HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

(Logistic Regression)

Manasi Khedekar	**Shruti Birajdar**	**Prof. Deepali Shah**
Student, NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai	Student, NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai	Assistant Professor (MCA), NCRD's Sterling Institute of Management Studies, Nerul

ABSTRACT

This paper presents an overview of Heart Disease Prediction using Machine learning algorithms. In recent years, Heart disease is one of the most common diseases frequently found in India. In 2016, a total of 17.9 million people died due to chronic heart diseases.**[1]**

There, therefore, exists a dire need to mitigate these risks.

The current study implements Logistic algorithms for the prediction of heart disease based on anatomical attributes. The machine learning model will generate an equation with which individuals can assess themselves to understand if they might get diagnosed with chronic heart disease.In this logistic algorithm 14 variables are added .Using that variables only we can predict that the person is diagnosed with chronic heart disease or not.If any other variables will be include to calculate that it will be a fail. Because this algorithm works on that particular 14 variables only. The 14 variables which were included -age,sex,chest pain, blood pressure when resting,,Cholesterol , fasting blood sugar, Electrocardiographic results, maximum heart rate, pain induced due to exercise, ST depression induced by exercise, slope of the peak, thal, fluoroscopy, Dependent variable.The motive of this research is to save the amount which we are spending alot to the healthcare systems which are sometimes we cant afford. By checking the right things at the right time we can save money and save lives too. This paper states that by putting the right value in the researcher's regression we will get to know the correct result whether the person is diagnosed with heart disease or not.

***Keywords*:**

Heart Disease Prediction, Logistic Regression, Machine learning.

INTRODUCTION:

Heart disease is the most common disease in India. Earlier people were not aware of the food they are eating which is supposed to be oily, sweet, salty, sour which in future may lead to affecting our body system by consumption of more. Earlier if people are eating so much oily food then too they were working hard so the chances of getting digest the food is more were as now more people are just sitting in one place or one chair whole with eating the same food which may lead to the disease. Earlier people were not aware of this but now as they are educated and the health system is also improvised so they are getting to know about this in earlier phases. Then too some people are ignoring this and eating a lot of unhealthy and junk food which is causing the disease. In some cases, before the person gets to know about this the people lose their lives.

Heart disease refers to any condition affecting the heart. There are many types, some of which are preventable while in some cases the exact cause of CVD isn't clear, but there are lots of things that can increase your risk of getting it. The main risk factors for CVD are High Blood Pressure, High Cholesterol, Diabetes, Inactivity, and Obesity.0 Murray CJ and Lopez AD in their study 'Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study, Lancet. 24 May 1997, projected that the annual number of deaths due to cardiovascular diseases would rise from 2.26 million (1990) to 4.77 million (2020). **[2]**

REVIEW OF LITERATURE:

In a study undertaken by Gupta R et al., 'Epidemiology and causation of coronary heart disease

and stroke in India' in 2008, they projected that India would host more than half cases of heart diseases in the world in the coming 15 years. Estimated deaths due to cardiovascular diseases have been predicted and their prevalence in rural areas is ranged from 1.6% to 7.4% and 1% to 13.2% for urban areas. **[3]** A study by Rajiv et al., 'Prevalence of coronary heart disease and risk factors in an urban Indian population' in 2002 concluded that risks leading to heart diseases have become prevalent. **[5]**

A few of the reasons that contribute to heart disease are smoking habits, inactivity, hypertension, diabetes. Gupta R et al., in the study 'Younger age of escalation of cardiovascular risk factors in Asian Indian subjects.' in 2009 states how individuals with ages

ranging from 30-39 years are extremely vulnerable to cardiovascular diseases, and therefore there exists a need to mitigate this vulnerability.**[4]**

Frederic Commandeur et al., 2019 in their study were able to predict heart diseases better than the traditional clinical assessments. Another example of machine learning being used in the healthcare domain is by McKinney et al., (2020).**[6]** They have developed an ML algorithm that detects cancer tumors on mammograms. Likewise, this paper focuses on using basic anatomical factors to create a machine learning model that predicts if an individual is vulnerable to chronic heart disease. Multiple studies have incorporated the Naive Bayes algorithm and Decision Trees.

