Patients Risk Prediction for Cardiovascular Diseases

Aakash Dhruva

Anubhav Nehru

Deepti Hariyani Karthick Vel Kathirvel

Mayank Gupta

Zihao Zhu

# Project Goal:

The aim of this project is to create an efficient classification model for identifying the risk of cardiovascular disease (CVD) in patients. This initiative seeks to streamline the screening process, enabling healthcare professionals to detect CVD at an early stage. By doing so, the project aims to reduce waiting times for patients seeking appointments with heart specialists.

# Project Description:

In this project, we leverage the potential of machine learning to explore and uncover valuable insights in the realm of healthcare prediction tasks, with a specific focus on cardiovascular disease (CVD). By harnessing advanced analytical techniques, we aim to contribute to patient risk assessment, healthcare planning, and strategic decision-making within clinical settings.

Our dataset is sourced from the 2021 Behavioral Risk Factor Surveillance System (BRFSS) Dataset, available on Kaggle[[1](https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset)]. Comprising a substantial collection of approximately 309,000 observations, this dataset provides a comprehensive overview of each patient. It encompasses an array of diverse features including age, sex, self-reported general health, frequency of medical checkups, exercise habits, smoking history, as well as the presence of several prevalent diseases such as heart disease, skin cancer, other cancers, diabetes, and arthritis.

Our investigation employs a rigorous methodology that harnesses the power of machine learning to tackle this complex problem. We follow a structured approach that encompasses data preprocessing, feature engineering, model selection, and evaluation. Our investigation extends beyond mere inquiry, as we endeavor to identify and analyze intricate patterns that may exist across different demographic and health-related groups.

# Project Motivation:

Approximately 695,000 individuals in the United States lost their lives to heart disease in 2021, which translates to 1 in every 5 death[[2](https://www.cdc.gov/heartdisease/facts.htm)]. Also, the average wait time of 26.6 days for cardiology appointment in USA in 2022[[3](https://www.healthleadersmedia.com/clinical-care/physician-appointment-wait-times-have-increased-significantly-survey-finds)]. This project is motivated by the pressing requirement to develop an advanced screening tool for timely identification of cardiovascular disease risk, aiming to mitigate the burden of heart disease fatalities and enhance patient access to crucial heart specialist consultations.

Our project aims to create an accurate classification model and deepen our understanding of the complex interplay between various patient attributes and the likelihood of heart disease. By shedding light on these intricate connections, our project can empower healthcare professionals with an additional tool to make informed decisions, optimize resource allocation, and elevate patient care and outcomes.

As our investigation unfolds, we anticipate that our findings will not only contribute to the realm of predictive healthcare but will also underscore the immense potential of data-driven approaches to drive advancements in medical research and public health campaigns.

# Exploratory Data Analysis:

# This dataset contains 19 features, we separate them into numerical features and categorical features. The numerical features are height, weight, BMI (Body Mass Index), Fruit Consumption, Alcohol Consumption, Fried Potato Consumption, and Green Vegetable Consumption. The categorical data contains Sex, Age Category, Exercise, General Health, Checkup, Smoking History, Diabetes, Depression, Arthritis, Skin Cancer, Heart Disease (CVD in following passage), and Other Cancer.

# During the exploratory data analysis (EDA) phase, we thoroughly examined both the numerical and categorical features of the dataset. We observed that many patients rated their general health as either "very good" or "good." Additionally, most patients had undergone a medical checkup within the last year. However, a critical and noteworthy finding was the high level of class imbalance within the dataset. Approximately 90% of the patients in our dataset reported no occurrence of heart disease, while only 10% of the patients indicated suffering from heart disease. [[Univariate Analysis](#bookmark=id.3j3u1jpnws5v)]

# This significant class imbalance has implications for our subsequent analyses and modelling, as it may impact the accuracy and reliability of predictions related to heart disease.

# After conducting basic statistical analyses, we examined the impact of various features on the likelihood of developing heart disease. [[Bivariate Analysis](#bookmark=id.whdzsq4g3a5g)]

# Age vs. CVD: We observed a clear exponential increase in the likelihood of heart disease with advancing age. As patients get older, their risk of CVD rises notably.

# BMI vs. CVD: Patients with a BMI in the underweight category (0–18.5) had a low probability of developing heart disease, approximately 1.36%. However, the risk significantly escalated for overweight individuals (BMI 24.9-29.9) and obese individuals (BMI 29.9+), with probabilities of 35.96% and 40.92%, respectively.

# Smoking History vs. CVD: Smoking emerged as a major risk factor for heart disease. Smokers exhibited nearly double the likelihood of developing CVD compared to non-smokers, with probabilities of 11.64% and 5.66%, respectively.

