# **Exploratory Data Analysis (EDA) and Model Evaluation Report**

## **1️ Introduction**

This report presents the exploratory data analysis (EDA) and evaluation of various machine learning models used to predict heart disease risk. The analysis includes key insights from the dataset and performance comparisons among different models.

## **2️Exploratory Data Analysis (EDA)**

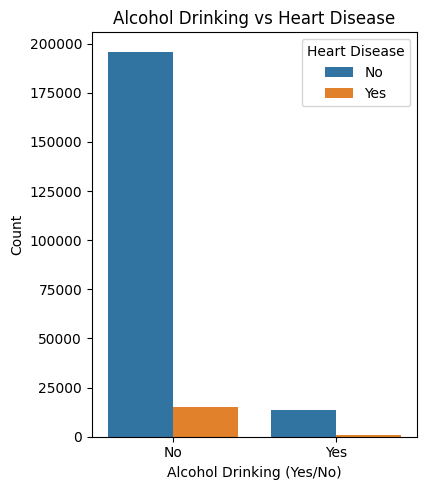
### **Key Variables in the Dataset**

The dataset includes various patient health attributes:

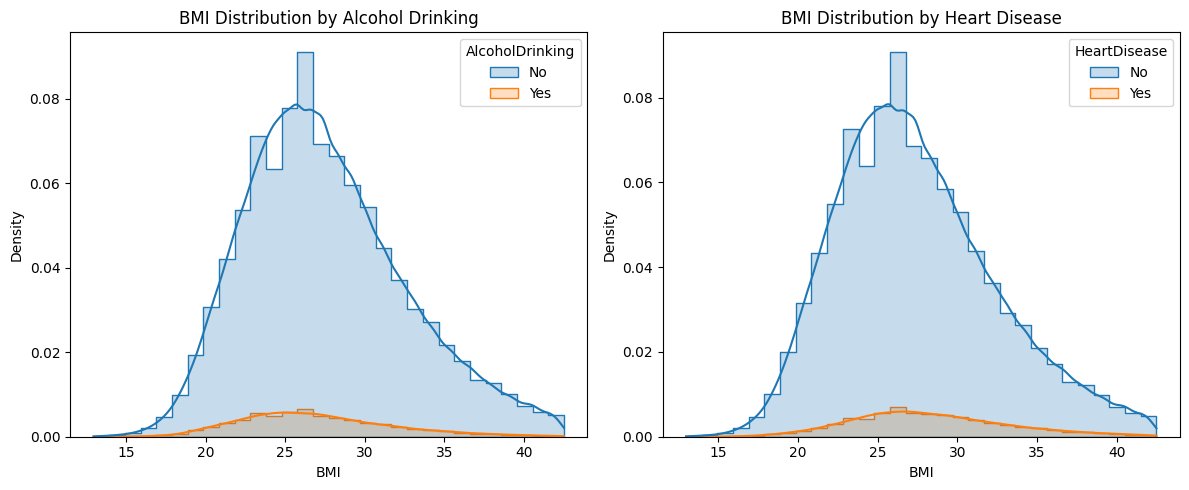
* **BMI (Body Mass Index):** Numeric value representing weight-to-height ratio.
* **Smoking & Alcohol Consumption:** Binary indicators for smoking and drinking habits.
* **Stroke, Diabetes, Kidney Disease, and Asthma:** Presence or absence of chronic conditions.
* **Physical & Mental Health Scores:** Number of poor health days in the past month.
* **General Health Status:** Categorical variable indicating overall health.
* **Physical Activity:** Whether the patient exercises regularly.
* **Age Category & Gender:** Demographic features.

### **Key Insights from EDA**

#### **1️ Alcohol Drinking vs. Heart Disease**

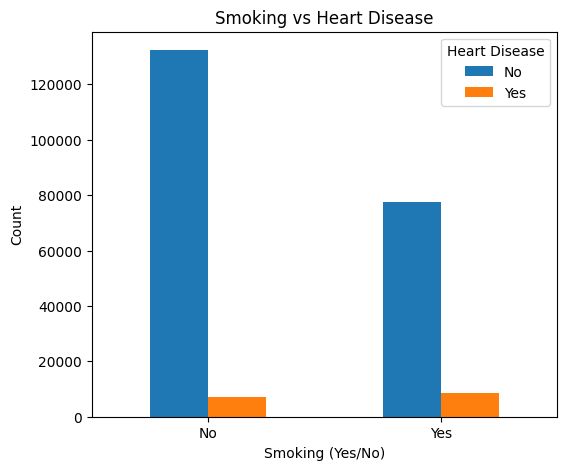
* The majority of individuals **do not consume alcohol**, and among them, heart disease cases are more prevalent.
* Among those who **drink alcohol**, heart disease cases are significantly lower, but the sample size is also smaller.

