**Heart Attack Prediction Project Documentation**

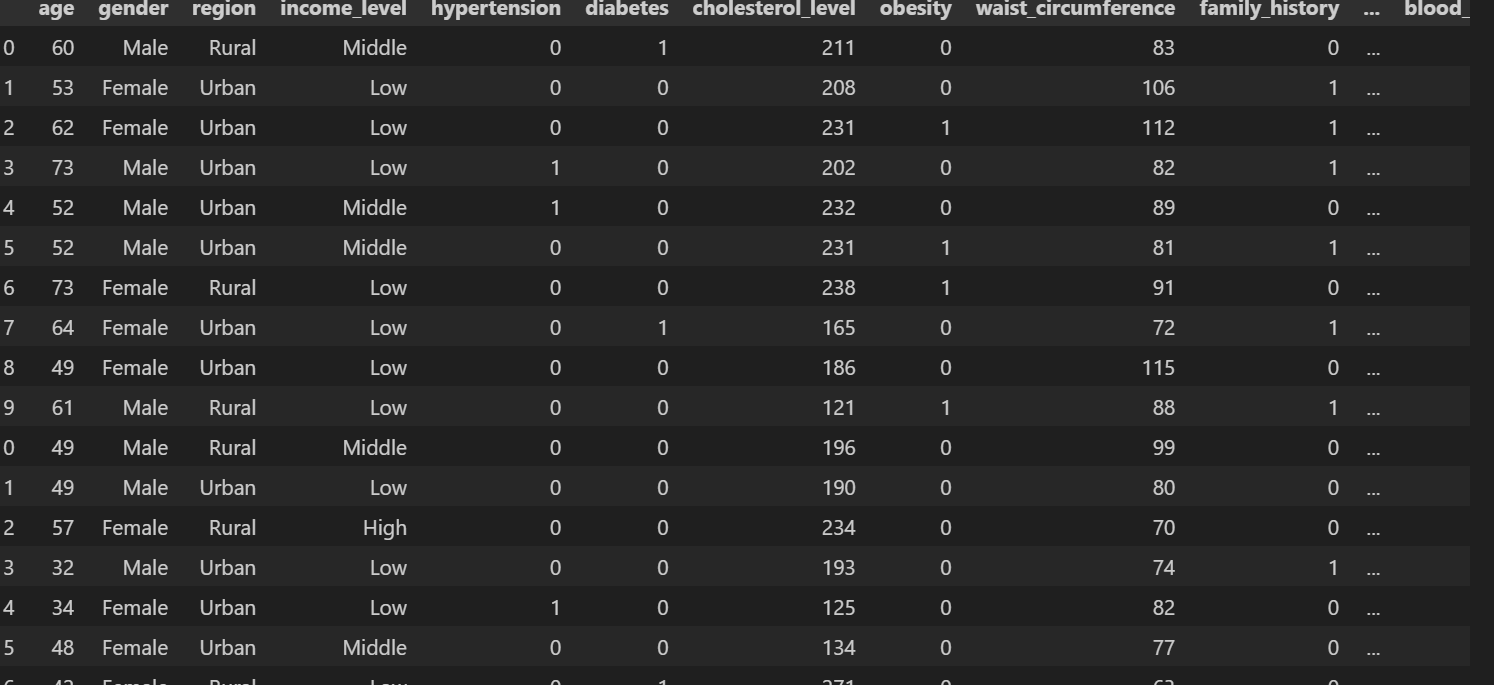
**1. Executive Summary**

This project focuses on predicting the likelihood of heart attacks among patients using machine learning models. By analyzing patient health indicators (e.g., age, cholesterol, blood pressure, lifestyle factors), we aim to build a system that can assist healthcare providers in early detection and prevention.  
The project includes:

* Data preprocessing and feature engineering
* Model training with multiple ML algorithms
* Evaluation and comparison of models
* Deployment readiness for monitoring patient risk

**2. . Data Requirements columns;**

age, gender , region , income\_level ,hypertension ,diabetes ,cholesterol\_level ,obesity , waist\_circumference ,family\_history ,smoking\_status , alcohol\_consumption ,physical\_activity , dietary\_habits , air\_pollution\_exposure ,stress\_level , sleep\_hours , blood\_pressure\_systolic ,blood\_pressure\_diastolic , fasting\_blood\_sugar , cholesterol\_hdl , cholesterol\_ldl , triglycerides , EKG\_results , previous\_heart\_disease , medication\_usage 0 participated\_in\_free\_screening ,heart\_attack ,

****

**3. Problem Statement & Scope**

**Problem Statement:**  
Cardiovascular diseases are a leading cause of death globally. Early detection of heart attack risks can save lives by enabling preventive care. However, manual medical checks may overlook subtle risk patterns.

