DEVELOPMENT OF MACHINE LEARNING MODEL TO PREDICT FUTURE POSSIBILITY OF HEART DISEASE.

Dr.M. Rajeswari[1]
Assistant Professor,
Department of B. Com (Business Analytics)
PSGR Krishnammal College for Women, Coimbatore, India
rajeshwarim@psgrkcw.ac.in

P. Kausika[2]
UG Scholar,
Department of B. Com (Business Analytics)
PSGR Krishnammal College for Women, Coimbatore, India. kausikapalanisamy2001@gmail.com

ABSTRACT

Heart disease is any problem in cardiovascular system. Coronary heart disease, Arrhythmia, Heart valve disease, Heart failure are some common types of heart disease, then it is said that the symptom of heart disease depends upon, type of heart disease of a patient. It is said that the leading cause of death in USA is because of heart disease. In this project the medical history of patients is taken as dataset. The cholesterol, diabetes, chest pain type, slope and many other attributes are taken into consideration for predicting heart disease in machine learning(ML). Machine Learning (ML) has proved to be vital in predicting heart disease. Linear regression and logistic regression algorithum with machine learning are used to predict the heart disease. The visualization is made for the dataset. Using logistic regression is used for predict further possibility of heart disease. The predicted output is displayed and the accuracy is measured.

Keywords – Machine learning model for Prediction of Heart Disease, machine learning (ML), Logistic regression, artificial intelligence(AI),

I. INTRODUCTION

Problem in cardiovascular system or blood vessels is heart disease. People above 60 age are more likely to get heart disease. There are four levels of chest pain type asymptomatic, atypical angina, non-angina pain, typical angina. Many people have died due to severe heart disease. Main symptoms of heart disease are Chest pain, chest tightness, chest pressure, chest discomfort, shortness of breath, pain, numbness, pain in the neck, jaw, throat, upper abdomen or back these are some usual symptoms of heart disease. There are more types of heart disease, but commonly heart disease refers to coronary heart disease. Machine learning(ML) is an artificial intelligence(AI) which allows software applications to become more accurate in predicting outcomes. It uses algorithms that use historical data as input to predict new output values. It could be applied in machine learning to determine if a person is likely to get heart disease. Linear regression is supervised machine learning, it finds the best fit linear line between independent and dependent variable. Logistic regression is used to calculate or predict the probability of a heart disease (yes/no) event occurring.

II. OBJECTIVE

This objective is framed in the goal as for development of machine learning model to predict the future possibility of heart disease. The risk factor for heart disease is having diabetes, cholesterol level above 250, slope test, hypertension. Heart disease can be identified through ECG, slope test, thallium, old peak, blood pressure level, blood sugar, prevalent stoke and many other datasets are collected for further prediction of heart disease. Changing in lifestyle and taking proper medicine reduces the risk of death due to heart disease. The cholesterol, diabetes, chest pain type, slope and many other attributes are taken into consideration for predicting heart disease in machine learning(ML). Machine Learning (ML) has proved to be vital in predicting heart disease. For understanding few charts are made. Linear regression and logistic regression algorithum with machine learning are used to predict the heart disease. After applying algorithum, the visualization is made for the dataset. Using logistic regression is used for predict further possibility of heart disease. Using this logistic regression, the prediction of heart disease is easier. The predicted output is displayed and the accuracy is measured.

III. RELATED WORK

The term “heart disease” refers to several types of heart conditions. The most common type of heart disease in the United States is coronary artery disease (CAD), which affects the blood flow to the heart [10]. High blood pressure, high blood cholesterol, and smoking are key risk factors for heart disease [11]. Heart disease in diabetic individuals appears earlier in life, affects women almost as often as men, and is more often fatal [2]. Adults with diabetes are more likely than those without diabetes to have heart disease risk factors, especially high blood pressure, low levels of HDL cholesterol, and high levels of triglycerides [4].

The incidence of heart disease is significantly increased among diabetic individuals in five occupational/population-based studies after adjustment for major heart disease risk factors [5]. The cardiovascular system consists of the heart and blood vessels. [6] Cigarette smoking, body fatness and relative bodyweight did not seem to explain population differences in incidence of the disorder, but there was a tendency for incidence to be related to the prevalence of hypertension, serum cholesterol values and saturated fatty acids in the diet. There were no statistically significant relations between habitual physical activity and the incidence of coronary heart disease. [1]