# Exercise vs. CVD: Engaging in regular exercise demonstrated a notable protective effect against heart disease. Individuals who exercised had a significantly lower chance of developing CVD, with probabilities of 6.67% compared to 12.95% for non-exercisers.

# A majority takeaway from Health condition vs different type disease is that general health condition is a good indicator of less like to get heart disease, but for disease like skin cancer, the likelihood is the same across five different health condition.

# The exploratory data analysis (EDA) on gender (male vs. female) in relation to cardiovascular disease (CVD) revealed significant differences in the likelihood of developing CVD between the two groups. The dataset comprised of a roughly equal number of male and female count. However, when considering the presence of heart disease, 60.4% of males were affected compared to 39.6% of females. Thus, it was found that males are 50% more likely to develop CVD compared to females.

The analysis of drinking frequency in relation to cardiovascular disease (CVD) revealed interesting insights. Among the different drinking categories, moderate drinkers (7-14 days) showed the lowest percentage of individuals with CVD, at 4.8%. In contrast, non-drinkers had a higher percentage of CVD at 10.6%, followed by heavy drinkers (15 days and above) with 7.2%, and light drinkers (1-6 days) with 6.1%. This suggests that moderate drinkers tend to have a relatively lower risk of developing cardiovascular disease compared to non-drinkers, light drinkers, and heavy drinkers.

# Feature Engineering:

The approach to predictive modeling encompassed data quality measures and feature engineering, including handling duplicates, and outlier treatment using height-to-weight ratio, and standard deviation criteria. Interaction terms, BMI categories, and encoding techniques were crafted, along with mapping health checkup frequency and applying one-hot encoding for gender. A lifestyle score was created using an autoencoder with LeakyReLU and ReLU functions, balanced with SMOTE, and proven more effective than Near Miss algorithms. Logistic regression, random forest, and gradient-boosting classifiers were implemented. Despite the extensive work to integrate the lifestyle score, it did not have a notable impact on performance, so it was excluded from the features used in further modeling.

**Modeling:**

Based on the insights from the EDA we used 16 columns(general\_health, age\_category, depression, arthritis, alcohol\_consumptions, etc), and given that our focus was on building a model with a high recall rate in order to classify a patient at risk of heart disease, we leveraged the Logistic Regression model to achieve a recall score of 0.81. We also built tree-based models - Random Forest and XGBoost, but the logistic regression model performed better, maybe because of the linear trend in the data.

The important variables were age\_category, general\_health, smoking\_history and gender. The important variables align with our observations in the preliminary EDA. [[Model Output :](#bookmark=id.7m1it395rdwo)]

# Insights and Recommendations:

1. The patient screening tool we propose must include "Age category", "general health", "smoking history", and "gender" since they emerged as important variables in predicting patients' susceptibility to CVD.
2. Health professionals can use this tool to improve patient screening and shorten the time patients must wait to see specialists.
3. Individuals: self-assessment / increase awareness and emergency handling

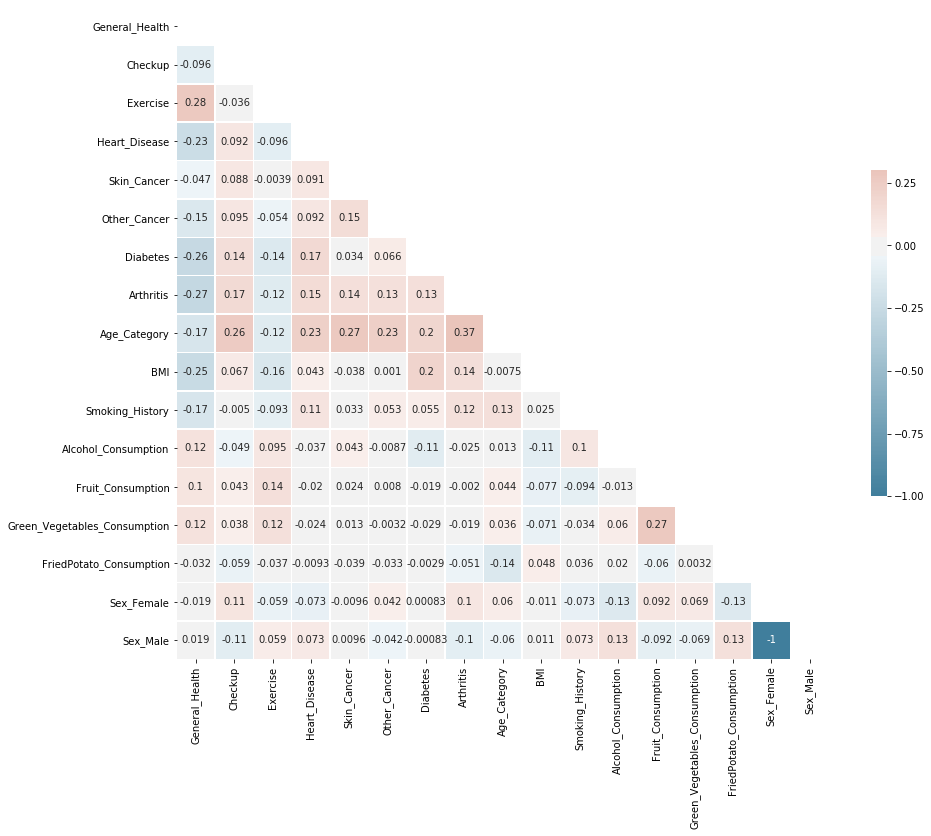
# Appendix:

