# Heart Disease Prediction Model Using Machine Learning

Lovekush, Himanshu Arya, Piyush Saini, Mr. Devendra Kumar
*School of Computer Application and Technology, Galgotias University, Greater Noida, Uttar Pradesh*

***Abstract:*** **Heart failure is one of the major diseases worldwide, accounting for nearly one-third of all deaths globally. The growing burden of cardiovascular diseases particularly coronary artery disease, presents a significant challenge to human health life, especially in low and poor countries. This paper explores the major risk factors contributing to heart disease like cholesterol Blood pressure smoking etc. The Logistic Algorithm helps us in prediction, according to provided data which is taken by different countries' Research Labs like AHA, WHO, etc. In the future, we will look towards different algorithms like decision tree, and random forest to improve our project and research paper. But still, we have achieved an 82% accuracy rate in prediction. Moreover, this research paper will go in-depth and give the answer of why how, and what, why heath disease problems occur, how we can defend ourselves from it as well as what are mistakes people are making in their life which is the cause of heart disease. last but not least we have taken help from different researchers like AHA and WHO as well as we read some more projects and reaches which are relevant to our Research paper.**

***Keywords*** **– Logistic Regression, Kaggle data set (BP, Debates, smoking, alcohol, Chest pain, etc)**

## INTRODUCTION

Heart disease in particular is one of the biggest health issues that the world's population is currently dealing with. Heart disease still claims millions of lives annually, despite advances in medical research, placing a tremendous burden on global healthcare systems. Reducing the burden of this disease still depends on early diagnosis and intervention, but the progressive and subtle nature of its important facts like high blood pressure, smoking, cholesterol, and age—often makes it difficult to identify it in time. The application of machine learning techniques to aid in the early identification and prediction of heart disease has garnered increasing attention in recent years. These predictive models can offer vital insights into a person's risk of getting heart disease by examining patient data and finding patterns. This allows for more individualized and efficient preventative care. In particular, logistic regression has shown to be a dependable and comprehensible approach to solving binary classification problems such as the prediction of heart disease using logistic regression.

## RESEARCH PAPER PROBLEM STATEMENT

At present Heart Attack is of the bigger problem even with substantial improvements in medical care and public health campaigns. Improving patient outcomes and lowering death rates from heart disease requires early detection and prevention. Using machine learning techniques presents a potential to increase the efficiency and accuracy of cardiac disease prediction given the growing availability of patient health data. Still, there are difficulties in creating predictive models that are useful for application in actual clinical settings and that are not only accurate but also comprehensible and interpretable. In order to forecast the risk of heart disease, a strong and trustworthy machine learning-based model that can evaluate important risk factors including blood pressure, cholesterol, smoking status, and age is needed. This research attempts to fill that gap. The study's focus on logistic regression aims to produce a model that is practical and simple to use, giving medical professionals a useful tool for the early detection and prevention of heart disease.

## LITERATURE REVIEW

heart pumps about 5 to 6 Liters of blood per minute at rest. This can amount to approximately 7,200 Liters

2018 AHA

Guideline on the Management of Blood Cholesterol:

Fibers and omega-3 fatty acids are frequently advised, particularly in situations when there is a high risk of diabetes. In order to lower cholesterol levels, the guidelines also stress the significance of controlling other lifestyle factors like food, exercise, and weight loss. At all ages, a healthy lifestyle lowers the risk of cardiovascular disease. Healthy living is

the principle of ASCVD risk reduction in younger people, and it can slow the development of risk factors. An evaluation of lifetime risk for young individuals between the ages of 20 and 39 helps to focus on intensive lifestyle efforts and enhances the clinician-patient risk discussion.

2022 AHA

RESEARCH for the Management of Heart Failure

1. Prevention and Early Detection :

at-risk" and "pre-HF are two stages of heart Failure A and B for preventing, lifestyle changes and managing risk factors like hypertension and diabetes is recommended for individuals at these stages.

