HEART DISEASE ANALYSIS REPORT

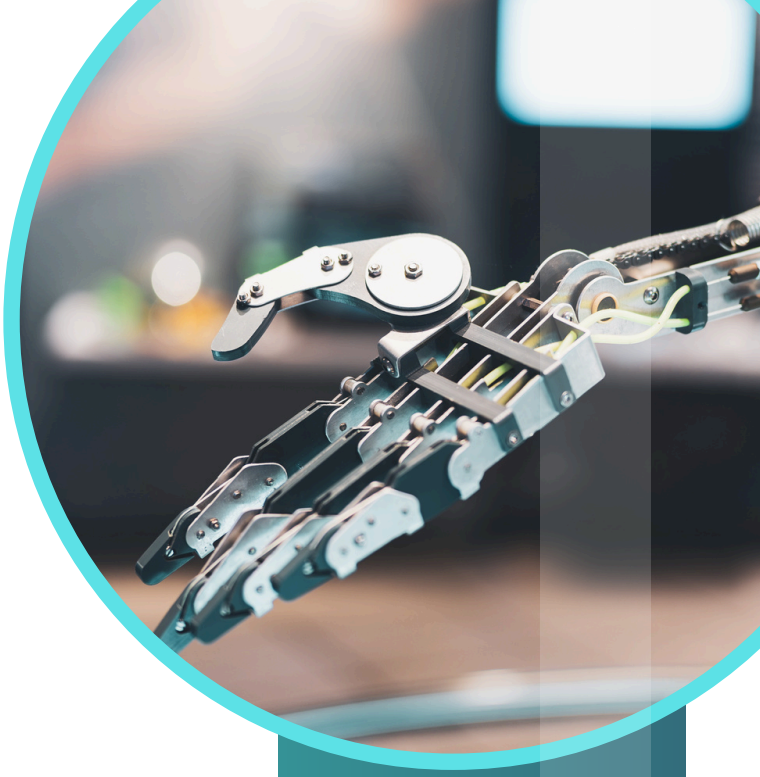
Machine Learning

2024

Prepared By:

Team Epsilon

Table of Content



Introduction

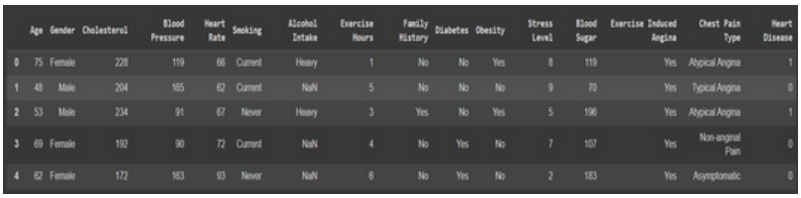
Despite significant advancements in healthcare, heart disease continues to be a leading cause of mortality worldwide. The ability to accurately predict heart disease is crucial for early detection and timely intervention, which can significantly improve patient outcomes and save lives. In this project, we aim to develop a robust predictive model to identify individuals at high risk of developing heart disease based on key demographic and health-related factors.

Our approach leverages a comprehensive dataset encompassing various attributes such as age, gender, cholesterol levels, blood pressure, smoking habits, alcohol intake, exercise patterns, family history of heart disease, diabetes status, obesity, stress levels, blood sugar levels, exercise-induced angina, and chest pain type. By utilizing advanced machine learning techniques, we strive to build a model that can effectively discern patterns and correlations within these variables to predict the likelihood of heart disease.

The goal of this project is not only to enhance our understanding of the risk factors associated with heart disease but also to contribute to the development of practical tools for healthcare providers. By identifying high-risk individuals early, we can facilitate preventive measures and personalized treatment plans, ultimately reducing the burden of heart disease on individuals and healthcare systems globally.

