Let’s talk about health

Heart Disease Prediction

Analyzing risk factors to forecast likelihood of developing heart conditions.

Problem Statement

The Problem

. Cardiovascular disease (CVD) is the leading global cause of death, disproportionately affecting low- and middle-income countries, where it accounts for three-quarters of all fatalities (Kontchou, McCracy, & Schulman, 2019).

In Kenya, CVD remains a significant public health burden, with a mortality rate of 13.8% in 2019.

The disease primarily affects individuals aged 30–70 years, with at least 2.5 heart surgeries performed per 1,000 individuals at a corresponding labor cost of US$526.

The goal

Despite growing awareness, early detection and risk assessment remain challenging due to limited local research and data-driven insights.

Rapid urbanization, lifestyle changes, and inadequate access to preventive healthcare further exacerbate the situation. This study aims to analyze key health indicators, identify major risk factors, and uncover trends to support healthcare professionals, policymakers, and researchers. By leveraging data-driven insights, this research seeks to improve early diagnosis, develop targeted interventions, and enhance public health strategies to combat heart disease in Kenya.

Objectives

i) To develop a Predictive Model by creating a machine learning that accurately predicts the likelihood of heart disease based on patient data.

ii) To identify Key Risk Factors which are the most significant predictors of heart disease.

iii) To provide healthcare professionals with a tool that aids in the early detection and management of heart disease.

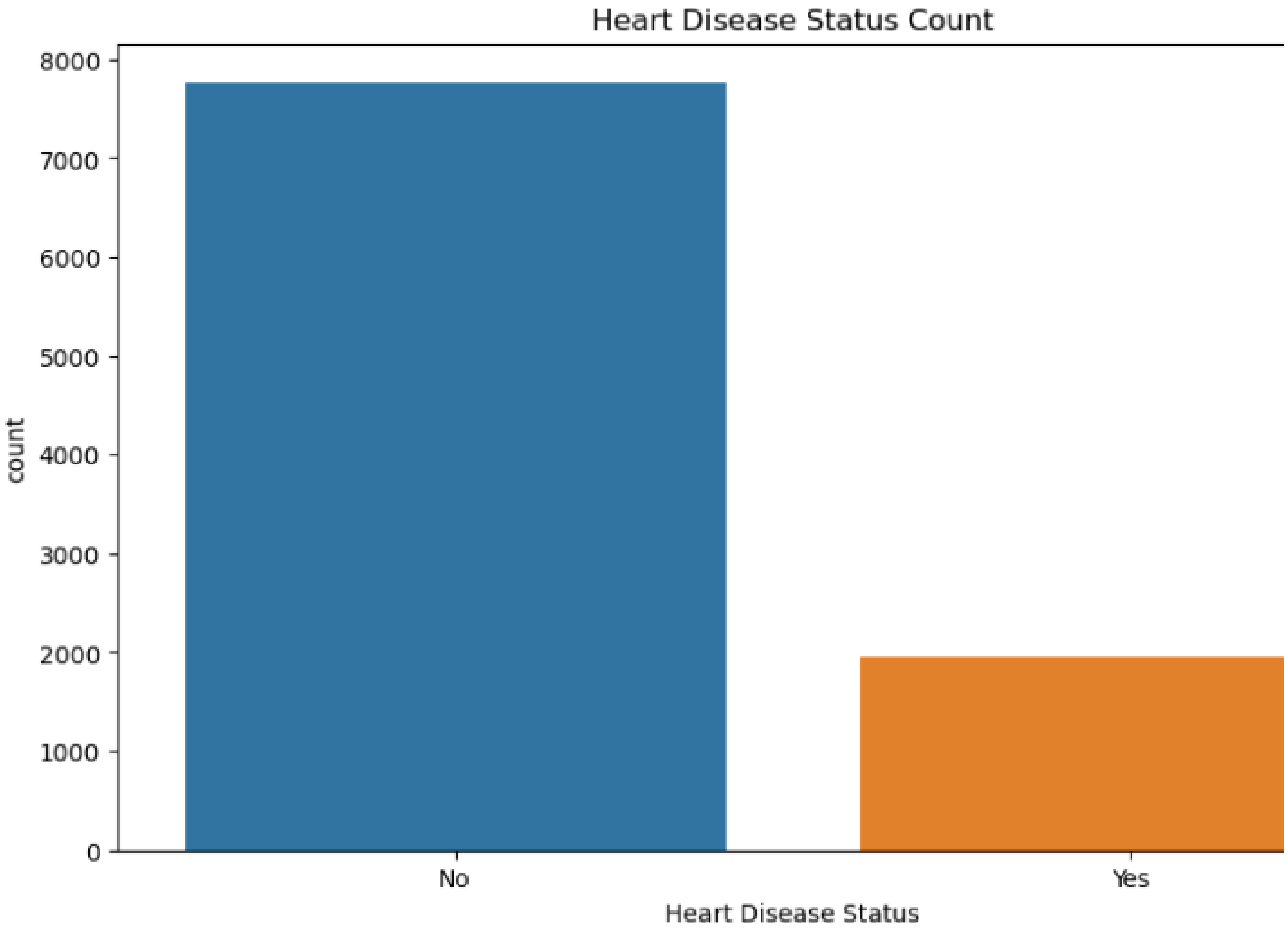
iv) To make informed decisions regarding resource allocation and preventive care strategies based on model predictions.