2. Medication Updates:

It is now advised that symptomatic individuals with heart failure with reduced ejection fraction use SGLT2 inhibitors or SGLT2i.

2019 ACC and AHA Report of Solution on Main Prevention of CVD Problem

1. Healthy Lifestyle
2. Managing Risk Factors
3. Aspirin Use
4. Cholesterol and Blood Pressure
5. Smoking Cessation

The major factor of Heart Disease is cardiovascular CVD which is caused by several things. According to WHO World Health Organization CVD problem generated by -

- Smoking
- Cholesterol level
- Blood pressor
- Diabetes
- Depression
- Anxiety
- Sleep
- Alcohol

Paper 1: Major Research View to Know about CVD Problem-

1. Cardiovascular diseases are the leading cause of death globally, accounting for 17.9 million deaths each year.
2. High blood pressure is the most significant risk factor for CVDs and affects nearly 1.4 billion people worldwide.
3. Lack of Physical Activity contributes to approximately 3.2 million deaths annually, making it a significant public health concern.
4. Smoking cessation reduces the risk of CVD by nearly half within one year of quitting.
5. Diabetes mellitus doubles the risk of developing CVD complications as compared to Non Diabetic humans.
6. A healthy Diet rich in vitamins and Protin can reduce the significant CVD problem

Paper 2: Cardiovascular disease its risk factors among older adults in six low- and middle-budget countries.

1. The prevalence of angina symptoms ranged from 5.7% in India to 14.7% in Ghana, showing significant regional variability.
2. Older adults in low-income countries often have limited access to healthcare services for CVD management.
3. Hypertension was present in over 40% of study participants, highlighting a major unaddressed health burden.
4. Physical inactivity and high body mass index (BMI) were common among participants, contributing to elevated CVD risk.
5. Men were less likely to report symptoms of angina but had higher rates of hypertension compared to women.
6. Tobacco use remains a pervasive risk factor, especially among men, exacerbating cardiovascular risks in these populations.

Paper 3: The main study of the Researchers on humans living with cardiovascular disease and their View on a digital health platform for self-management

1. Patients with CVD show the need for a user-friendly digital platform that provides clear and actionable health insights.
2. Regular feedback from healthcare providers through digital platforms was identified as a critical feature for improving adherence.
3. Many participants emphasized the importance of personalized health data, such as tailored dietary and activity recommendations.
4. The study highlighted significant concerns regarding data security and privacy in digital health platforms.
5. Digital platforms can play a vital role in bridging gaps in healthcare access, particularly for remote or underserved populations.

6. The integration of gamification elements was suggested as a strategy to enhance user engagement in self-management practices.

## OBJECTIVES

Predictive Power: Using logistic regression to its fullest potential in heart disease prediction is its main objective. Healthcare providers can determine a patient's risk of getting heart disease by examining a variety of patient data, including age, cholesterol, and lifestyle factors. This prediction model facilitates focused healthcare intervention to reduce risk and is an essential tool for early intervention.

Risk Factor Analysis:-To better understand risk factors use logistic regression as a magnifying glass. It measures how each factor affects the course of cardiac disease. For example, if the model indicates a substantial positive correlation between higher cholesterol and patient education and care.

Exploring Data

Note : If these values go above the Given ratio then there are significant chances of heart disease. According to research in the below table, how many people are affected by it .

THIS IS THE RATIO OF 2024. THE HUMANS WHO ARE DIED CAUSE OF HEART DISEASE.

| CAUSES | INDIA | CHINA | AMERICA | RUSSIA | JAPAN |
|---|---|---|---|---|---|
| SMOKING | 1.35M | 1.3M | 11K | 25K | 4K |
| CHOLESTROL | 4.4M | 1.1M | 120K | 250K | 500K |
| BLOOD PRESSOR | 2.6M | 2.5M | 130K | 300K | 610K |
| DIABETETES | 1.2M | 1.4M | 250K | 220K | 100K |
| SLEEP | 1.1M | 1.2M | 150K | 200K | 700K |

This data specified by the World Health Organization (WHO):

Cholesterol Levels:

- Total Cholesterol: A concentration of less than 200 mg/dL is considered optimal for cardiovascular health.
- Low-density lipoprotein: Levels below 100 mg/dL are recommended to decrease risk.
- High-Density Lipoprotein: A concentration of 60 mg/dL or higher is associated with protective cardiovascular effects.
- Triglycerides: Levels under 150 mg/dL are indicative of a healthy profile.