Data Acquisition

TThe dataset utilized for identifying heart disease in our tests was sourced from the Kaggle platform and can be freely accessed at kaggle/input/Heart-disease/new_model.csv. It comprises 1000 instances with 16 features, carefully selected to encompass a wide range of variables that are indicative of heart disease. The variables include Age, Gender, Cholesterol, Blood Pressure, Heart Rate, Smoking, Alcohol Intake, Exercise Hours, Family History, Diabetes, Obesity, Stress Level, Blood Sugar, Exercise-Induced Angina, Chest Pain Type, and the target variable, Heart Disease.



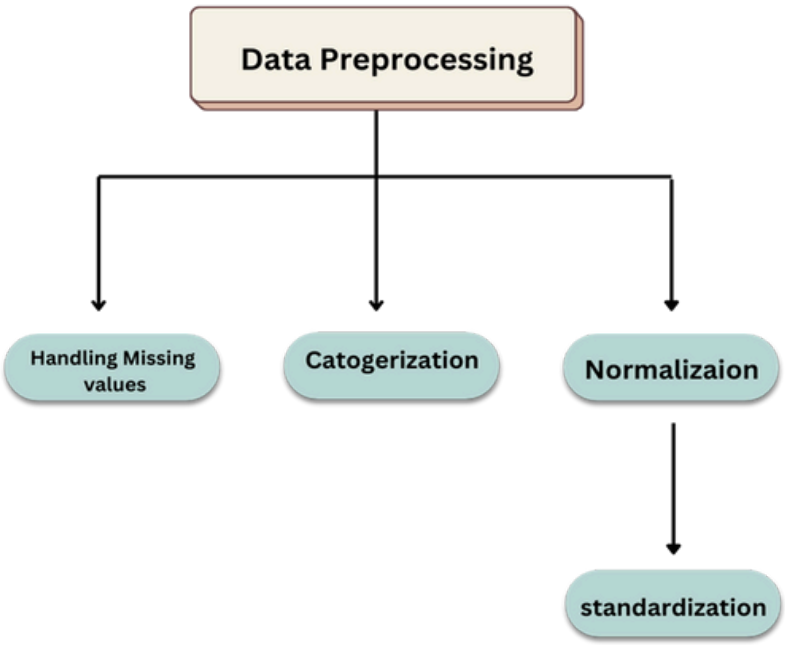
	Age	Gender	Cholesterol	Blood Pressure	Heart Rate	Smoking	Alcohol Intake	Exercise Hours	Family History	Diabetes	Obesity	Stress Level	Blood Sugar	Exercise Induced Angina	Chest Pain Type	Heart Disease
0	75	Female	228	119	66	Current	Heavy	1	No	No	Yes	8	119	Yes	Atypical Angina	1
1	48	Male	204	165	62	Current	NaN	5	No	No	No	9	70	Yes	Typical Angina	0
2	53	Male	234	91	67	Never	Heavy	3	Yes	No	Yes	5	196	Yes	Atypical Angina	1
3	69	Female	192	90	72	Current	NaN	4	No	Yes	No	7	107	Yes	Non-anginal Pain	0
4	62	Female	172	163	93	Never	NaN	6	No	Yes	No	2	183	Yes	Asymptomatic	0

These variables can be categorized into numerical and categorical data. Numerical variables in the dataset include Age, Cholesterol, Blood Pressure, Heart Rate, Exercise Hours, Blood Sugar, and Stress Level. Categorical variables include Gender, Smoking, Alcohol Intake, Family History, Diabetes, Obesity, Exercise-Induced Angina, Chest Pain Type, and Heart Disease. By leveraging this comprehensive dataset, our goal is to uncover patterns and correlations that will enable us to predict the likelihood of heart disease, thereby aiding in the early identification and management of high-risk individuals.