#### **2️ BMI Distribution by Alcohol and Heart Disease**

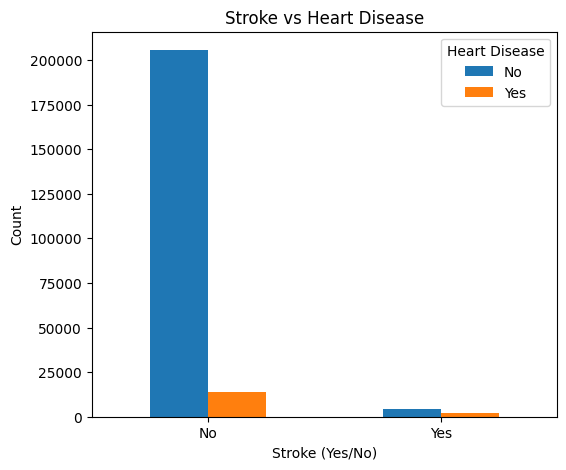
* BMI is **similarly distributed across drinkers and non-drinkers**.
* BMI **does not show a strong correlation** with heart disease risk alone.

#### **3️ Smoking vs. Heart Disease**

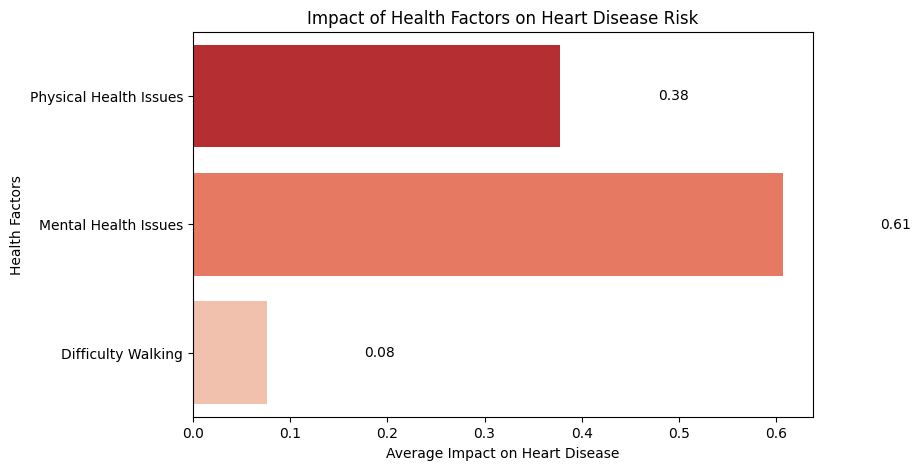
* Smokers **have a higher incidence of heart disease** compared to non-smokers.
* The total number of smokers is lower than non-smokers.



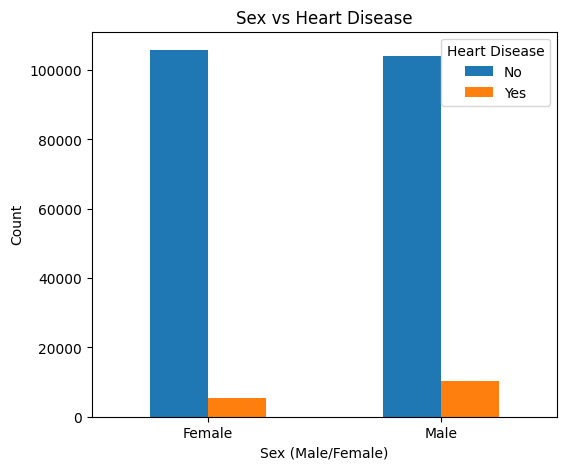
#### **4️ Stroke vs. Heart Disease**

* A history of stroke **significantly increases the likelihood of heart disease**.
* Heart disease is more common among stroke survivors than in the general population.