Blood Pressure Categories:

- Normal: Blood pressure below 120 mm Hg and diastolic pressure below 80 mm Hg.
- Elevated: Blood pressure around 129 mm Hg with diastolic pressure below 80 mm Hg.
- Stage 1: Systolic pressure between 130-139 mm Hg or diastolic pressure between 80-89 mm Hg.
- Stage 2: Blood pressure of 140 mm Hg or higher, or diastolic pressure of 90 mm Hg or higher.

Blood Sugar Levels:

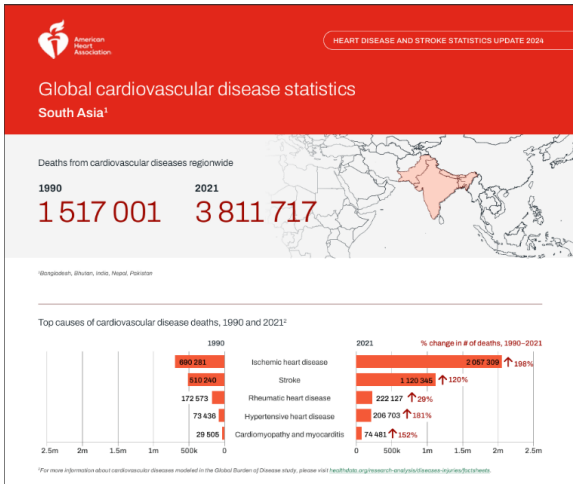
- Normal: Fasting blood glucose concentration below 140 mg/dL.
- Prediabetes: Blood glucose levels ranging from 140-199 mg/dL.
- Diabetes: Blood glucose levels of 200 mg/dL or higher, indicating diabetes mellitus.

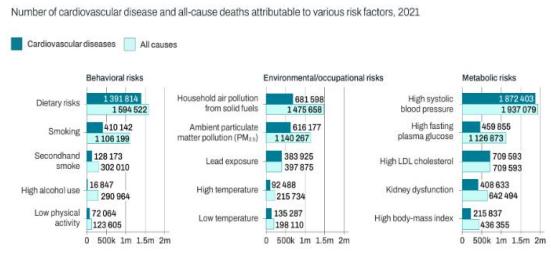
Sleep Duration:

- Adults: 6-8 hours per night.
- Older Adults (65+): 7-8 hours per night.

Reference This research has been certified by the WHO.

## STATISTICS




## PROPOSED SYSTEM

Introduction to the System:-The system aims to predict that a person has heart disease based on key health factors such as cholesterol levels, blood pressure, age, and smoking status. The logistic regression model will be used to classify the risk into two categories: high risk and low risk.

Data Collection

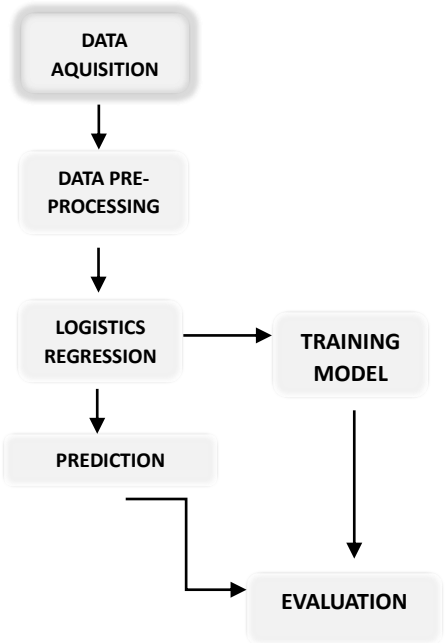
- Source of Data: The system will use real-world datasets Kaggle that include health records of patients, including age, cholesterol levels, blood pressure, smoking history, and other related factors.