One of the most important advances is the availability of standardized criteria for the definition of diabetes, promulgated by the U.S. National Diabetes Data Group [3]. These findings have been translated into health promotion programs by the American Heart Association with emphasis on seven recommendations to decrease the risk of CVD: avoiding smoking, being physically active, eating healthy, and keeping normal blood pressure, body weight, glucose, and cholesterol levels.[7] Strokes can leave people with severe disabling sequelae like dysarthria or aphasia, dysphagia, focal or generalized muscle weakness or paresis that can be temporal or cause permanent physical disability.[8]

The most feared complication from CVD is death and, as explained above, despite multiple discoveries in the last decades CVD remains in the top leading causes of death all over the world owing to the alarming prevalence of CVD in the population. [9] Machine Learning is used to make exact decisions based on observations and predictions. Machine Learning examines the areas of algorithms that can make high-end predictions on data. [13]

The learning process in Machine Learning is classified into Training and Testing. If the model is to be built, the training data has to be utilized and this model will also be validated using testing data [15]. Logistic Regression (LR) is one of the most important statistical and datamining techniques employed by statisticians and researchers for the analysis and classification of binary and proportional response data sets [12] Logistic regression designs the best-fitting function with the help of the maximum likelihood method in order to maximize the probability of classifying the recognized data into the proper division [14].

IV. METHODOLOGY

A. PROPOSED SYSTEM

STEP 1: Imported the dataset from kaggle dataset, modified the dataset and saved in Excel.csv format.

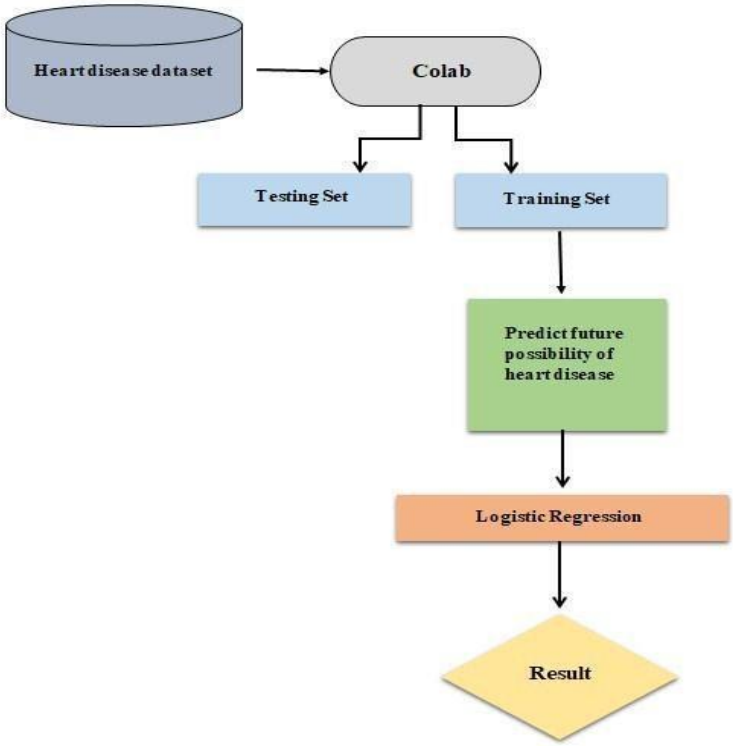
STEP 2: Used google colab for executing python coding.

STEP 3: Preprocessing is done by removing all unwanted data from dataset.

STEP 4: Then dataset is separated into training dataset and testing dataset.

STEP 5: Visualization are made in google colab for better understanding of dataset. **STEP 6:** Logistic algorithm is used to predict the outcome of dataset, accuracy is calculated.

V. WORK FLOW



A. DATA PREPROCESSING

- In data preprocessing first the data set is imported in google colab.
- Data processing is a process in which unwanted data are removed from the imported dataset.
- The data like hypertension, sysBP, diaBP, max heart rate, prevalentHyp, current smoker, cigs per day, mental health, physical health.
- The unwanted data/column is removed using code “drop” in colab. We are using logistic regression to predict heart disease dataset.

B. LOGISTIC REGRESSION

- Logistic regression is a model that helps to determine the probability or to predict the outcome. As a result, it helps in making a better decision by prediction of heart disease.
- A logistic approach it fits the best when the task in machine learning is based on two values, or a binary classification.
- In medicine, this analytics approach can be used to predict the disease or illness for a given population.