1. <https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset>
2. <https://www.cdc.gov/heartdisease/facts.htm>
3. <https://www.healthleadersmedia.com/clinical-care/physician-appointment-wait-times-have-increased-significantly-survey-finds>

Data Schema

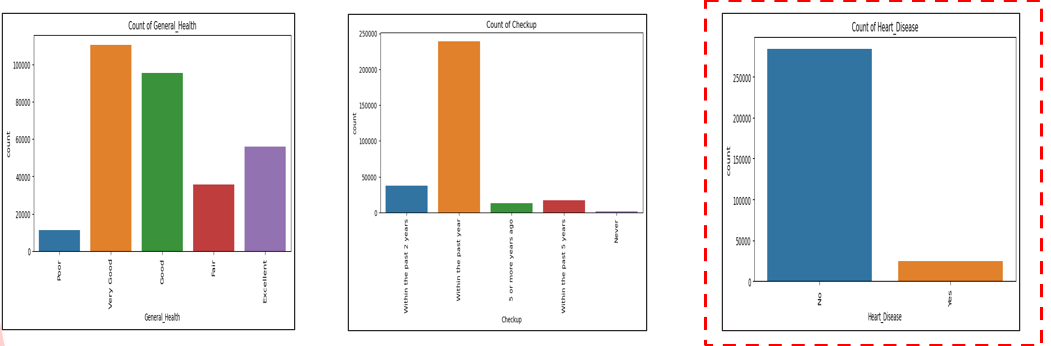
| **Predictor Variables** | **Variable Source** | **Description** |
| --- | --- | --- |
| General\_Health | Independent | How does a person feel about general health level |
| Checkup | Independent | When did a person had last checkup |
| Exercise | Independent | Does a person Exercise |
| **Heart Disease** | **Dependent** | Whether a person has heart disease |
| Skin Cancer | Independent | Whether a person has Skin Cancer |
| Diabetes | Independent | Whether a person has Diabetes |
| Arthritis | Independent | Whether a person has Arithritis |
| Sex | Independent | Gender of the patient |
| Age Category | Independent | Age of the patient |
| BMI | Independent | Body Mass Index of the patient |
| Smoking History | Independent | Whether a person smokes or not |
| Alcohol Consumption | Independent | No of times alcohol consumed per month |
| Fruit Consumption | Independent | No of times fruits consumed per month |
| Green\_Vegetables\_Consumption | Independent | No of times green vegetables consumed per month |
| FriedPotato\_Consumption | Independent | No of times fried potato consumed per month |