| 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 |
| 50 | 0 | 2 | 120 | 219 | 0 | 1 | 158 | 0 | 1.6 | 1 | 0 | 2 | 1 |
| 58 | 0 | 2 | 120 | 340 | 0 | 1 | 172 | 0 | 0 | 2 | 0 | 2 | 1 |
| 66 | 0 | 3 | 150 | 226 | 0 | 1 | 114 | 0 | 2.6 | 0 | 0 | 2 | 1 |
| 43 | 1 | 0 | 150 | 247 | 0 | 1 | 171 | 0 | 1.5 | 2 | 0 | 2 | 1 |
| 69 | 0 | 3 | 140 | 239 | 0 | 1 | 151 | 0 | 1.8 | 2 | 2 | 2 | 1 |
| 59 | 1 | 0 | 135 | 234 | 0 | 1 | 161 | 0 | 0.5 | 1 | 0 | 3 | 1 |
| 44 | 1 | 2 | 130 | 233 | 0 | 1 | 179 | 1 | 0.4 | 2 | 0 | 2 | 1 |
| 42 | 1 | 0 | 140 | 226 | 0 | 1 | 178 | 0 | 0 | 2 | 0 | 2 | 1 |
| 61 | 1 | 2 | 150 | 243 | 1 | 1 | 137 | 1 | 1 | 1 | 0 | 2 | 1 |

Data Preprocessing

- Data Cleaning: Handle missing values and remove duplicates. For example, if the cholesterol or blood pressure data is missing, either remove those rows or use imputation techniques to fill them. Python libraries help to manipulate it.
- Normalization: Numerical Normalized data features like cholesterol bold pressure will help to improve the model in performance.

DATA FLOW DIAGRAM




METHODOLOGY

Data Collection

- Source of Data: Data was taken from Kaggle and included patient details such as age cholesterol levels, blood pressure, and lifestyle habits.

Data Preprocessing

- Data Cleaning: Handle missing or incomplete data through techniques such as removing rows with missing values or imputing them using statistical methods like the mean or median.
- Normalization: Scale numerical data (cholesterol, blood pressure, etc.) to bring all features to the same scale, which helps logistic regression perform better.

Exploratory Data Analysis (EDA)

- Visualize Data: Use graphs (histograms, scatter plots, box plots) to visualize the relationships between features and the outcome.
- Correlation Analysis: Check correlations between independent variables (age, cholesterol) and the dependent variable. This helps in identifying important features.
- Statistical Summary: Generate summary statistics mean, and median to understand the difference of the data.

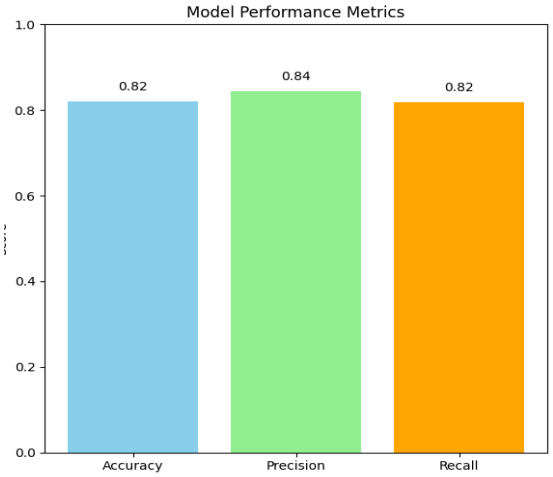
Model Development

- Logistic Regression: Logistic regression was chosen for its prediction effectively. This model was trained on 80% of the data and validated on the remaining 20%