PROBLEM DEFINITION:

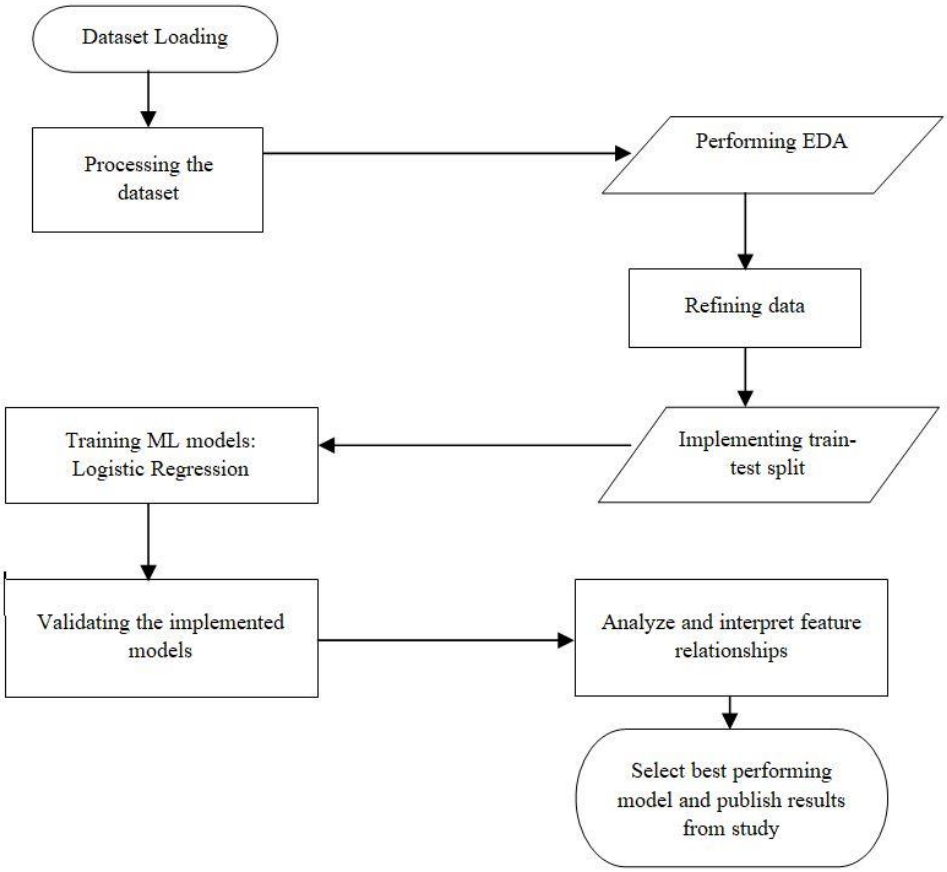
India is burdened with the population explosion. With the COVID 19 pandemic, it was evident that our healthcare domain needed restructuring with the best services and smart solutions to tackle the overwhelming surge of patients. Also, there exists a need to mitigate the over-increasing heart risks. A cure for heart disease is not a magic pill but an improvement in lifestyle. A healthy lifestyle would mitigate the risk of heart disease. And thus, this study becomes extremely crucial in achieving the said goal. Researchers are using the Heart disease dataset to curate a machine learning model that precisely predicts if an individual has a chance of getting diagnosed with chronic heart disease. The dataset comprises of the mentioned factors of the individuals; age, sex, chest pain, resting blood pressure, cholesterol level, Sugar level when fasting, heart rate when resting, etc. With a systemic model development process, we will be able to generate a machine learning model that classifies the individuals into the ones with a higher risk of chronic heart disease and the ones with minimal or no risk of chronic heart disease.