SR NO	VARIABLES
1	AGE
2	GENDER
3	CHOLESTEROL
4	BLOOD PRESSURE
5	HEART RATE
6	SMOKING
7	ALCOHOL INTAKE
8	EXERCISE HOURS
9	FAMILY HISTORY
10	DIABETES
11	OBESITY
12	STRESS LEVEL
13	BLOOD SUGAR
14	EXERCISE INDUCED ANGINA
15	CHEST PAIN

05

Data PREPROCESSING

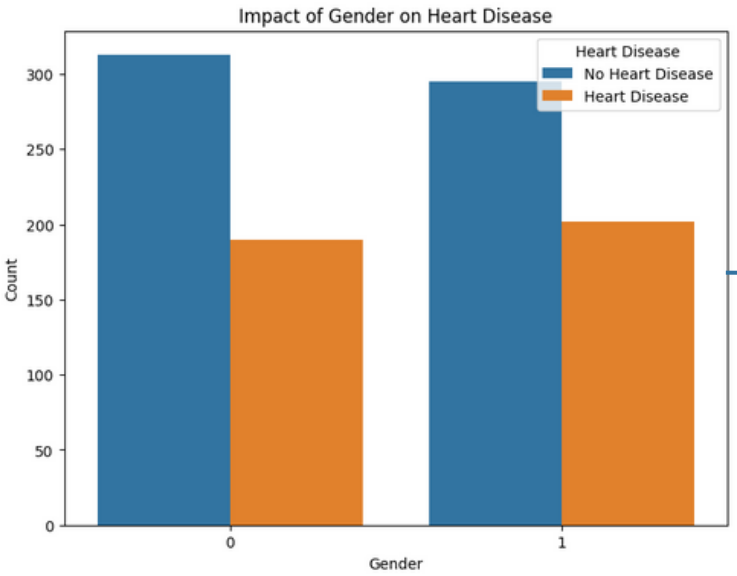


1.Handling Missing Values:
Handling missing values involves imputing them with estimates or removing incomplete rows/columns to ensure a complete dataset for model training.

2.Handling Imbalance Data:
By balancing the dataset, you ensure that your model does not become biased towards the majority class and performs well across all classes. This leads to more reliable and fair predictions, especially in critical applications where the minority class is of particular interest.

3.Standardizing and Normalizing Data:
Numerical features were standardized to have a mean of zero and a standard deviation of one. This process ensured that the data was on a similar scale, which is particularly important for algorithms that rely on distance measurements. Alternatively, normalization was applied to scale the numerical data within a specific range, typically between 0 and 1. This helped in speeding up the convergence of gradient-based algorithms

