**Heart Disease Detection Using Machine Learning**

**ABSTRACT**

This project focuses on predicting the likelihood of heart disease in patients by analysing health metrics through machine learning techniques. The dataset includes features like age, gender, cholesterol levels, and blood pressure. Key steps include data preprocessing, visualization, feature correlation analysis, and evaluating various machine learning models to identify the most accurate approach for prediction. This comprehensive approach aims to improve diagnostic accuracy and assist healthcare professionals in making data-driven decisions.

**INTRODUCTION**

Heart disease remains a significant global health concern, responsible for millions of deaths annually. Early detection is crucial for reducing mortality rates and improving quality of life. Advances in machine learning offer promising solutions for analyzing complex clinical data to identify potential heart disease cases. This project leverages multiple machine learning techniques to predict heart disease, focusing on improving diagnostic accuracy and uncovering patterns in patient health data. By comparing various algorithms, the project identifies the most effective model for precise and reliable predictions.

**DATASET INFORMATION**

* **Source**: The dataset comprises anonymized clinical data collected from a diverse patient population. It includes essential health metrics and lifestyle factors, enabling a comprehensive analysis of heart disease risk.
  + id: Patient ID (removed during preprocessing).
  + age: Patient age (in days), later converted to years for better interpretability.
  + gender: Patient gender (1: Male, 2: Female).
  + height: Height (in cm), an important factor for BMI calculation.
  + weight: Weight (in kg), used for BMI and health assessments.
  + ap\_hi: Systolic blood pressure, a critical indicator of cardiovascular health.
  + ap\_lo: Diastolic blood pressure, complementing systolic measurements.
  + cholesterol: Cholesterol levels (1: Normal, 2: Above Normal, 3: Well Above Normal).
  + gluc: Glucose levels (1: Normal, 2: Above Normal, 3: Well Above Normal).
  + smoke: Smoking status (0: Non-smoker, 1: Smoker), a major risk factor.
  + alco: Alcohol consumption (0: No, 1: Yes), associated with lifestyle risk factors.
  + active: Physical activity (0: No, 1: Yes), crucial for overall health.
  + cardio: Target variable (0: No heart disease, 1: Heart disease).
* **Number of Records**: 70,000 entries, providing a robust dataset for analysis and modeling.
* **Preprocessing Steps**:
  + Removed irrelevant id column to focus on meaningful features.
  + Verified the absence of missing values to ensure data completeness.
  + Standardized features using StandardScaler for consistent scaling and improved model performance.

**DATA ANALYSIS AND VISUALIZATION**

**1. Feature Distributions**

* Histograms and density plots were generated to explore the distribution of each feature. These visualizations helped identify potential outliers and skewed data, particularly in blood pressure and weight metrics.

**2. Correlation Analysis**

* A heatmap was created to visualize the correlation matrix, highlighting relationships between features and the target variable (cardio).
* Key observations included:
  + Positive correlations between cholesterol levels, glucose levels, and the presence of heart disease.
  + Negative correlations between physical activity and the risk of heart disease.

**3. Key Insights**

* Patients with elevated cholesterol and glucose levels are significantly more prone to heart disease.
* Lifestyle factors, such as smoking and physical inactivity, contribute heavily to cardiovascular risk.
* The importance of systolic and diastolic blood pressure in predicting heart disease was evident.

**MACHINE LEARNING MODELS AND EVALUATION**

The dataset was split into training (70%) and testing (30%) subsets to evaluate model performance. Features were standardized using StandardScaler. Various machine learning algorithms were tested to identify the most accurate model:

**1. Logistic Regression (LR)**

* **Accuracy**: 0.74
* Logistic Regression provided a solid baseline, excelling in linear classification tasks.

**2. Support Vector Machine (SVM)**

* **Accuracy**: 0.77 (Linear Kernel).
* SVM demonstrated improved accuracy but required higher computational resources.

**3. K-Nearest Neighbors (KNN)**

* **Accuracy**: 0.72
* The performance of KNN varied based on the number of neighbors (k) and the size of the dataset. While simple, it struggled with scalability.