V) Increase awareness of heart disease risk factors among patients and the general public through educational materials and outreach programs.

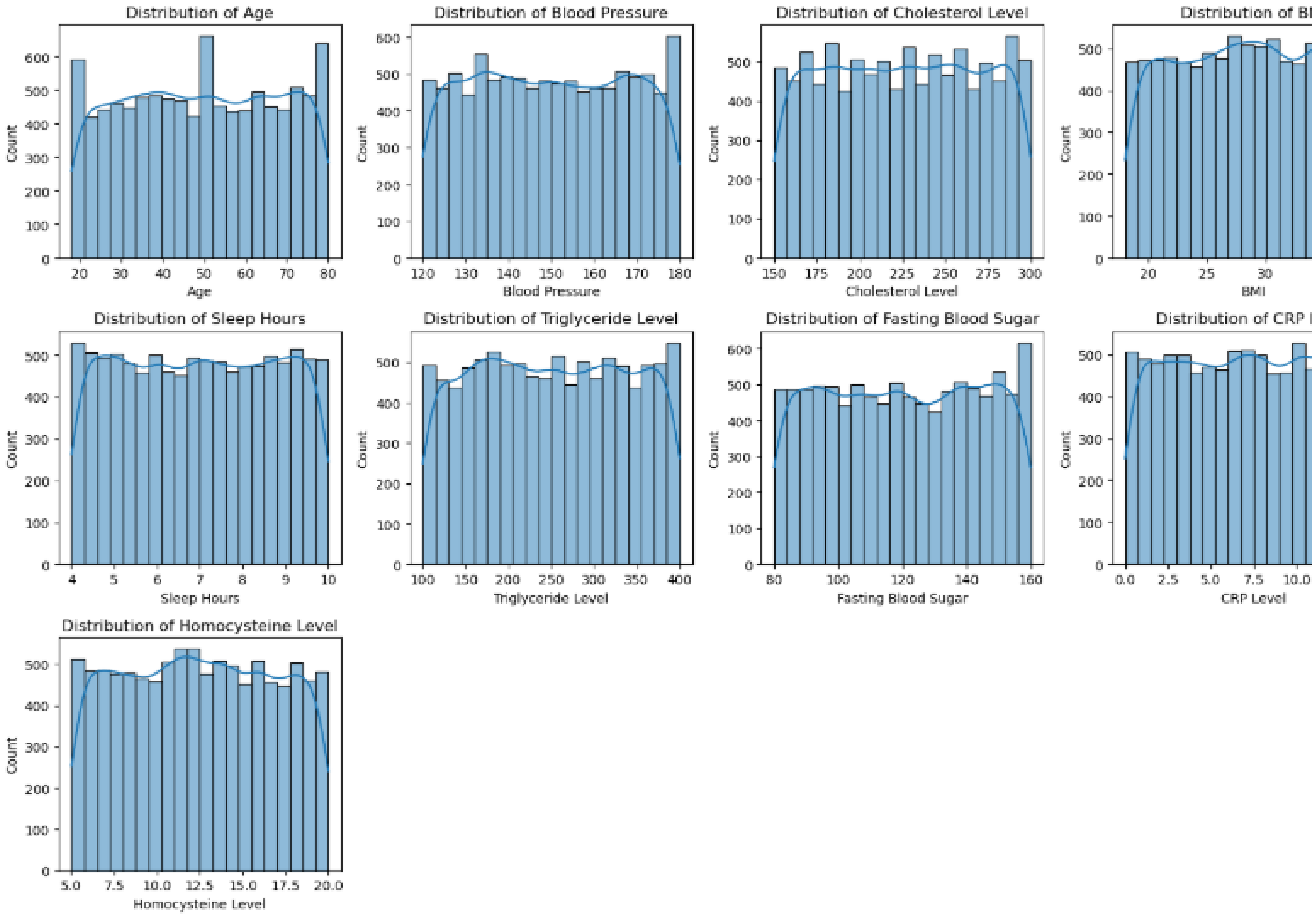
Data Understanding

- Our data was extracted from Kaggle (https://www.kaggle.com/datasets/oktayrdeki/heart-disease) which has 21 columns and 10,000 rows. The data has both numerical and categorical features. The label encoding was for used to assign unique numbers on categorical features because of its nominal data to ensure that the model can interpret the data, detect patterns, and make predictions effectively. Data cleaning was done to handle missing values and also check for outliers which showed there were no both missing values and outliers. The visualizations for outliers is as shown below:

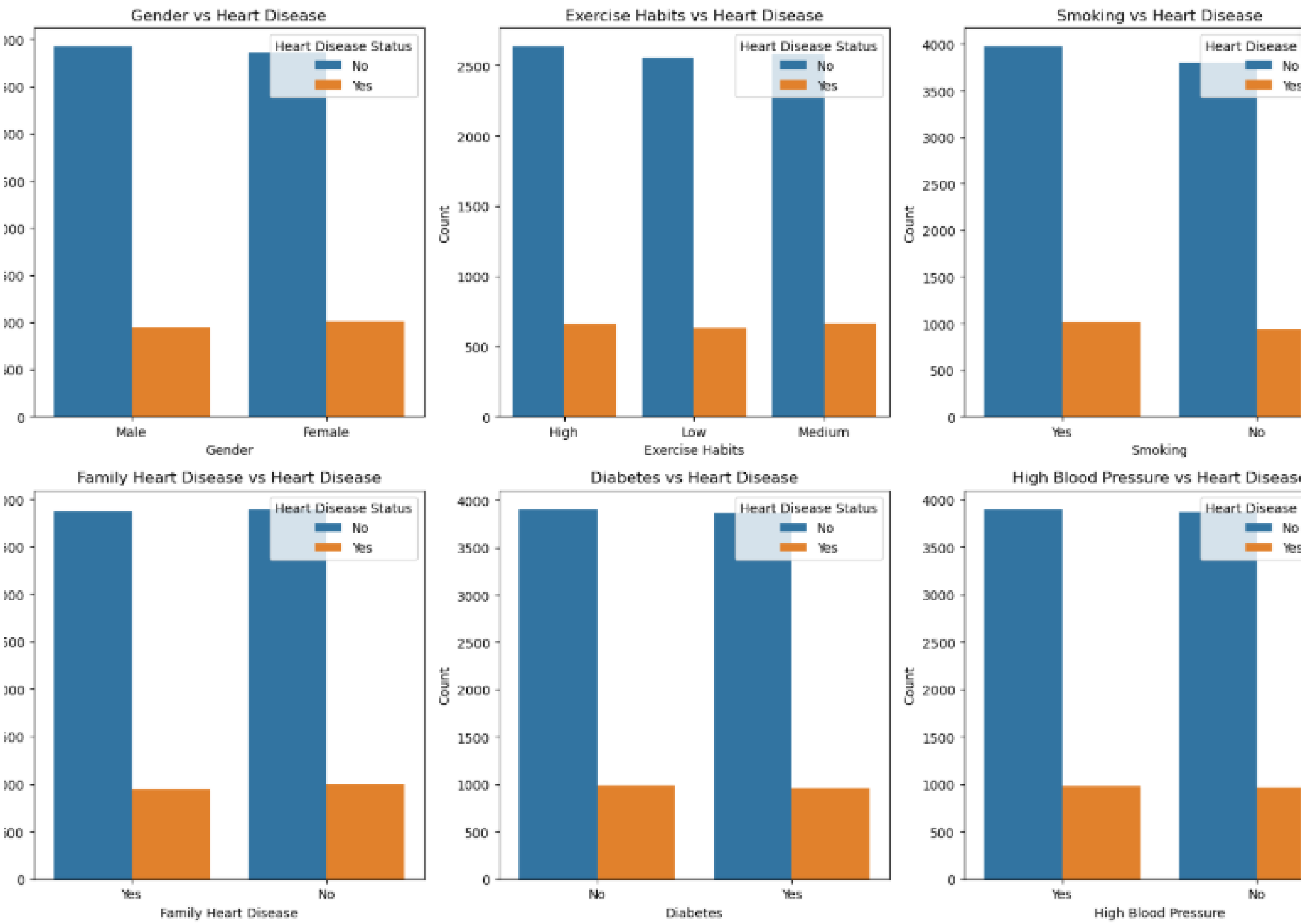
Visualization



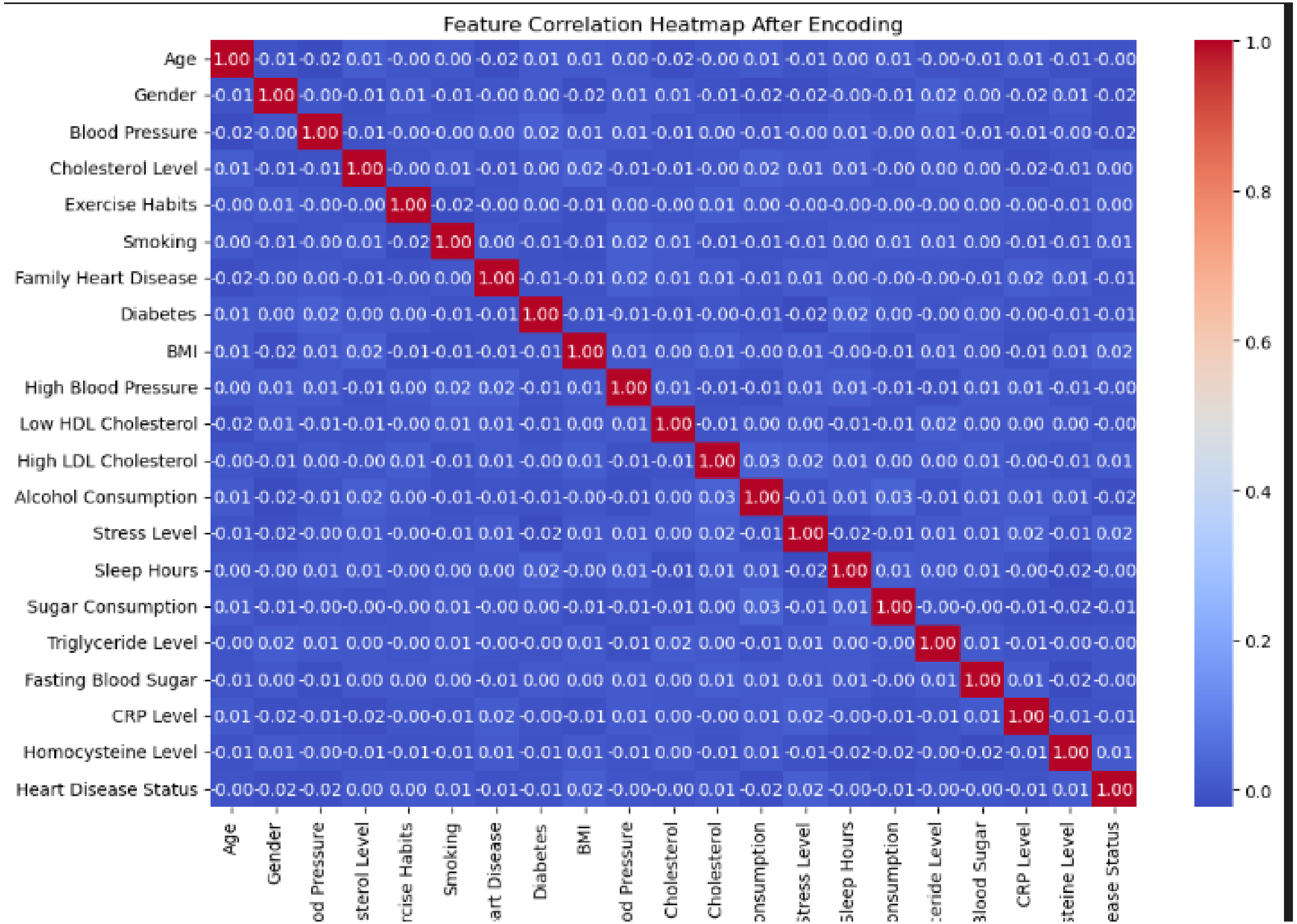
- **A count plot visualizes the distribution of patients with and without heart disease. The visualization showed that our data is imbalanced thus it will need to be balanced.**