#### **5️ Impact of Health Factors on Heart Disease**

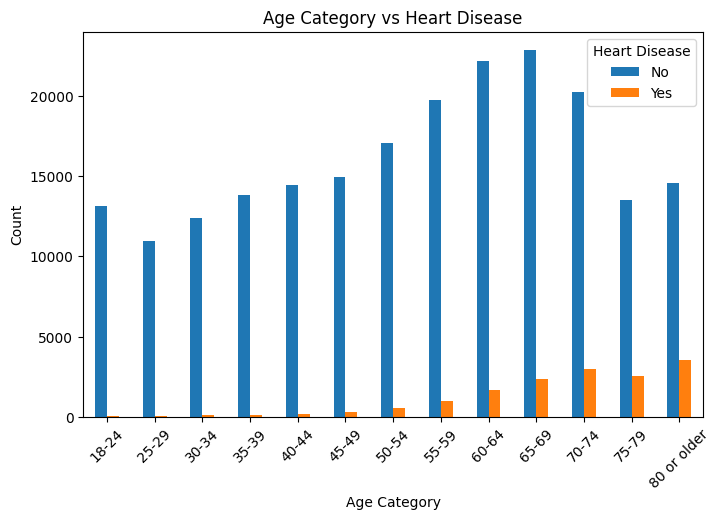
* **Mental health issues** have the highest impact on heart disease risk (0.61 correlation).
* **Physical health problems** also play a role (0.38 correlation).
* **Difficulty walking** has a smaller but notable impact (0.08 correlation).

#### **6️ Sex vs. Heart Disease**

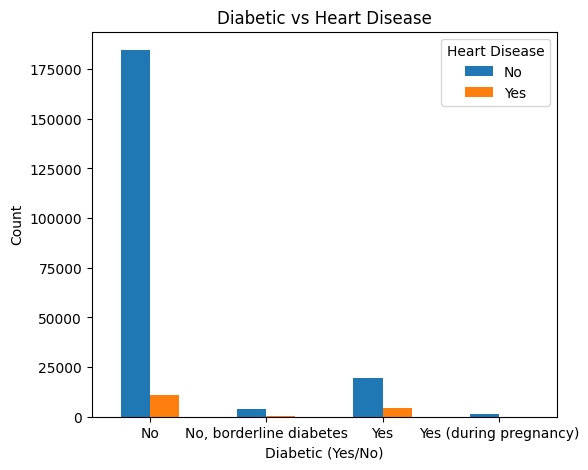
* **Men are more likely to have heart disease than women**, based on dataset trends.
* **Chi-squared test**
  + Chi-squared Statistic: 1394.5271147751523
  + P-value: 3.2486303915567704e-305
  + Degrees of Freedom: 1

#### **7️ Age vs. Heart Disease**

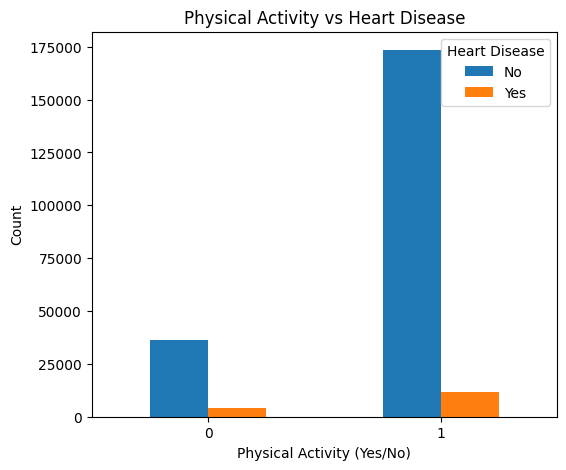
* **Heart disease prevalence increases significantly with age**, especially in those **above 50 years old**.



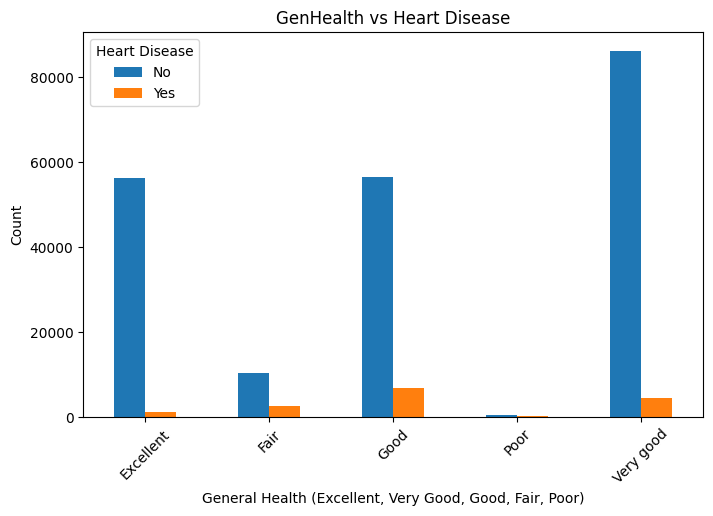
#### **8️ Diabetes vs. Heart Disease**

* **Diabetic individuals are more likely to have heart disease** than non-diabetic individuals.