VI. RESULT

C. VISUALIZATION

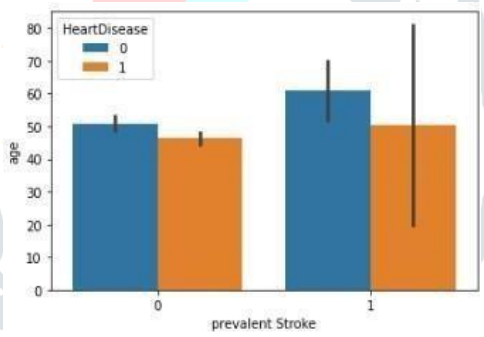


Fig 6.1
prevalent stroke with heart disease

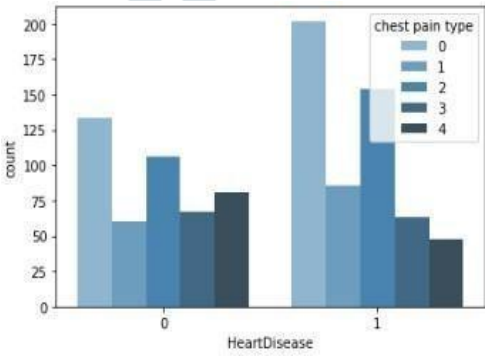


Fig 6.2
chest pain type and heart disease

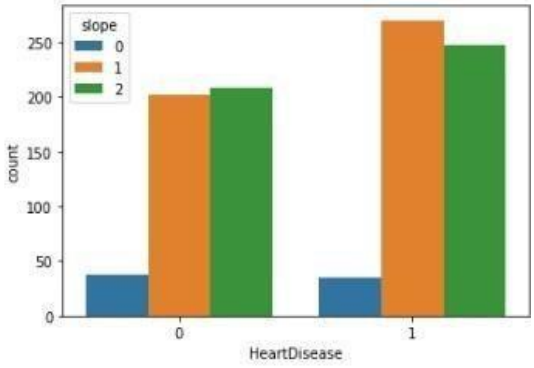


Fig 6.3
Slope and heart disease

- In fig 6.1 the common attributes are prevalent stroke, age and heart disease are compared x-axis represent prevalent stroke, y-axis represent age and in index 0 represent negative heart diseases, 1 represent positive of heart disease. From this chart it is understood that people who had prevalent stroke are likely to get heart disease.

- In fig 6.2 the attributes chosen are chest pain type, count and heart disease. The xaxis represents heart disease, y-axis represent count, index shows chest pain type. From this chart it is understood that people who have chest pain at level 0 are likely to get heart disease.

- In fig 6.3 the attributes chosen are slope, count, heart disease. The x-axis represents heart disease, y-axis represent count and index shows slope rate level. From this chart it is understood that people who had slope rate level at 1 are likely to get heart disease

D. PREDICTION OF HEART DISEASE BY MACHINE LEARNING

- One of the main reason for heart disease is high cholesterol, BMI, lack of exercise, high diabetes.
- Above age 50, 85% of the people are getting heart disease. After the testing the slope it could be able to easily predict the heart disease.

	predicted_value	KnowO/P
163	273	242
519	210	216
226	220	243
52	220	237
483	210	176
...	...	...
592	258	232
823	210	178
905	300	309
9	220	225
742	210	196

300 rows × 2 columns

```
#STATUS
#accuracy for HeartDisease
metrics.accuracy_score(test_Y,pr
```

```
1.0
```

Fig 6.4

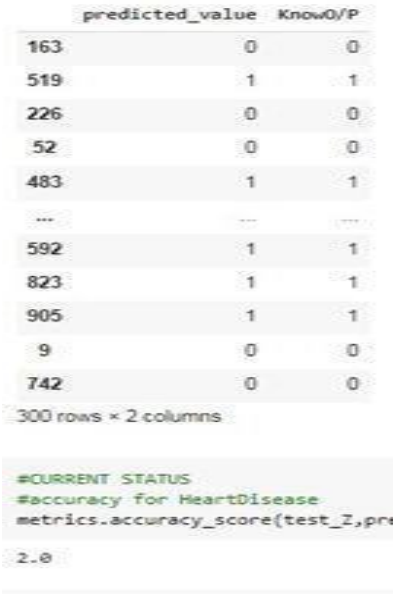
prediction with cholesterol

	predicted_value	KnowO/P
163	1	1
519	2	2
226	2	2
52	2	2
483	2	2
...	...	...
592	1	1
823	2	2
905	0	0
9	1	1
742	2	2

300 rows × 2 columns

Fig 6.5

prediction with slope



	predicted_value	KnowO/P
163	0	0
519	1	1
226	0	0
52	0	0
483	1	1
...	...	...
592	1	1
823	1	1
905	1	1
9	0	0
742	0	0

300 rows × 2 columns

```
#CURRENT STATUS
#accuracy for HeartDisease
metrics.accuracy_score(test_Z,pr
```

```
2.0
```

Fig 6.6

prediction with heart disease

- In fig 6.5, When the slope rate in 1, 95% of people are heart disease. The commonly affected heart disease is coronary heart disease.
- Where 80% of people get affected by this coronary heart disease.
- Using logistical regression, cholesterol and slope are taken as predicting variable. In the above picture, the known output (know O/P) the prediction is done with slope and cholesterol level and it is predicted and the output is shown.
- In fig 6.4, When the cholesterol level is above 200 the person is likely affected by heart disease. By analyzing all these dataset, it is predicted he/she maybe heart disease is positive in future.