Histograms were plotted for all numerical features to identify skewness, normality, and the presence of outliers. Histograms showed that our numerical features are normally distributed.



Count plots were used to analyze categorical variables to determine the association of certain features with heart disease. It was observed that patients who smoke are higher than those who don't smoke in respect to heart disease. The same applies to those with high blood pressure.



Feature Correlation Heatmap After Encoding

	Age	Gender	od Pressure	sterol Level	rcise Habits	Smoking	art Disease	Diabetes	BMI	od Pressure	Cholesterol	Cholesterol	onsumption	Stress Level	Sleep Hours	onsumption	ceride Level	Blood Sugar	CRP Level	steine Level	ease Status
Age	1.00	-0.01	-0.02	0.01	-0.00	0.00	-0.02	0.01	0.01	0.00	-0.02	-0.00	0.01	-0.01	0.00	0.01	-0.00	-0.01	0.01	-0.01	-0.00
Gender	-0.01	1.00	-0.00	-0.01	0.01	-0.01	-0.00	0.00	-0.02	0.01	0.01	-0.01	-0.02	-0.02	-0.00	-0.01	0.02	0.00	-0.02	0.01	-0.02
Blood Pressure	-0.02	-0.00	1.00	-0.01	-0.00	-0.00	0.00	0.02	0.01	0.01	-0.01	0.00	-0.01	-0.00	0.01	-0.00	0.01	-0.01	-0.01	-0.00	-0.02
Cholesterol Level	0.01	-0.01	-0.01	1.00	-0.00	0.01	-0.01	0.00	0.02	-0.01	-0.01	-0.00	0.02	0.01	0.01	-0.00	0.00	0.00	-0.02	-0.01	0.00
Exercise Habits	-0.00	0.01	-0.00	-0.00	1.00	-0.02	-0.00	0.00	-0.01	0.00	-0.00	0.01	0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.01	0.00
Smoking	0.00	-0.01	-0.00	0.01	-0.02	1.00	0.00	-0.01	-0.01	0.02	0.01	-0.01	-0.01	-0.01	0.00	0.01	0.01	0.00	-0.01	-0.01	0.01
Family Heart Disease	-0.02	-0.00	0.00	-0.01	-0.00	0.00	1.00	-0.01	-0.01	0.02	0.01	0.01	-0.01	0.01	0.00	-0.00	-0.00	-0.01	0.02	0.01	-0.01
Diabetes	0.01	0.00	0.02	0.00	0.00	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.00	-0.01	-0.02	0.02	0.00	-0.00	0.00	-0.00	-0.01	-0.01
BMI	0.01	-0.02	0.01	0.02	-0.01	-0.01	-0.01	-0.01	1.00	0.01	0.00	0.01	-0.00	0.01	-0.00	-0.01	0.01	0.00	-0.01	0.01	0.02
High Blood Pressure	0.00	0.01	0.01	-0.01	0.00	0.02	0.02	-0.01	0.01	1.00	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	-0.01	-0.00
Low HDL Cholesterol	-0.02	0.01	-0.01	-0.01	-0.00	0.01	0.01	-0.01	0.00	0.01	1.00	-0.01	0.00	0.00	-0.01	-0.01	0.02	0.00	0.00	0.00	-0.00
High LDL Cholesterol	-0.00	-0.01	0.00	-0.00	0.01	-0.01	0.01	-0.00	0.01	-0.01	-0.01	1.00	0.03	0.02	0.01	0.00	0.00	0.01	-0.00	-0.01	0.01
Alcohol Consumption	0.01	-0.02	-0.01	0.02	0.00	-0.01	-0.01	-0.01	-0.00	-0.01	0.00	0.03	1.00	-0.01	0.01	0.03	-0.01	0.01	0.01	0.01	-0.02
Stress Level	-0.01	-0.02	-0.00	0.01	-0.00	-0.01	0.01	-0.02	0.01	0.01	0.00	0.02	-0.01	1.00	-0.02	-0.01	0.01	0.01	0.02	-0.01	0.02
Sleep Hours	0.00	-0.00	0.01	0.01	-0.00	0.00	0.00	0.02	-0.00	0.01	-0.01	0.01	0.01	-0.02	1.00	0.01	0.00	0.01	-0.00	-0.02	-0.00
Sugar Consumption	0.01	-0.01	-0.00	-0.00	-0.00	0.01	-0.00	0.00	-0.01	-0.01	-0.01	0.00	0.03	-0.01	0.01	1.00	-0.00	-0.00	-0.01	-0.02	-0.01
Triglyceride Level	-0.00	0.02	0.01	0.00	-0.00	0.01	-0.00	-0.00	0.01	-0.01	0.02	0.00	-0.01	0.01	0.00	-0.00	1.00	0.01	-0.01	-0.00	-0.00
Fasting Blood Sugar	-0.01	0.00	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	-0.00	0.01	1.00	0.01	-0.02	-0.00
CRP Level	0.01	-0.02	-0.01	-0.02	-0.00	-0.01	0.02	-0.00	-0.01	0.01	0.00	-0.00	0.01	0.02	-0.00	-0.01	-0.01	0.01	1.00	-0.01	-0.01
Homocysteine Level	-0.01	0.01	-0.00	-0.01	-0.01	-0.01	0.01	-0.01	0.01	-0.01	0.00	-0.01	0.01	-0.01	-0.02	-0.02	-0.00	-0.02	-0.01	1.00	0.01
Heart Disease Status	-0.00	-0.02	-0.02	0.00	0.00	0.01	-0.01	-0.01	0.02	-0.00	-0.00	0.01	-0.02	0.02	-0.00	-0.01	-0.00	-0.00	-0.01	0.01	1.00