#### **9️ Physical Activity vs. Heart Disease**

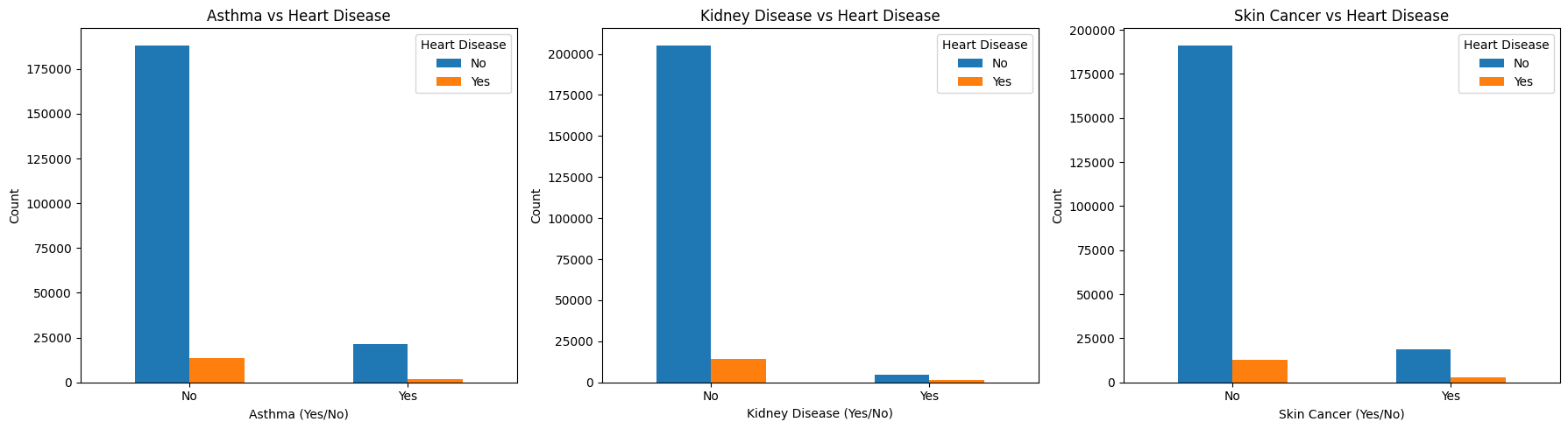
* **Regular physical activity is associated with a lower risk of heart disease**.

#### **10 General Health Status vs. Heart Disease**

* Individuals reporting **poor general health have a higher prevalence of heart disease**.

#### **1️1️ Other Conditions (Asthma, Kidney Disease, Skin Cancer) vs. Heart Disease**

* **Kidney disease has the strongest association** with heart disease among these conditions.
* Asthma and skin cancer show a weaker correlation. # Define the correct path to the scaler file



## **3️ Model Evaluation**

### **Models Used**

We tested multiple machine learning models:

* **Logistic Regression**
* **Random Forest**
* **XGBoost**
* **Support Vector Machine (SVM)**
* **Artificial Neural Network (ANN)**

### **Performance Comparison**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (0)** | **Recall (0)** | **Precision (1)** | **Recall (1)** | **F1-score** |
| **Logistic Regression** | 86% | 0.82 | 0.93 | 0.92 | 0.80 | 0.86 |
| **Random Forest** | 85% | 0.84 | 0.87 | 0.86 | 0.84 | 0.85 |
| **XGBoost** | **89%** | 0.85 | 0.95 | 0.94 | 0.84 | **0.89** |
| **SVM** | 86% | 0.83 | 0.92 | 0.91 | 0.81 | 0.86 |
| **Artificial Neural Network (ANN)** | **89%** | 0.83 | 0.97 | 0.96 | 0.81 | **0.89** |

### **Key Findings**

* **XGBoost and ANN performed the best**, achieving **89% accuracy**.
* **ANN achieved the highest F1-score (0.90 for class 0, 0.88 for class 1)**.
* **Logistic Regression and SVM performed similarly (86%)**, showing reliability for structured datasets.
* **Random Forest performed slightly lower (85%)**, making it less preferable than XGBoost.