VII. CONCLUSION AND FUTURE WORK

In this paper, Google colab notebook is used to analyze and predict the future possibilities of heart disease. First the dataset is imported from kaggle dataset and for executing coding and for visualization of charts are done in google colab. Logistic regression is mainly used for further prediction in medical purpose, so for prediction of heart disease status (positive/negative) logistic regression is used for prediction among people. In USA many people died due to heart disease. From this paper it is understood that people above age 60 are likely getting heart disease. People who have chest pain are getting heart disease and people who have slope rate at level 1 are getting heart disease. To avoid the death due to heart disease proper medicine should be taken with doctors' advice, should have healthy diet and regular exercise, and should change the lifestyle.

REFERENCES

1. ADA Consensus Panel: Role of cardiovascular risk factors in prevention and treatment of macrovascular disease in diabetes. Diabetes Care 12:573-79, 1993

2. Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, Chiuve SE, Cushman M, Delling FN, Deo R, de Ferranti SD, Ferguson JF, Fornage M, Gillespie C, Isasi CR, Jiménez MC, Jordan LC, Judd SE, Lackland D, Lichtman JH, Lisabeth L, Liu S, Longenecker CT, Lutsey PL, Mackey JS, Matchar DB, Matsushita K, Mussolino ME, Nasir K, O'Flaherty M, Palaniappan LP, Pandey A, Pandey DK, Reeves MJ, Ritchey MD, Rodriguez CJ, Roth GA, Rosamond WD, Sampson UKA, Satou GM, Shah SH, Spartano NL, Tirschwell DL, Tsao CW, Voeks JH, Willey JZ, Wilkins JT, Wu JH, Alger HM, Wong SS, Muntner P., American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association. Circulation. 2018 Mar 20;137(12):e67-e492.
3. Carvalho-Pinto BP, Faria CD. Health, function and disability in stroke patients in the community. Braz J Phys Ther. 2016 Jul-Aug;20(4):355-66.
4. Centers for Disease Control and Prevention, National Center for Health Statistics. About Multiple Cause of Death, 1999–2019. CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2019. Accessed February 1, 2021.
5. Farley A, McLafferty E, Hendry C. The cardiovascular system. 2012 Oct 31-Nov 6Nurs Stand. 27(9):35-9. [PubMed]
6. Hoffman, Julien IE, and Samuel Kaplan. "The incidence of congenital heart disease." Journal of the American college of cardiology 39.12 (2002): 1890-1900.
7. Lloyd-Jones DM, Hong Y, Labarthe D, Mozaffarian D, Appel LJ, Van Horn L, Greenlund K, Daniels S, Nichol G, Tomaselli GF, Arnett DK, Fonarow GC, Ho PM, Lauer MS, Masoudi FA, Robertson RM, Roger V, Schwamm LH, Sorlie P, Yancy CW, Rosamond WD., American Heart Association Strategic Planning Task Force and Statistics Committee. Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond. Circulation. 2010 Feb 02;121(4):586-613.
8. Manson JE, Colditz GA, Stampfer MJ, Willett WC, Krolewski AS, Rosner B, Arky RA, Speizer FE, Hennekens CH: A prospective study of maturity-onset diabetes mellitus and risk of coronary heart disease and stroke in women. Arch Intern Med 151:1141-47, 1991
9. Mark Club and George Michailidis, " Graph-Based SemiSupervised Learning", IEEE Transaction On Pattern Analysis And Machine Intelligence, Vol 30, NO.1, January 2008
10. National Diabetes Data Group: Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. Diabetes 28:1039-57, 1979
11. R. Berk. Statistical Learning from a Regression Perspective. Springer, 1st edition,2008.
12. Sunpreet Kaur, Sonalika Jindal, "A Survey on Machine Learning Algorithms", November 2016.
13. Vijay N. Kalbande, Dr. C.C.Handa, "Developing A Model To Predict Employability Of Engineering Students In Campus Placement For IT Sector", IJAREST Vol 2, Issue 6, June 2015
14. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics—2021 update: a report from the American Heart Associationexternal icon. Circulation. 2021;143:e254–e743.
15. Wingard, Deborah L., and Elizabeth Barrett-Connor. "Heart disease and diabetes." Diabetes in America 2.1 (1995): 429-448.