**Name:** Ahmed Hazem Ibrahim **ID:** 211001984

**Heart Disease Prediction**

Heart Disease Data for Health Research

Dataset link: <https://www.kaggle.com/datasets/oktayrdeki/heart-disease/data>

GitHub repo link: <https://github.com/Ahmedhazem-2/Machine-Learning-Project>

Link to the Video:

<https://nileuniversity.sharepoint.com/sites/database240/_layouts/15/stream.aspx?id=%2Fsites%2Fdatabase240%2FShared%20Documents%2FGeneral%2FRecordings%2FMeeting%20in%20General%2D20250605%5F051704%2DMeeting%20Recording%2Emp4&referrer=StreamWebApp%2EWeb&referrerScenario=AddressBarCopied%2Eview%2E45b80f88%2D7019%2D460b%2D8847%2Da3ceffac4a1d>

**1. Introduction**

Heart disease remains one of the leading causes of mortality worldwide, impacting individuals across diverse age groups and demographics. Early identification and accurate prediction of heart disease risk are essential for implementing timely interventions and improving long-term health outcomes. Traditionally, clinicians rely on a combination of medical history, physical examinations, and laboratory tests to assess cardiovascular health. However, with the growing availability of large-scale health datasets and advancements in machine learning (ML), data-driven models are increasingly being utilized to enhance diagnostic accuracy and personalize risk assessment. These models can uncover complex patterns and interactions among various risk factors, offering a more nuanced understanding of individual susceptibility to heart disease.

**2. Objective**

The primary objective of this project is to develop a machine learning model that predicts the likelihood of heart disease in patients using key clinical and lifestyle factors, including age, blood pressure, cholesterol levels, BMI, smoking status, family history, and other relevant health indicators.

**3. Dataset Description**

This dataset includes a variety of health indicators and risk factors associated with heart disease. It consists of 10,000 records and 21 columns. Data on parameters such as age, gender, blood pressure, cholesterol levels, smoking habits, and exercise patterns have been gathered to assess heart disease risk and support health research. Healthcare professionals, researchers, and data analysts can utilize this dataset to explore heart disease trends, pinpoint risk factors, and conduct diverse health-related analyses.

Columns:

* Age: The individual's age.
* Gender: The individual's gender (Male or Female).
* Blood Pressure: The individual's blood pressure (systolic).
* Cholesterol Level: The individual's total cholesterol level.
* Exercise Habits: The individual's exercise habits (Low, Medium, High).
* Smoking: Whether the individual smokes or not (Yes or No).
* Family Heart Disease: Whether there is a family history of heart disease (Yes or No).
* Diabetes: Whether the individual has diabetes (Yes or No).
* BMI: The individual's body mass index.
* High Blood Pressure: Whether the individual has high blood pressure (Yes or No).
* Low HDL Cholesterol: Whether the individual has low HDL cholesterol (Yes or No).
* High LDL Cholesterol: Whether the individual has high LDL cholesterol (Yes or No).
* Alcohol Consumption: The individual's alcohol consumption level (None, Low, Medium, High).
* Stress Level: The individual's stress level (Low, Medium, High).
* Sleep Hours: The number of hours the individual sleeps.
* Sugar Consumption: The individual's sugar consumption level (Low, Medium, High).
* Triglyceride Level: The individual's triglyceride level.
* Fasting Blood Sugar: The individual's fasting blood sugar level.
* CRP Level: The C-reactive protein level (a marker of inflammation).
* Homocysteine Level: The individual's homocysteine level (an amino acid that affects blood vessel health).
* Heart Disease Status: The individual's heart disease status (Yes or No).

**4. Methodology**

The methodology for this project involves a systematic approach to building and evaluating a machine learning model that predicts the likelihood of heart disease in patients. This approach includes data collection, preprocessing, exploratory data analysis, feature engineering, model selection, training, evaluation, and interpretation.

1. Data Collection

The dataset used in this study was obtained from a real-world source and contains comprehensive health-related information for a large number of patients. The dataset includes features such as age, gender, systolic and diastolic blood pressure, cholesterol levels (low-density lipoprotein (LDL), high-density lipoprotein (HDL)), lifestyle factors (e.g., smoking status, diabetes), family history of heart disease, Body Mass Index (BMI), serum creatinine levels, ejection fraction, and other relevant clinical metrics. These variables are crucial for understanding cardiovascular health and identifying risk factors associated with heart disease.

2. Data Preprocessing

Before analysis, the dataset underwent several preprocessing steps:

- Handling Missing Values: Any missing values in the dataset were imputed using appropriate statistical measures, such as the KNN-imputer.

- Encoding Categorical Variables: Categorical features such as gender, smoking status, diabetes, family history, and others were encoded into binary values (0 or 1) using techniques like label encoding.

- Normalization/Standardization: Numerical features such as age, blood pressure, BMI, and cholesterol levels were standardized to ensure that all features contributed equally to the model’s performance.

3. Exploratory Data Analysis (EDA)

An exploratory data analysis was conducted to understand the distribution of the target variable (presence or absence of heart disease) and its relationship with other features.

Key aspects included:

- Visualization of Feature Distributions: Histograms and box plots were used to visualize the distributions of key features and identify potential outliers.

- Correlation Analysis: Correlation matrices were generated to identify relationships between numerical features and their influence on the target variable.

- Class Imbalance Check: The class distribution of the target variable (heart disease vs. no heart disease) was analyzed to determine if resampling techniques were necessary.

4. Feature Engineering

Based on insights from EDA, new features were derived where applicable, such as:

- Discretization of Continuous Variables: Some continuous features (e.g., age) were binned into meaningful categories (e.g., young, middle-aged, elderly) to simplify interpretation.

5. Model Selection

Several machine learning algorithms were considered for classification due to the binary nature of the target variable (heart disease: Yes / No). The following models were evaluated:

- Logistic Regression

- Random Forest Classifier

- Gradient Boosting Machines (e.g., CatBoost, AdaBoost)

- Support Vector Machine (SVM)

6. Training and Validation Strategy

The dataset was split into training and testing sets using an 80:20 ratio. Cross-validation techniques, particularly stratified k-fold cross-validation (k=5), were employed to ensure robustness in model evaluation. Hyperparameter tuning was performed using grid search and randomized search methods to optimize model performance.

7. Evaluation Metrics

To assess the effectiveness of the predictive models, the following evaluation metrics were used:

- Accuracy: Proportion of correctly classified instances.

- Precision: Ability of the model to not label a healthy patient as diseased.

- Recall (Sensitivity): Ability of the model to correctly identify patients with heart disease.

- F1-Score: Harmonic means of precision and recall, especially useful for imbalanced datasets.

- ROC-AUC: Area under the Receiver Operating Characteristic curve, which evaluates the model's ability to distinguish between classes at different probability thresholds.

8. Interpretation and Insights

Post-training interpretability techniques, such as feature importance plots, were used to explain model predictions and identify the most influential features contributing to heart disease prediction. This step is critical for translating the model into actionable clinical insights.

9. Deployment Considerations

Finally, considerations for deploying the model in a real-world setting were discussed, including:

- Integration into electronic health record (EHR) systems.

- Real-time prediction capabilities.

- Periodic retraining with updated data to maintain accuracy over time.