- **Correlation heatmap used to examine features with a strong correlation to heart disease e.g., cholesterol, exercise habits, smoking, etc.. These features will help us know the main risk factors leading to heart disease.**

Modeling

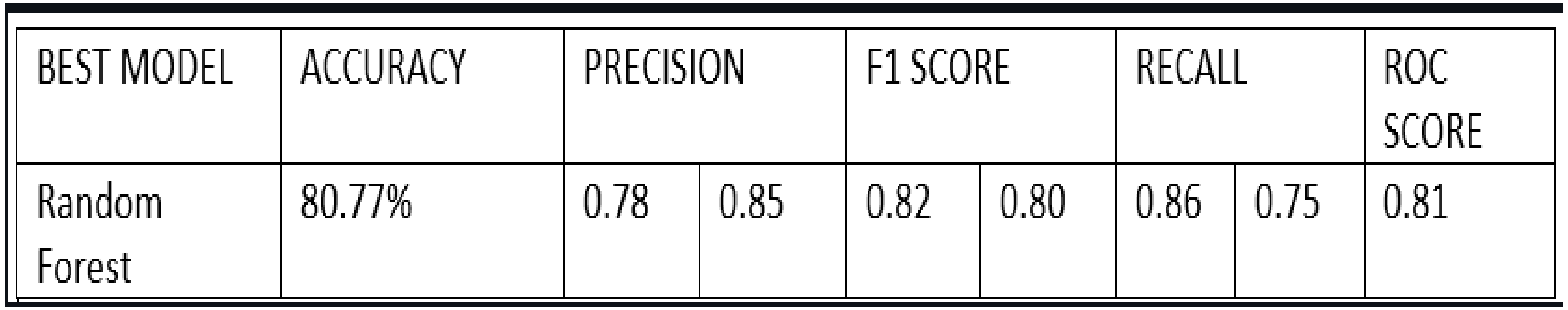
Modeling

In our project, we created 4 predictive models namely: Logistic Regression, Random Forest, GradientBoosting and Support Vector Classifier

The Random Forest model achieved an accuracy of 80.26%, which is slightly better than the 80% accuracy of the Logistic Regression model.