OBJECTIVE/SCOPE:

The dataset currently has 14 anatomical factors that are going to be used to create a machine learning model. However, when undergoing assessment for heart disease, there are multiple factors that need to be reviewed. Like high cholesterol levels, stress levels play a significant role in heart disease detection. Food habits constitute the risk of heart diseases. This study would generate an equation that individuals can use to assess themselves and understand where they stand in terms of heart disease. The 14 factors from the dataset are not empirical

and therefore the solution provided by the study would be one of the solutions for detecting heart disease and not the only solution.

RESEARCH METHODOLOGY:



INTRODUCTION:

The study consists of implementing a machine learning model to predict probable subjects that may get diagnosed with heart disease. The study will start with loading the data in a jupyter notebook, implementing suitable transformations to it, and then making it eligible for analysis. Once this is done, we will curate the machine learning models. Finally, Researchers choose the most efficient model that provides expected results.

Dataset Description:

The proposed research will work on the dataset repository, It has 303 anatomical records of patients across 14 variables. Following the variable description:-

1. **Age –** Age description of the patient.

2. **Sex –** Gender of the patient
3. **Chest pain -** 1: typical angina, 2: atypical angina, 3: non-anginal pain, and 4: asymptomatic.
4. **trestbps -** The blood pressure when resting
5. **chol –** Cholesterol level in mg/dl
6. **FBS -** Fasting blood sugar > 120 mg/dl; Yes or No.
7. **restecg -** Electrocardiographic results while resting (values 0,1,2)
8. **thalach -** maximum heart rate achieved
9. **exang –** If pain was induced due to exercise.
10. **old peak -** ST depression induced by exercise relative to rest
11. **slope -** the slope of the peak exercise ST segment
12. **ca -** number of major vessels (0-3) colored by fluoroscopy
13. **Thal-** 3= normal; 6 = fixed defect; 7 = reversible defect
14. **target –** Dependent variable

Data Pre-processing:

In this step, Researchers will transform the dataset suitably and make it appropriate for model development. They need to implement this step prior to model creation or even the univariate or bivariate analysis. The dataset processing includes the following steps:

Null value treatment – In this step, they check through the dataset if they are able to find any null values. When found, they deal with each variable individually and treat them by imputing Mean or Median in the case of numerical variables and Mode or category creation for Categorical variables.

Outlier treatment – There are two steps to conduct this step. In the first approach, Researchers check the statistics for the numerical variables. If the difference between the maximum value and 75% quartile value is extremely high, then they say that outliers exist. In the second approach, they plot a box plot for the variable. If there exist any points beyond the whiskers, it's evident that outliers exist.

Data type formatting – If variables in the dataset are incorrect or are in a different format, Researchers need to convert them to an appropriate format. Say, the sex variable was written

as F and M for female and Male respectively, then they have to transform the variable by imputing 0 for female and 1 for male or vice versa.

Dataset Splitting - In this step, the dataset is split into 70%-30% split, where 70% is for training and the rest 30% is for validating the results. If Researchers were to train the model on the entire data, it will learn all the values and thus lead to overfitting.

Model creation and Evaluation:

Now that data is transformed and ready for analysis, they go ahead with model creation. Multiple models will be created and one of the efficient ones will be finalized. Recursive feature elimination or RFE will also be used to understand the variables suitable for analysis.

Evaluation metrics:

F1 score is the harmonic mean of precision and recall. Researchers have used the F1 score as an evaluation metric. More focus is given to precision and recall as the study is based on predicting heart disease. Even the tiniest possibility of heart vulnerability shouldn't be missed and therefore F1 score will provide the best analysis of the model.

