**Heart Disease Prediction Using Machine Learning**

**ABSTRACT**  
This project aims to predict the likelihood of heart disease by analyzing key health indicators through machine learning techniques. The dataset contains attributes such as age, gender, cholesterol, and blood pressure. The workflow involves cleaning and preprocessing the data, generating visual insights, analyzing feature relationships, and testing multiple machine learning algorithms to determine the most effective predictive model. By leveraging these techniques, this study aims to enhance diagnostic accuracy and support healthcare professionals in early identification of heart disease.

**INTRODUCTION**  
Heart disease is one of the leading causes of death worldwide, claiming millions of lives annually. Detecting potential cases early can drastically improve outcomes and reduce mortality rates. Machine learning has emerged as a powerful tool for analyzing complex health data, offering new opportunities for early diagnosis and treatment. This project utilizes machine learning algorithms to predict heart disease by examining critical health parameters. By systematically comparing different approaches, the goal is to identify the most reliable and accurate model, enabling data-driven decision-making in healthcare.

**DATASET OVERVIEW**

* **Source**: The dataset consists of anonymized clinical records from a diverse group of patients, providing essential health metrics for heart disease risk analysis.
  + **Attributes:**
    - **age**: Patient age, converted to years for better readability.
    - **gender**: Gender of the patient (1 for Male, 2 for Female).
    - **height**: Patient height (in cm), used for BMI calculations.
    - **weight**: Patient weight (in kg), relevant for BMI and health assessments.
    - **ap\_hi**: Systolic blood pressure, a critical cardiovascular indicator.
    - **ap\_lo**: Diastolic blood pressure, complementary to systolic readings.
    - **cholesterol**: Cholesterol levels categorized as Normal (1), Above Normal (2), or High (3).
    - **gluc**: Blood glucose levels (1: Normal, 2: Elevated, 3: High).
    - **smoke**: Smoking status (0: Non-smoker, 1: Smoker), a significant risk factor.
    - **alco**: Alcohol consumption (0: No, 1: Yes), linked to cardiovascular health.
    - **active**: Physical activity level (0: Inactive, 1: Active).
    - **cardio**: Target variable (0: No heart disease, 1: Heart disease).
* **Dataset Size**: 70,000 entries, providing a large and balanced dataset for reliable training and testing.
* **Preprocessing Steps:**
  + Dropped irrelevant columns such as patient ID.
  + Verified the dataset for missing or inconsistent values.
  + Scaled numerical features using StandardScaler to ensure uniformity and improve model performance.

**DATA ANALYSIS AND VISUALIZATION**

1. **Feature Distributions**:
   * Histograms and density plots were created for each feature to assess data distributions. Outliers were particularly noted in blood pressure and weight metrics.
2. **Correlation Analysis**:
   * A heatmap of the correlation matrix revealed significant relationships between features and the target variable. Key observations included:
     + Positive correlation between high cholesterol and glucose levels and the presence of heart disease.
     + Negative correlation between physical activity and the likelihood of heart disease.
3. **Insights Gained**:
   * Patients with higher cholesterol and glucose levels exhibit a higher risk of heart disease.
   * Smoking and inactivity are significant contributors to cardiovascular issues.
   * Blood pressure metrics (both systolic and diastolic) are critical indicators of heart disease risk.

**MACHINE LEARNING MODELS AND PERFORMANCE EVALUATION** The dataset was split into 70% training and 30% testing subsets to evaluate the performance of various models. Features were standardized to improve model consistency. The following algorithms were implemented:

1. **Logistic Regression (LR):**
   * **Accuracy**: 74%
   * Provided a straightforward baseline with good results for linearly separable data.
2. **Support Vector Machine (SVM):**
   * **Accuracy**: 77%
   * The linear kernel produced improved accuracy but required longer computation times.
3. **K-Nearest Neighbors (KNN):**
   * **Accuracy**: 72%
   * Sensitive to dataset size and the choice of k, with declining performance on larger datasets.
4. **Decision Tree (DT):**
   * **Accuracy**: 70%
   * While interpretable, decision trees showed overfitting tendencies.
5. **Random Forest (RF):**
   * **Accuracy**: 80%
   * By combining predictions from multiple decision trees, Random Forest achieved the highest accuracy and generalization ability.

**RESULTS AND ANALYSIS** The Random Forest Classifier was identified as the optimal model based on its superior accuracy and reliability. Feature importance rankings revealed that cholesterol levels, glucose levels, and blood pressure were the most influential predictors. The model’s balanced precision, recall, and F1-score further reinforced its robustness and suitability for heart disease prediction.

**CONCLUSION** This project highlights the potential of machine learning in improving heart disease detection. By leveraging advanced data analysis techniques and systematically evaluating multiple algorithms, Random Forest emerged as the most effective model. This predictive framework can assist healthcare providers in identifying at-risk individuals and devising early intervention strategies, ultimately improving patient outcomes.

**FUTURE DIRECTIONS**

* Incorporate additional variables, such as family medical history and dietary patterns, to enhance prediction capabilities.
* Investigate deep learning approaches to handle complex and non-linear relationships in data.
* Develop user-friendly applications, such as web or mobile platforms, to integrate the model into real-world healthcare workflows.
* Conduct validation studies across diverse demographics and timeframes to ensure generalizability and long-term reliability.