Correlation Matrix



*Figure 1: Correlation of Each Variables on Complete Dataset*

Univariate Analysis



*Figure 1: Findings from univariate analysis*

Bivariate Analysis

A graph of people with numbers and symbols

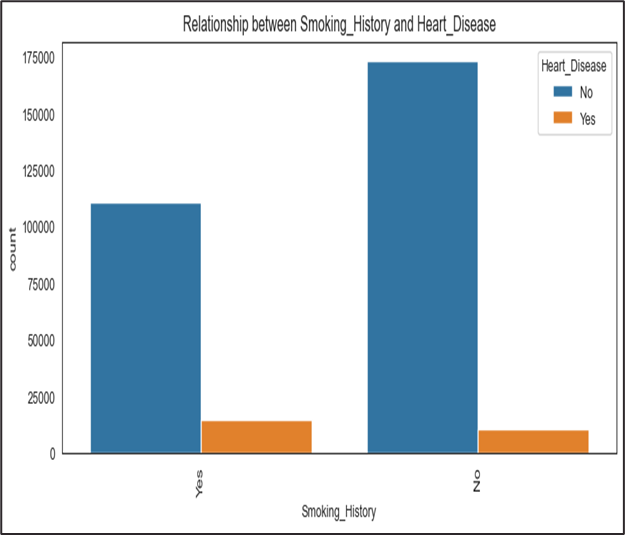
Description automatically generated

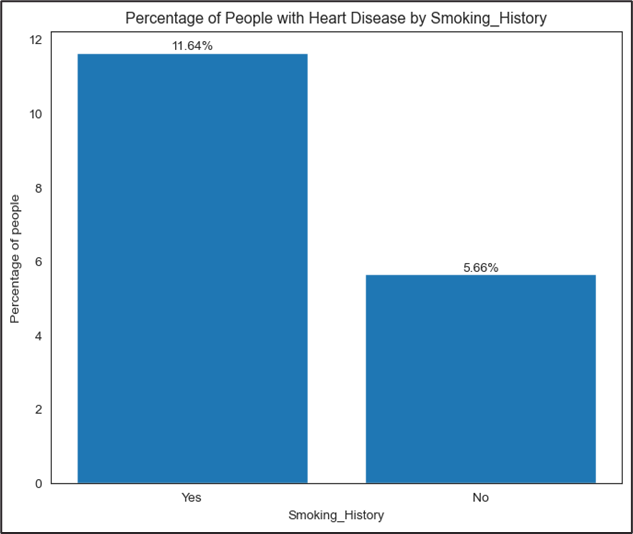
*Figure 2: Percentage of people in each Age Category v/s Heart Disease*

A graph of a number of blue rectangular objects

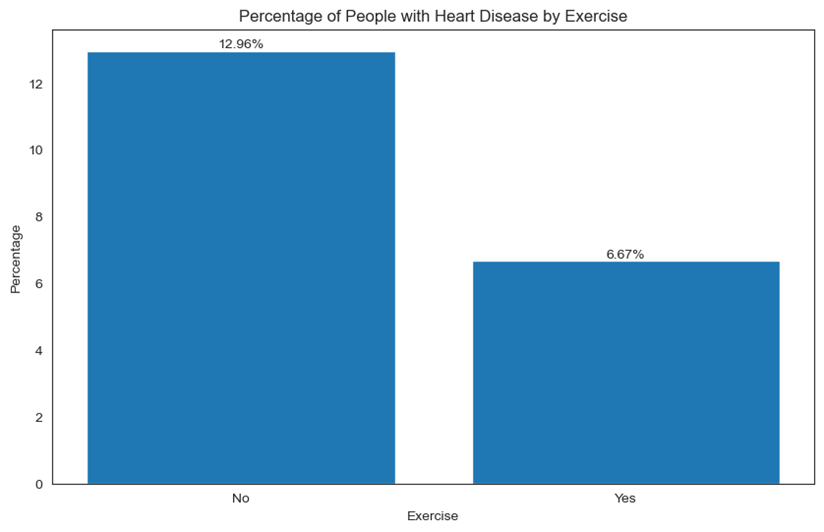
Description automatically generated

*Figure 3: Percentage of people in each BMI Category v/s Heart Disease*





*Figure 4: Smoking vs Heart Disease*

**

*Figure 5: Exercise vs Heart Disease*

Model Output :

| Model | Accuracy | Precision | Recall | Parameter |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.71 | 0.81 | 0.19 | Optimal Threshold = 0.47 |
| Random Forest Classifier | 0.74 | 0.19 | 0.70 | n\_estimators = 30, max\_depth = 3 , min\_samples\_split = 5, criterion = 'entropy' |
| XG Boost Classifier | 0.69 | 0.18 | 0.81 | learning\_rate = 0.01 ,n\_estimators = 70,max\_depth = 3, eval\_metric='logloss',use\_label\_encoder=False |

*A graph of a number of people

Description automatically generated*

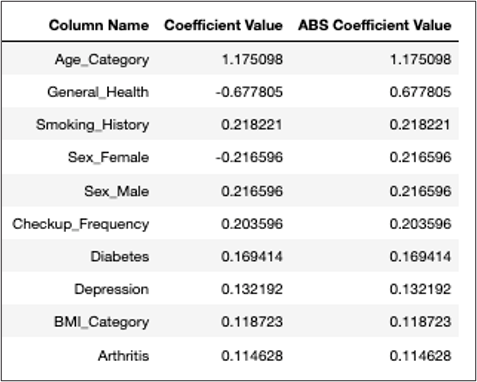
*Fig 17 : Important Variables - Random Forest*

*A graph with text on it

Description automatically generated*

*Fig 17 : Important Variables - XGBoost*

Logistic Regression Coefficients:

**

*g 19 : Coefficient Relationship for Complete Dataset*