Data:

From dataset first 15 records are given below:

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1

ANALYSIS AND FINDINGS :

Deaths due to heart diseases have increased as per the WHO statistics and so does the need for tackling this ever-growing chronic disease. Machine learning now is doing great in the healthcare domain. Ziegler LE (2017) emphasizes the need to detect chronic diseases while they are in the early phase of diagnosis. Furthermore, he explains how delaying the diagnosis and treatment worsens the health condition of the patient. Kenneth Smith, James Kirby, and Anthony Kang (2018) stated how analysis by AI and Human intervention proved crucial than AI or Human intervention alone. With this study researchers were able to decipher that chest pain showed a greater impact on the heart disease target variable followed by the slope variable that recorded peak ST-segment rise. Emphasis is more on the variable that recorded the chest pain and therefore it is evident that the type of chest pain suffered will in turn be able to predict if the subject is vulnerable to heart disease. The optimum cutoff for the Sensitivity-Specificity trade-off came as 0.6. This is the threshold for the calculation of heart disease risk. The below mentioned equation was obtained with which the individual

TargetVariable=1.511+ cp_2(2.66)+slope(1.984)+cp_3(1.792)-sex(1.22)-thal_3(1.942)-ca_1(2.607)-ca_2(2.842)-oldpeak(3.412)-ca_3(3.823)

LIMITATIONS AND SCOPE :

The study is about predicting heart disease based on anatomical factors. Variables related to stress, diabetic condition, or previous history of stroke are not covered. Individuals could use the equation from this study to assess themselves. Additionally, clinicians and medical centers could also use this equation as one of the assessments to detect heart-related diseases. Previous researches have not used logistic regression for the analysis of the dataset. The researchers will therefore try to implement the said technique and refine the model to improve the research in this domain.

CONCLUSION :

Based on the anatomical factors, researchers were able to predict the target variable, i.e. whether an individual is vulnerable to a heart-related disease with an outstanding F1 score of 88%. Healthcare is one of the largest domains and therefore for someone with a chronic disease condition, medical treatment and assessments would cost a fortune. The study, therefore, comes in as a savior by giving the individual an equation that he can use to self-

assess and avoid undergoing any further expensive assessments. Since there exists no cure for chronic disease, anyone with even the slightest vulnerability should be treated and given utmost attention. This is the reason the study has emphasized the F1 score metric and not the Accuracy. Analysis and prediction of heart disease through logistic regression was successful and could be refined further by adding multiple variables and improving dataset size.

REFERENCES :

1.Cardiovascular diseases-(cvds) https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2.Murray CJ, Lopez AD. Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study. Lancet. 1997 May 24;349(9064):1498-504. DOI: 10.1016/S0140-6736(96)07492-2. PMID: 9167458.

3.Gupta R, Joshi P, Mohan V, Reddy KS, Yusuf S. Epidemiology and causation of coronary heart disease and stroke in India. Heart. 2008 Jan;94(1):16-26. DOI: 10.1136/hrt.2007.132951. PMID: 18083949.

4.Gupta, R., Misra, A., Vikram, N. K., Kondal, D., Gupta, S. S., Agrawal, A., & Pandey, R. M. (2009). Younger age of escalation of cardiovascular risk factors in Asian Indian subjects. BMC cardiovascular disorders, 9, 28. https://doi.org/10.1186/1471-2261-9-28

5.Rajeev Gupta, V P Gupta, MukeshSarna, Smita Bhatnagar, JyotiThanvi, Vibha Sharma, A K Singh, J B Gupta, Vijay Kaul., (2002) Prevalence of coronary heart disease and risk factors in an urban Indian population: Jaipur Heart Watch-2. Available at - https://pubmed.ncbi.nlm.nih.gov/11999090/

6.Frederic Commandeur, Piotr J Slomka, Markus Goeller, Xi Chen, Sebastien Cadet, AryabodRazipour, Priscilla McElhinney, Heidi Gransar, Stephanie Cantu, Robert J H Miller, Alan Rozanski, Stephan Achenbach, Balaji K Tamarappoo, Daniel S Berman, DaminiDey, Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study, Cardiovascular Research, Volume 116, Issue 14, 1 December 2020, Pages 2216–2225, https://doi.org/10.1093/cvr/cvz321