DATA ANALYSIS

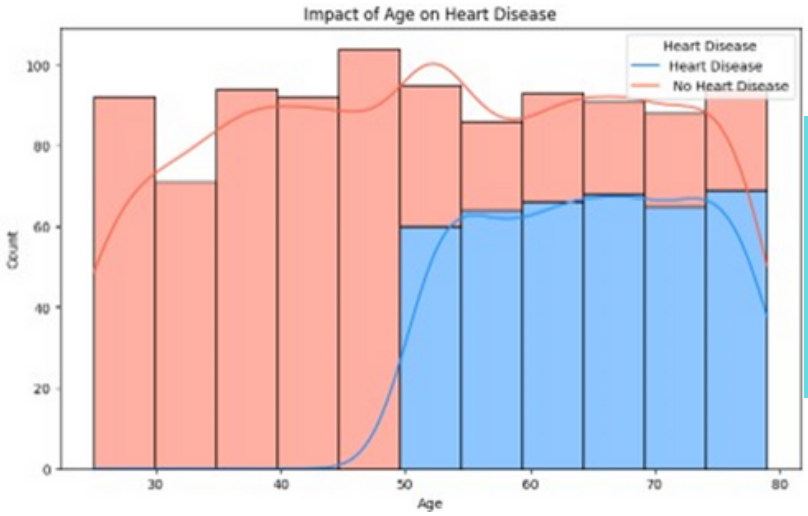


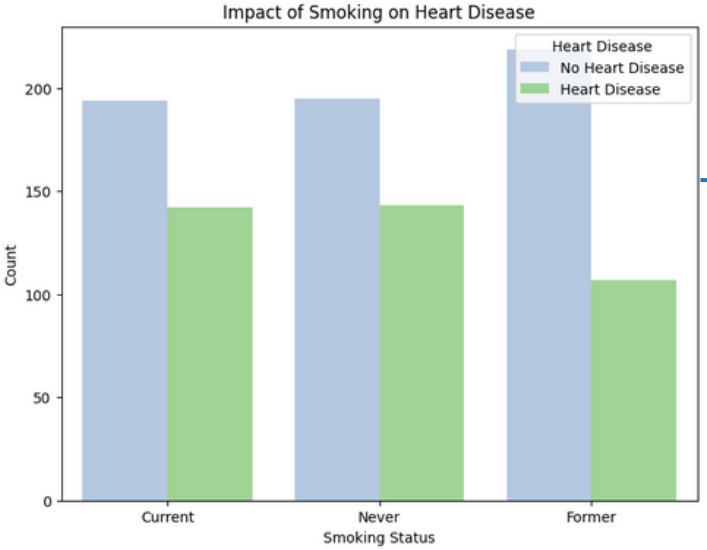
IMPACT OF GENDER ON HEART DISEASE?

The count of Female not suffering from Heart disease is greater than that of male. The count of Male suffering from Heart disease is greater than that of female.

IMPACT OF AGE ON HEART DISEASE?

People older than 50 are at greater risk of Heart Disease Age less than 50 have less count of people with Heart Disease

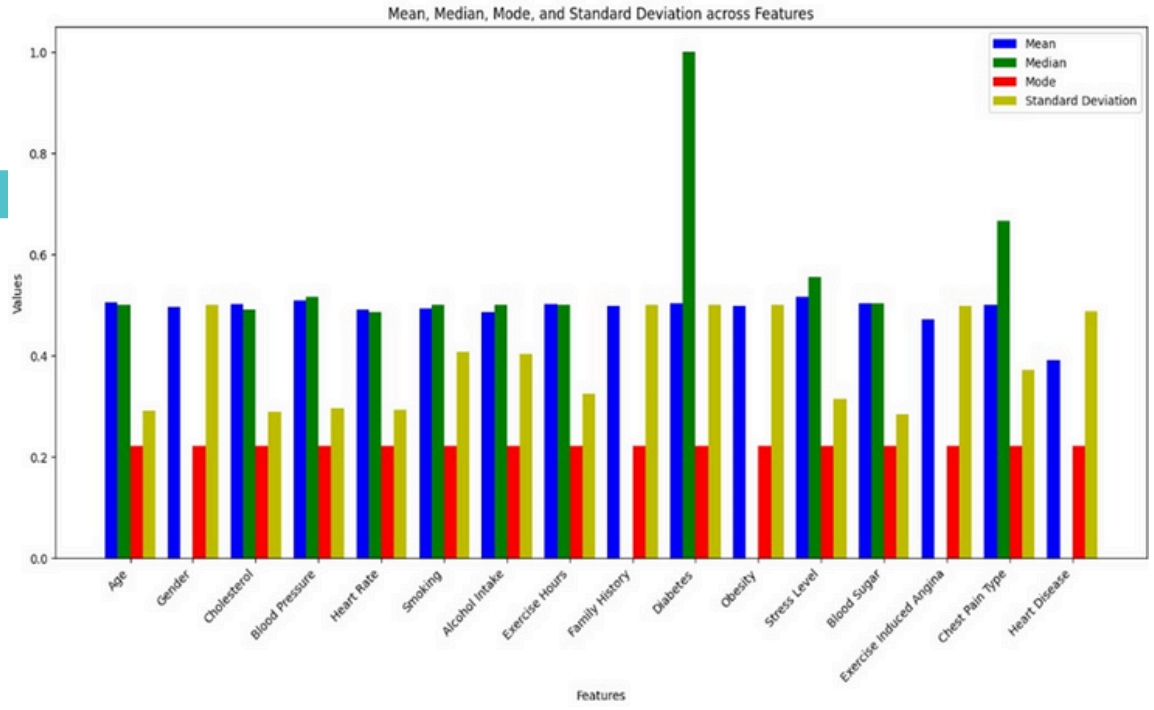




IMPACT OF SMOKING ON HEART DISEASE?