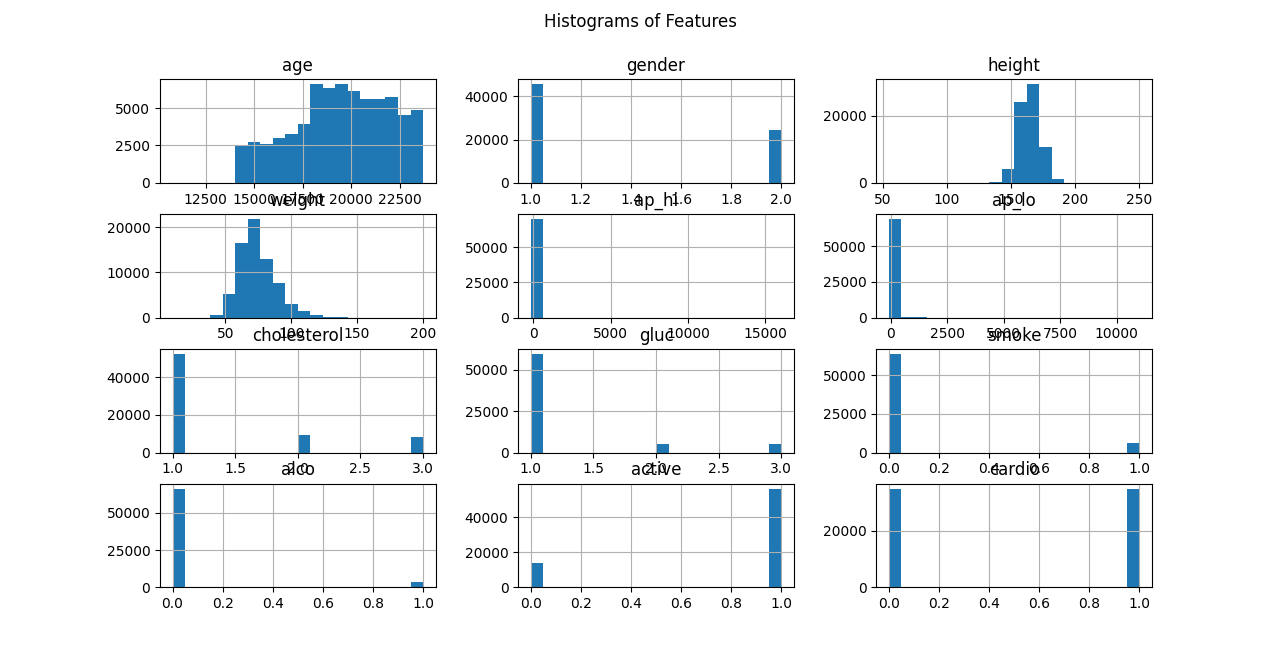
**4. Decision Tree (DT)**

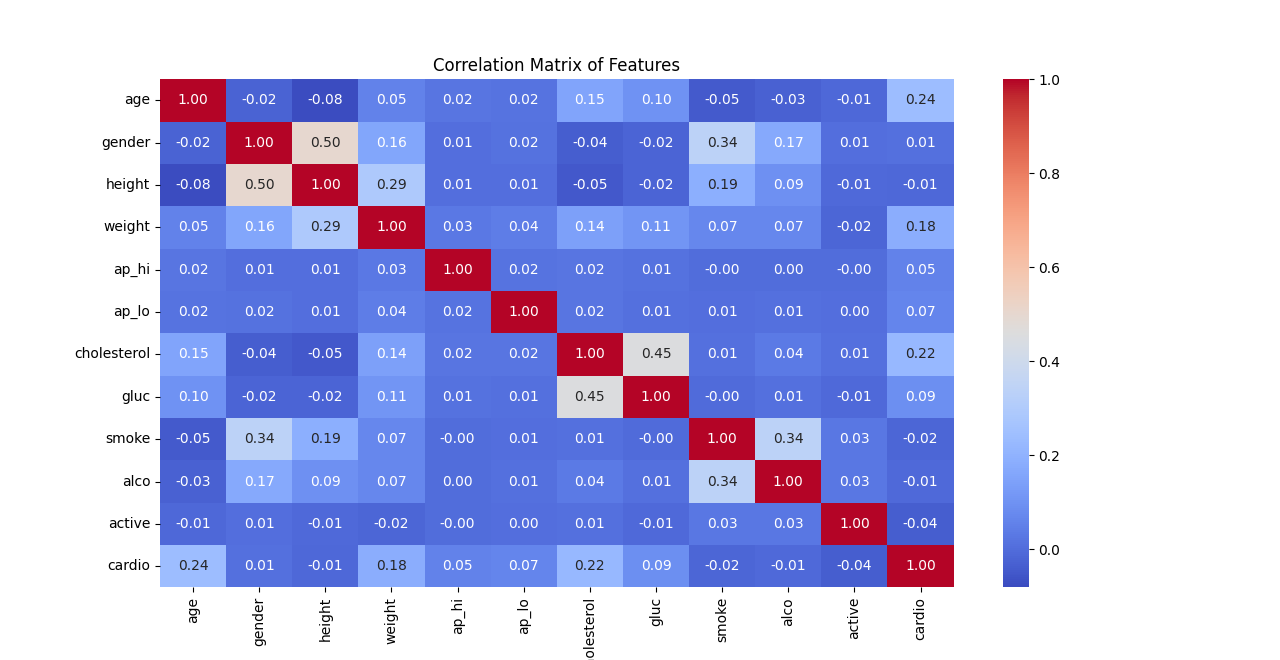
* **Accuracy**: 0.70
* Decision Trees were easy to interpret but prone to overfitting, limiting their generalizability.

**5. Random Forest (RF)**

* **Accuracy**: 0.80
* Random Forest outperformed all other models by aggregating predictions from multiple decision trees, reducing overfitting and improving accuracy.

**OUTPUT IMAGES:**

****

****

**FINAL MODEL**

The **Random Forest Classifier** emerged as the optimal choice for heart disease detection due to its superior accuracy and robustness. Key performance metrics included:

* **Accuracy**: 80%
* **Precision, Recall, F1-Score**: Balanced results across all metrics, confirming its reliability and consistency in predictions.

Additionally, the Random Forest model provided feature importance rankings, highlighting cholesterol, glucose levels, and blood pressure as the most critical predictors.

**CONCLUSION**

This project demonstrates the potential of machine learning in advancing heart disease detection. By integrating data preprocessing, visualization, and algorithm evaluation, the Random Forest Classifier was identified as the most effective model. This approach supports healthcare professionals in making informed decisions, paving the way for early diagnosis and personalized treatment plans.

**FUTURE WORK**

* Incorporate additional features, such as family medical history and dietary habits, to enhance prediction accuracy.
* Explore advanced machine learning techniques, including deep learning, for more nuanced analysis of complex datasets.
* Develop an interactive web or mobile application to deploy the model for real-world, real-time use, bridging the gap between research and practical healthcare applications.
* Perform longitudinal studies to evaluate model performance over time and across diverse populations.