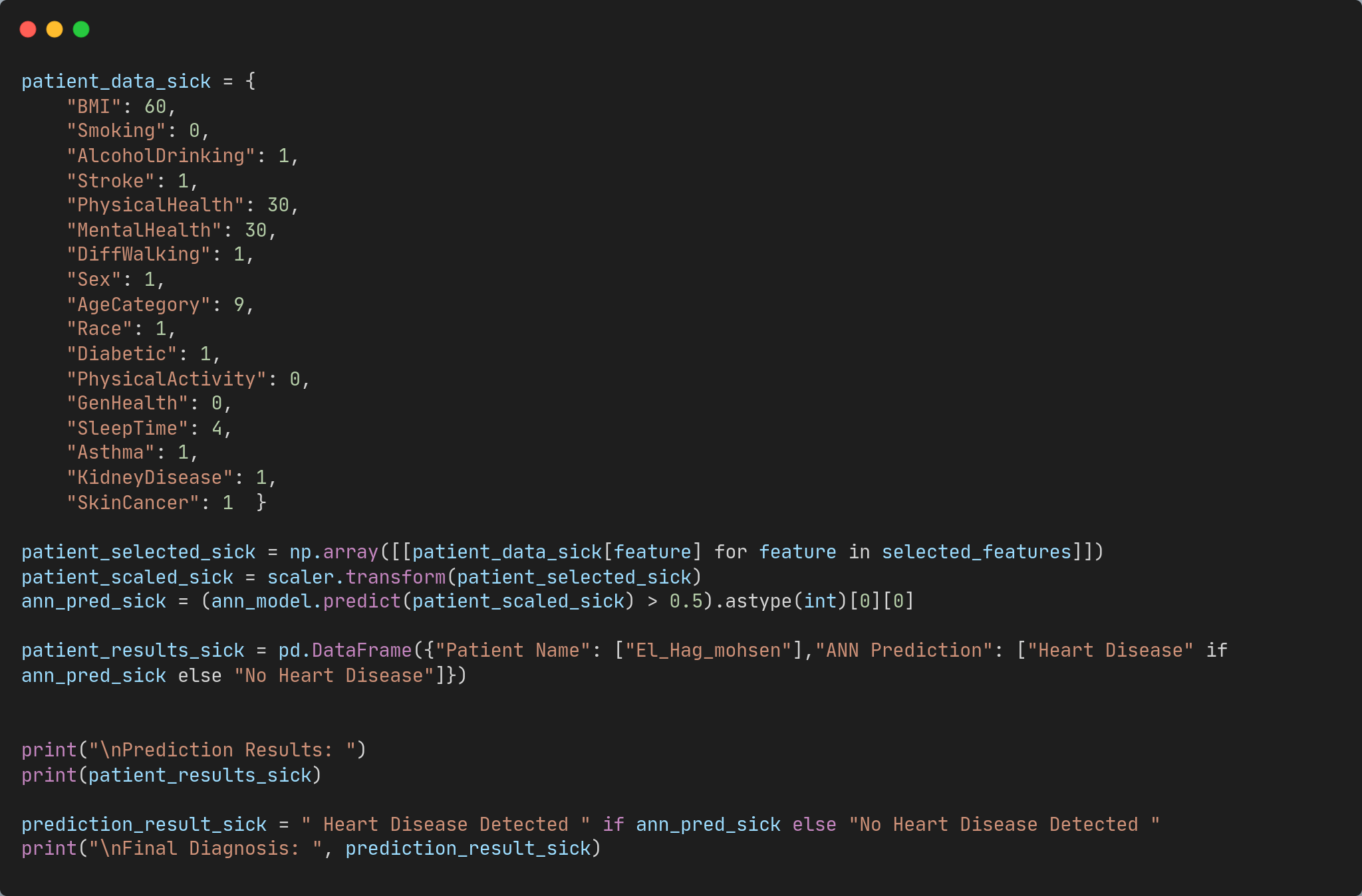
## **4️ Conclusion & Next Steps**

**Heart disease is strongly linked to age, mental health, physical inactivity, diabetes, and kidney disease.** **XGBoost and ANN provide the most reliable predictions.**

**Further improvements could include hyperparameter tuning and feature engineering.**

**Next Steps:**

* Conduct real-world testing with new patient data**………………( I'm doing it now)**



Prediction Results:

Patient Name ANN Prediction

0 El\_Hag\_mohsen

Final Diagnosis: **Heart Disease Detected**

**Prepared by:** *[shady mohammed ]***Date:** *[23/2/2025]*