TP/FN+TP)*100**
- **SPECIFICITY = (TN/TN+*FP*)*100**

F1 Score:- The F1 score is defined as the multiple of 2 of the arithmetic mean of precision and recall. It is very much useful for computing an average rate. In the F1 score, the author computes the average precision and recall. For, finding the F1 score authors have to find the first precession and recall.

Precession = TP / (TP+FP) Recall = TP /(TP+FN)

F1 Score = $(2*\mathit{Recall}*\mathit{precession}) / (\mathit{Recall}+\mathit{Precession})$

5.4) **Flow Chart:**


Data Set
Select Dependent & Independent Attributes
Identify which independent variables are most important for finding the value on Dependent variables by applying SVM Algorithm and recursively checked one by one attributes
Generate Multiple Linear Regression Model
Trained-Test Partition on 80%, 66% and 50% of data
10 Fold Cross Validation
Doing Confusion Matrix and find the Accuracy, Sensitivity, Specificity, Precision & F1_Score.
Prepared Model

6. RESULTS :

Table-2
Taking66Percent of Data as Training and 34 Percent as a Test:

Confusion Matrix	4 8 3 30
Accuracy	75.55
Sensitivity	57.14
Specificity	78.94
Precision	33.3333
F1Score	0.4210

Table-3
Taking 50 Percent of Data as Training and 50 Percent as a Test:

Confusion Matrix	11 14 5 36
Accuracy	71.2121
Sensitivity	68.75
Specificity	72.0
Precision	44.0
F1 Score	0.5365

Table-4
Taking 80 Percent of Data as Training and 20 Percent as a Test:

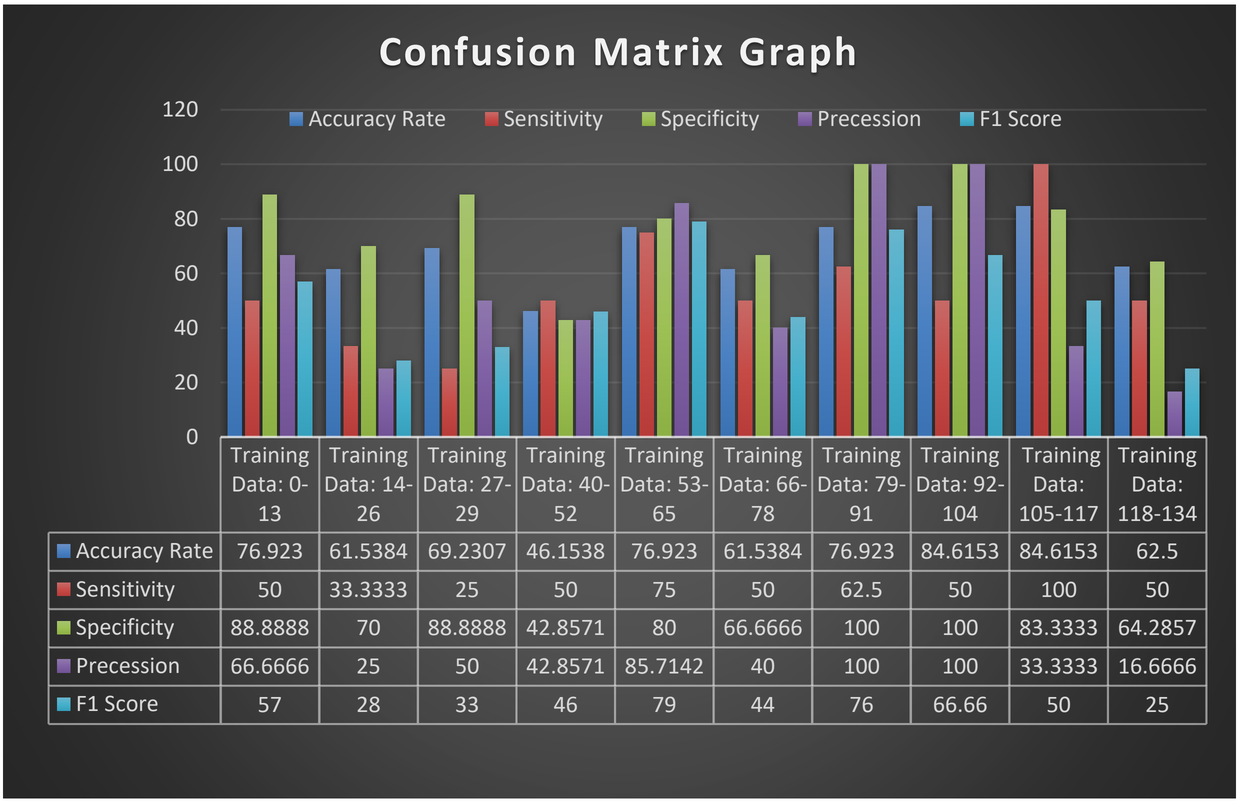
Confusion Matrix	2 7 0 17
Accuracy	73.0769
Sensitivity	100
Specificity	70.8333
Precision	22.22
F1 Score	0.3636

For 10-fold Cross-validation:

Table-5

Test Cases	Accuracy Rate	Sensitivity	Specificity	Precession	F1 Score
01	76.9230	50	88.8888	66.6666	0.5714
02	61.5384	33.3333	70.0	25.0	0.2857
03	69.2307	25.0	88.8888	50.0	0.3333
04	46.1538	50.0	42.8571	42.8571	0.4615
05	76.9230	75.0	80.0	85.7142	0.7999
06	61.5384	50.0	66.6666	40.0	0.4444
07	76.9230	62.5	100.0	100.0	0.7692
08	84.6153	50.0	100.0	100.0	0.6666
09	84.6153	100.0	83.3333	33.3333	0.5
10	62.5	50.0	64.2857	16.6666	0.25

CROSS-VALIDATION GRAPH :



	Training Data: 0-13	Training Data: 14-26	Training Data: 27-29	Training Data: 40-52	Training Data: 53-65	Training Data: 66-78	Training Data: 79-91	Training Data: 92-104	Training Data: 105-117	Training Data: 118-134
Accuracy Rate	76.923	61.5384	69.2307	46.1538	76.923	61.5384	76.923	84.6153	84.6153	62.5
Sensitivity	50	33.3333	25	50	75	50	62.5	50	100	50
Specificity	88.8888	70	88.8888	42.8571	80	66.6666	100	100	83.3333	64.2857
Precession	66.6666	25	50	42.8571	85.7142	40	100	100	33.3333	16.6666
F1 Score	57	28	33	46	79	44	76	66.66	50	25

7. Conclusions:

This paper uses multiple regressions (MLR) to predict High BP/Cholesterol. This Dataset has been collected from some districts in West Bengal, India based on the author made a relationship between the dependent variable and the independent variable after that author perform cross-validation for more accuracy. After checking the cross-validation, move to the Confusion matrix where the actual target values are compared with those predicted values by the machine learning model. Using these models, the author predicts the accuracy as well as sensitivity and specificity for 50 percent, 66 percent, and 80 percent sets of training data. In this project, the author defined some actual reasons for increasing High BP\Cholesterol. It will help any country's government to analyze their country's High BP/Cholesterol and they can also reduce the reasons behind increasing High BP/Cholesterol.

8. References:

Before doing that research paper some research is necessary so after analysis some paper some different concept is created and then a project can do so whatever papers are analyzed before doing this project those papers link are given in below

1) dozee-early-warning-score-dewsDozee | Hospital Care | Early Warning System, M. Knowledge Centre. Mar 10, 2021.

2) A. Sreeniwas Kumar[a,*] and Nakul Sinha[b](2020) Cardiovascular disease in India: A 360-degree overview,49-53

3) Ghani, I. M. M., & Ahmad, S. (2010). Stepwise multiple regression method to forecast fish landing. *Procedia-Social and Behavioral Sciences*, *8*, 549-554.Breiman,
L. (2001). Random forests. Machine learning, 45(1), 5-32.

4) Ben-Shakhar, G., Lieblich, I., & Bar-Hillel, M. (1982). An evaluation of polygraphers' judgments: A review from a decision-theoretic perspective. Journal of Applied Psychology, 67(6),701.

5) Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

6) Varshaa a; Vinitha V; Usha Nandhini D; Yogeshwaran R; Soundharya B M. "Artificial intelligence and its applications- A Review". *International Research Journal on Advanced Science Hub*, 1, 2, 2020, 1-4. doi: 10.47392/irjash.2019.11

7) Darmon, M., Vincent, F., Dellamonica, J., Schortgen, F., Gonzalez, F., Das, V., ... & Schlemmer, B. (2011). Diagnostic performance offractional.

8) .excretion of urea in the evaluation of critically ill patients with acute kidney injury: a multicenter cohort study. Critical care, 15(4),1-8.

9) Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal.

10) Samanta, Animesh, Akash Chowdhury, Dip Das, Arup Kumar Dey, and Mrs.
Sulekha Das. "Prediction through machine learning on the dependence of job prospects in the Afro-American community on proficiency in English."

11) Saha, Soumayadip, Joyitree Mondal, Arnam Ghosh, Sulekha Das, and Avijit
Kumar Chaudhuri. "Prediction on the Combine Effect of Population, Education, and
Unemployment on Criminal Activity Using Machine Learning."

12) Pal, Saikat Sundar, Soumyadeep Paul, Rajdeep Dey, Sulekha Das, and Avijit
Kumar Chaudhuri. "Determining the probability of poverty levels of the Indigenous
Americans and Black Americans in the US using Multiple Regression."

13) Ray, A., & Chaudhuri, A. K. (2021). Smart healthcare disease diagnosis and patient
management: Innovation, improvement and skill development. Machine Learning with Applications, 3, 100011.

14) Chaudhuri, Avijit Kumar, et al. "A multi-stage approach combining feature selection
with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis." Int J Intell Syst Appl 13.5 (2021): 46-63.

15) Naveenkumar S; Kirubhakaran R; Jeeva G; Shobana M; Sangeetha K. "Smart Health Prediction Using Machine Learning". International Research Journal on Advanced Science Hub, 3, Special Issue ICARD-2021 3S, 2021, 124-128. doi: 10.47392/irjash.2021.079

16) Chandrakala V; Surya Kumar M S R. "Intelligence slicing: A synthesized framework to integrate artificial intelligence into 5G networks". International Research Journal on Advanced Science Hub, 2, 8, 2020, 57-61. doi: 10.47392/irjash.2020.94