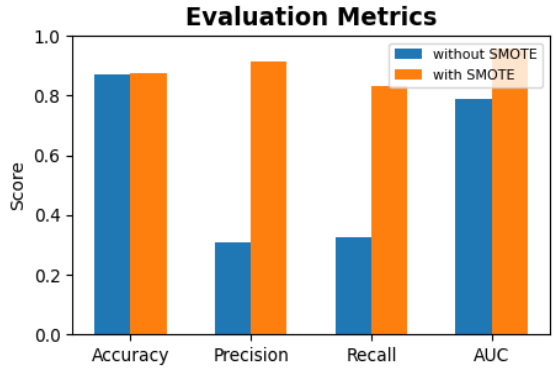
RESULT

The logistic regression model achieved an accuracy of 82% with a precision of 84% and recall of 82%. Some features like cholesterol and blood pressure were the most important predictors of heart disease risk.

1. Model Accuracy




Result: MR Auradee project CVD prediction. The logistic regression model achieved the same accuracy but our model did the better performance in precision and recall.




2. Precision and Recall

- Precision: Measures how many of the predicted positive cases (i.e., patients at high risk of heart disease) were correct.

Example: If the model predicts 50 patients to have heart disease, and 45 of them actually have it, the precision is 90%. According to our data set logistic regression has predicted 84% precision.

3. Practical Impact

- Risk Stratification: The system can help stratify patients into risk categories, allowing doctors to prioritize high-risk individuals for further tests or treatment.
- Patient Awareness: monitoring and prediction, patients can be more aware of their health status and take preventive actions.

## CONCLUSION

In the Present Time 2024 Heart disease has become a more significant problem and more increasing among people including our country. So Predicting the disease is very important to decrease the ratio of the risk of death. This study depicts, achieving an accuracy of 82% by analyzing patient data. This prediction is an area that is widely researched Our paper is part of the research on the detection and prediction of heart disease it is based on the application of a machine learning algorithm using logistic regression. Moreover, our Mindset is to decrease the rate of heart failure and its causes. Our research shows that if humans want to prevent themselves then they need to maintain our given ratio of body organs and chemicals. Apart from that our project can help humans from heart disease problems if a patient provides effective information about himself, on the basis of the provided parameter, we will predict whether the patient is in the red zone or green zone. in the last highlighting the urgent need for preventive measures and healthier lifestyle choices.

## REFERENCES

[1] World Health Organization (WHO). (2021, June 11). Cardiovascular diseases (CVDs). World Health Organization. Retrieved from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] American Heart Association (AHA). (2023). Heart disease and stroke statistics – 2023 update. American Heart Association. Retrieved from https://www.heart.org/en/health-topics/consumerhealthcare

[3] Kaggle. Find open datasets and machine learning projects. Retrieved from https://www.kaggle.com

[4] Samineni, P. Enhancing heart disease prediction using machine learning. Research contribution. Edward, N. Cardiovascular disease trends. Research contribution. Lloyd-Jones, D. M. Expert insights on cardiovascular disease at AHA. Research contribution.
McGowan, J. A. Cardiovascular disease prevention strategies across populations. Research contribution.

[5] Sharman, J., Tan, L., & Deane, F. (2023). Ten things to know about ten cardiovascular disease risk factors. *Journal of Cardiovascular Research*. Retrieved from

https://pmc.ncbi.nlm.nih.gov/articles/PMC9061634/

[6] Biritwum, R. B., & Minicuci, N. (2023). Cardiovascular disease and associated risk factors among older adults in six low- and middle-income countries. *BMC Public Health*. Retrieved from https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-018-5653-9

[7] Watkins, C. L., & Hughes, T. A. (2023). Perspectives on a digital health platform for cardiovascular disease self-management. *Journal of Digital Health Innovations*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC9628687/

[8] Auradee. (n.d.). *Heart Disease Prediction using Logistic Regression*. Kaggle. Retrieved December 22, 2024, from https://www.kaggle.com/code/auradee/heart-disease-prediction-using-logistic-regression