The classification report shows that the Random Forest model provides a good balance between precision and recall.

We shall therefore go ahead and tune the Random Forest Model to improve its performance



BEST MODEL	ACCURACY	PRECISION		F1 SCORE		RECALL		ROC SCORE
Random Forest	80.77%	0.78	0.85	0.82	0.80	0.86	0.75	0.81

- Our Model Random Forest had an overall accuracy score of 80.77%. There was a strong performance of class 0(Individuals whose heart disease status was NO) because of the recall 86%.However with an ROC score of 0.8 we can conclude that there was a strong predictive capability and that it only missed to detect 25% of heart disease cases

Conclusion

This project aimed to analyze and predict heart disease based on various health-related features. Through Exploratory Data Analysis (EDA), feature engineering, and model evaluation, we derived the following key insights:

1. Dataset Imbalance

- The count plot of patients with and without heart disease revealed an imbalanced dataset. Since an imbalanced dataset can negatively affect model performance, SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the classes before training machine learning models.

2. Data Distribution and Outliers

- Histogram analysis showed that the numerical features are normally distributed, which is beneficial for machine learning models that assume normality.
- Boxplots confirmed that there were no significant outliers, ensuring that extreme values would not distort predictions.

3. Key Risk Factors for Heart Disease

- A heatmap and feature correlation analysis revealed that features such as cholesterol levels, exercise habits, smoking, and stress levels had the strongest correlation with heart disease. These factors play a critical role in determining a patient's heart disease risk.

4. Feature Encoding

- Label Encoding was applied to transform categorical variables into numerical form. This was essential for incorporating ordinal categorical variables, such as exercise habits, into machine learning models.

5. Feature Importance and Correlation

- The correlation analysis showed BMI, Stress Level, and Homocysteine Level as the highest positively correlated features with heart disease, while Gender, Blood Pressure, and Alcohol Consumption had the lowest correlation.

6. Model Performance Comparison

- Four machine learning models—Random Forest, Logistic Regression, Gradient Boosting, and SVM—were trained and evaluated.
- Random Forest emerged as the best model with an accuracy of 80.25%.
- After hyperparameter tuning and advanced machine learning techniques, the Random Forest model’s accuracy improved from 80.25% to 80.76% with an ROC score of 0.8, indicating strong predictive capability.

Recommendation

Based on these findings, the following recommendations are suggested:

1. **Public Health Interventions**

- **Smoking, cholesterol level, exercise habits, BMI, stress level and Homocysteine Level are strong indicators of heart disease risk. Awareness programs and lifestyle modification campaigns should be implemented to reduce these risk factors in at-risk populations.**

2. **Government interventions**

- **The government should come up with strict punishment such as fines for people who smoke in undesignated places. They can also advice manufacturers to consider filters which are less harmful to the smokers.**
- **The government should allocate more funds in the health sector to facilitate frequent health camps and mobile clinics which can help in detecting heart diseases early on.**
- **Additionally, they should advice processors to correctly label ingredients such as quantity of sugar for consumer goods, this will aid consumers in their decision-making.**

3. **Clinical Application of the Model**

- **The improved Random Forest model (80.76% accuracy, 0.8 ROC score) can be integrated into a clinical decision support system (CDSS) to assist healthcare providers in identifying high-risk patients and making data-driven decisions.**

4. **Feature Refinement and Further Data Collection**

- **The correlation analysis showed that some features had minimal impact on heart disease prediction. Future research should focus on acquiring more relevant patient data, such as genetic factors or real-time heart monitoring metrics, to enhance model accuracy.**

5. **Deploying the Model in a Real-World Setting**

- **The model can be deployed as a web-based or mobile health application where users can input their health metrics and receive a risk assessment for heart disease. This could serve as an early warning system, prompting individuals to seek medical attention when necessary.**

6. **Continued Model Optimization**

- **Further improvements can be achieved through deep learning models, feature selection techniques, and real-time data analysis. Additionally, testing the model on external datasets will help validate its generalizability across different populations.**