Current Smokers have highest count of people with heart disease
Former smokers have highest count of people with no heart disease

STATISTICAL ANALYSIS

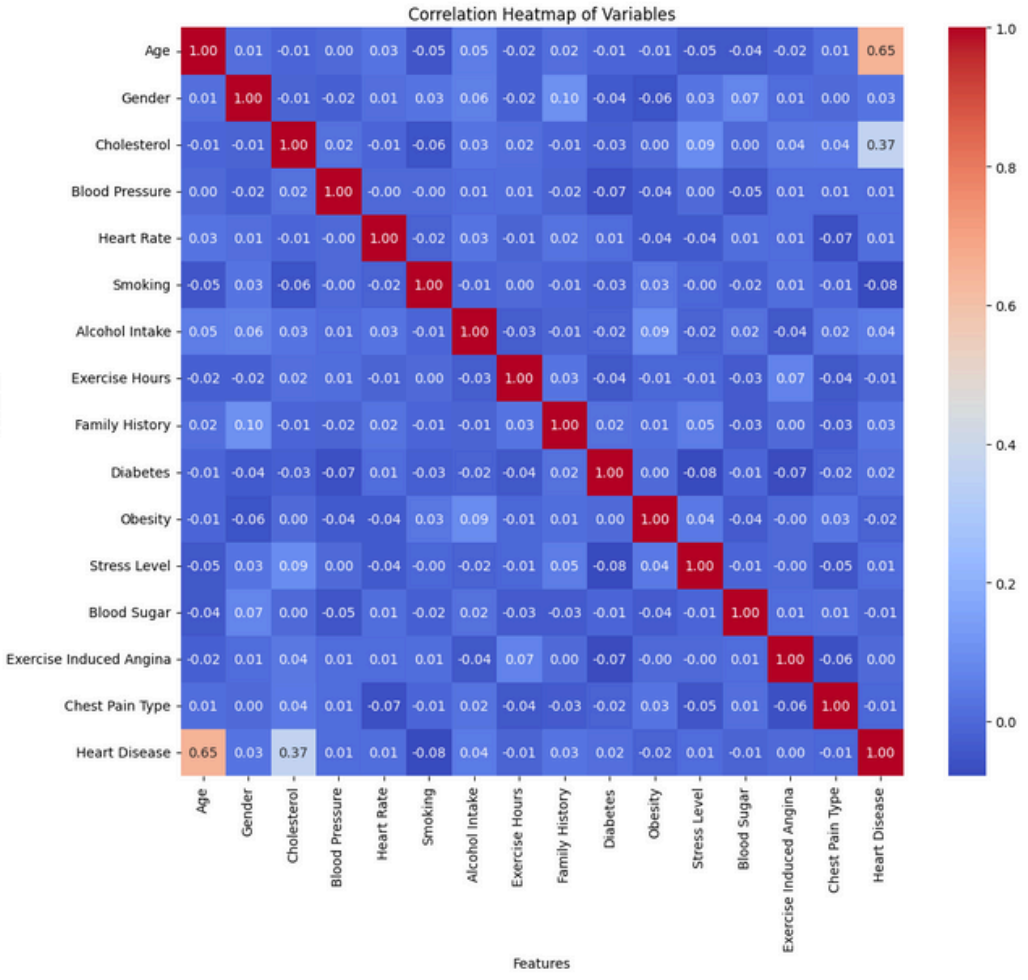


CORRELATION HEATMAP OF VARIABLES

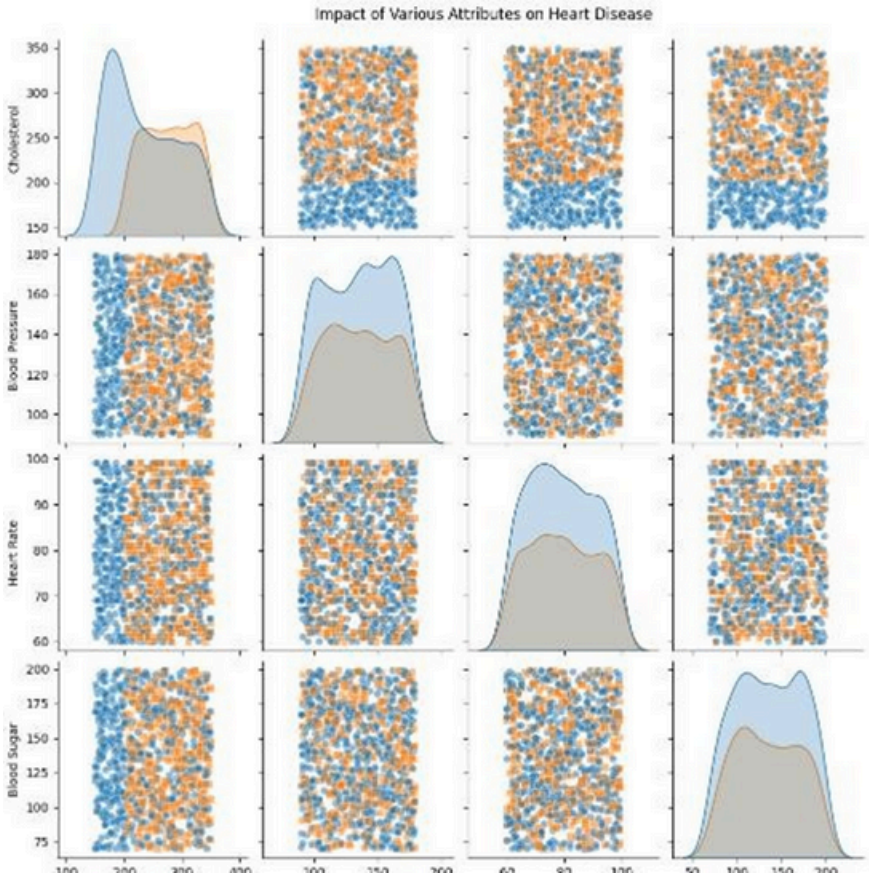
Cholesterol and Heart Disease: It shows moderate positive correlation with value 0.37

Smoking and Heart Disease: It shows negative correlation

Age and Heart Disease: light red color shows positive correlation between two variables with value 0.65



MEDICAL HISTORY ANALYSIS



Cholesterol has blue high peek near 200 in diagonal plots showing highest density of people with no heart disease

Distribution of cholesterol with Blood Pressure, Heart Rate and Blood sugar in non-diagonal plots indicates the X-axis variables (Heart Rate, Blood Sugar, Blood Pressure) are strong indicator of heart disease

FEATURE ENGNEERING

Creating New Features:

Log Blood Pressure: A logarithmic transformation of blood pressure was performed to normalize the distribution and reduce the impact of extreme values.

Interactive Terms:

Cholesterol_BloodPressure: A composite feature combining cholesterol and blood pressure levels to capture their combined effect on heart health.

Exercise_Stress: A measure of stress induced by exercise, which is a significant indicator of heart disease risk.