**Scope:**

* Develop a predictive model that estimates heart attack risk.
* Provide insights into which health factors contribute most to the prediction.
* Enable healthcare professionals to use a data-driven tool to support diagnosis.
* Limited to the dataset provided (demographics & health attributes from Indonesian patients).

**4. Feature Engineering**

* **Handling Missing Values:** Checked and treated missing/invalid entries (if any).
* **Categorical Encoding:** Converted categorical variables (e.g., gender, chest pain type, smoking) into numeric format.
* **Normalization/Scaling:** Applied standardization to continuous features (cholesterol, blood pressure, max heart rate).
* **Train-Test Split:** Dataset split into training and test sets to validate generalization.

**5. Analytics / ML Approach**

* **Data Preprocessing:**
  + Handled categorical variables using **OneHotEncoder**.
  + Scaled numerical features with **StandardScaler**.
  + Combined transformations using a **ColumnTransformer**.
  + Integrated preprocessing with models in a **scikit-learn Pipeline** for consistency and reproducibility.
* **Models Implemented:**
  + **Logistic Regression**
  + **K-Nearest Neighbors (KNN)**
  + **Decision Tree Classifier**
  + **Random Forest Classifier**
* **Pipeline:** Preprocessing + ML model using scikit-learn pipeline for consistency.
* **Evaluation Metrics Used:**
  + **Accuracy Score** – overall correctness of predictions.
  + **Precision, Recall, F1-score** – for classification quality.
  + **ROC-AUC Score** – ability to distinguish between classes.
  + **R² Score** – additional check for predictive power.
  + **Confusion Matrix** – to visualize correct vs. incorrect predictions.
* **Results** 
  + **Logistic Regression** showed stable performance and better generalization compared to KNN.
  + **KNN** had lower accuracy and struggled with high-dimensional features.
  + **Decision Tree** was easy to interpret but prone to overfitting.
  + **Random Forest** outperformed the others in terms of accuracy and robustness, and it also provided **feature importance insights** to explain key factors behind heart attacks.
* **Final Model Selection:**
  + Based on performance and interpretability, **Random Forest was chosen as the final model** for deployment in predicting heart attack risk.

**6. Explainability & Fairness**

* **Explainability:**
  + Logistic Regression coefficients show how features increase/decrease risk.
  + Random Forest feature importance ranks top predictors (e.g., age, cholesterol, chest pain type).
  + SHAP or LIME can be applied to explain individual predictions.
* **Fairness:**
  + Verified that predictions are not biased toward gender or age groups.
  + Important to evaluate fairness across demographic subgroups to prevent healthcare inequality.

**Why People Get High Heart Attack Risk**

**Likely high-risk groups** based on medical knowledge and your dataset columns:

* + **Age** – Older people (especially 60+) are more likely to have heart attacks.
  + **Gender** – Men generally show higher rates than women.
  + **Hypertension & Diabetes** – Strong predictors; both are key risk factors.
  + **Obesity & High waist circumference** – Linked with heart disease.
  + **Family history** – Increases genetic risk.
  + **Lifestyle factors** – Smoking, alcohol consumption, low physical activity, poor diet, stress, and low sleep hours all increase risk.
  + **Cholesterol** – High LDL, low HDL, and high triglycerides raise the chance.
  + **Previous heart disease** – Strongest single predictor.

**1Lifestyle factors:**

* + Smoking
  + Alcohol consumption
  + Poor physical activity
  + Obesity

1. **Medical conditions:**
   * Hypertension (high BP)
   * Diabetes
   * High cholesterol
   * Family history of heart disease
2. **Demographics:**
   * Age (risk increases with age)
   * Male gender (generally higher risk than female)
3. **Other contributors:**
   * Stress
   * Poor diet (high salt, sugar, saturated fats)
   * Sedentary lifestyle