Domain Specific Features:

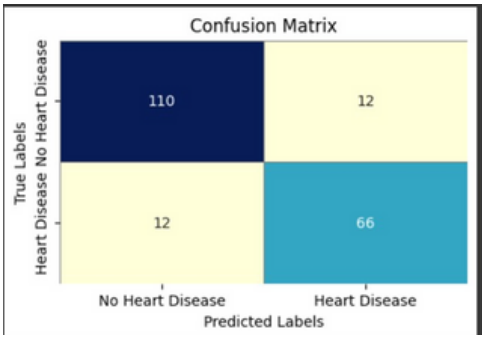
Cholesterol_Ratio: The ratio of HDL (good) cholesterol to LDL (bad) cholesterol, a well-known indicator of heart health.

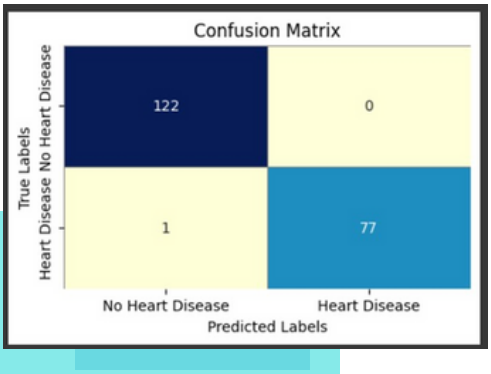
Mean Arterial Pressure: A calculated feature representing the average arterial pressure, providing insight into the overall blood pressure condition.

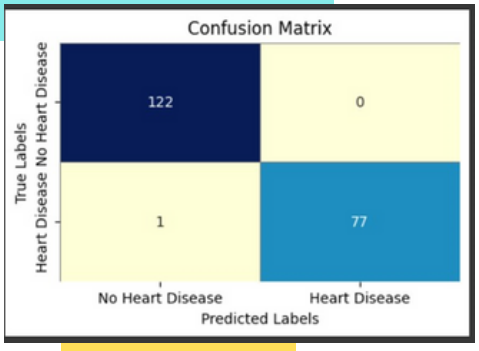
Risk Score: A composite risk score derived from multiple factors such as age, cholesterol levels, and blood pressure, which are known to influence heart disease risk

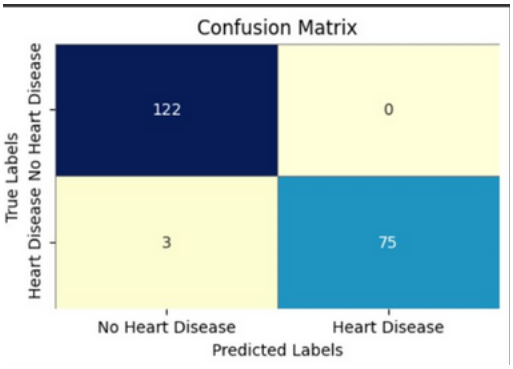
MODEL **BUILDING**

CONFUSION MATRIX









MODELS

Logistic Regression

Logistic Regression is a linear model widely used for binary classification tasks, such as predicting the presence or absence of heart disease. It works by estimating the probability that a given input belongs to a particular class using a logistic function. The model outputs probabilities, which are then thresholded to make a final classification decision. Logistic Regression is appreciated for its simplicity, ease of implementation, and interpretability, making it a strong baseline model for many classification problems.

Decision Tree

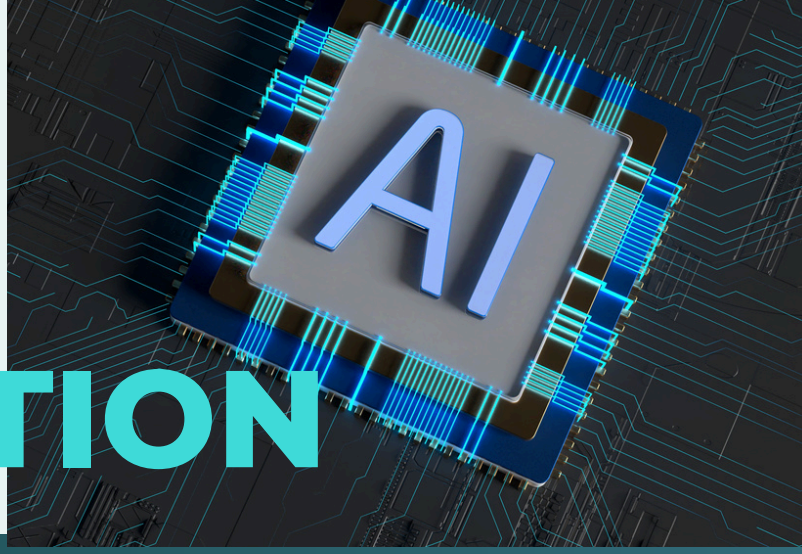
Decision Trees are non-linear models that recursively split the data into subsets based on feature values, forming a tree-like structure. Each node in the tree represents a feature, each branch represents a decision rule, and each leaf node represents an outcome. Decision Trees are intuitive and easy to visualize, making them useful for understanding the decision-making process. They handle both numerical and categorical data and are capable of capturing non-linear relationships in the data.

Random Forest

Random Forests are an ensemble learning method that constructs multiple decision trees during training and merges their outputs to improve accuracy and control overfitting. Each tree in the forest is trained on a random subset of the data and features, and the final prediction is made by averaging the predictions (for regression) or majority voting (for classification). Random Forests are robust, handle large datasets well, and provide insights into feature importance.

XGBOOST

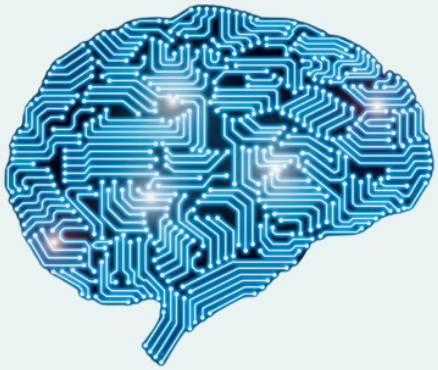
Gradient Boosting is another powerful ensemble technique that builds models sequentially, with each new model focusing on correcting the errors of the previous ones. Typically using decision trees as base learners, Gradient Boosting optimizes the model by minimizing a loss function in a stage-wise manner. This method is known for its high predictive accuracy and ability to handle a variety of data types, capturing complex relationships in the process.



MODEL EVALUATION

MODELS	F1 SCORE	Precision	Recall	Accuracy
Logistic Regression	0.90	0.91	0.89	0.905
XGBoost	0.98	0.99	0.98	0.985
Decision Tree	0.99	1.00	0.98	0.995
Random Forest	0.99	1.00	0.98	0.995

CONCLUSION



Random Forest and Decision Tree Outperforms other models giving us 99% Accuracy.

Chose Random Forest as it mitigate overfitting and work well on large dataset compared to Decision Tree.

Future Work

To further enhance the accuracy and robustness of the heart disease prediction model, expanding the dataset and employing advanced modeling techniques can be highly beneficial. The dataset can be increased by integrating additional data from public health databases and electronic health records, as well as utilizing data augmentation methods like SMOTE and GANs and Ensemble techniques such as stacking can be used for increased accuracy

RECOMMENDATIONS

1

Healthcare professionals should prioritize cardiovascular screenings and preventive measures for patients over the age of 50. Regular check-ups, lifestyle counseling, and early intervention can help mitigate risks associated with aging

2

Emphasize smoking cessation programs for current smokers to reduce their risk of heart disease. For former smokers, focus on maintaining healthy habits to prevent recurrence of heart disease

3

Conduct comprehensive assessments that include blood pressure monitoring, heart rate evaluation, and blood sugar tests. Understanding these interrelationships can aid in early detection and tailored treatment plans for individuals at risk

4

Educate patients about the impact of age, smoking habits, and cholesterol levels on heart health. Encourage lifestyle modifications such as healthy diet choices, regular exercise, and stress management.