# FINAL PROJECT DOCUMENTATION

Cardiovascular Disease Risk Prediction

s

## Table of Contents

# 1. Project Overview

**Objective**: Develop a predictive model to assess cardiovascular disease (CVD) risk using clinical, demographic, and lifestyle data. The project encompasses data analysis, feature engineering, model training, and deployment of a Streamlit application for real-time risk prediction.

**Dataset**:

- **Source**: Processed dataset (cardio_data_processed.csv) with 68,205 records and 14+ features.
- **Key Variables**:
    - *Demographics*: Age, gender, height, weight.
    - *Clinical Metrics*: Blood pressure (ap_hi, ap_lo), cholesterol, glucose.
    - *Lifestyle*: Smoking, alcohol consumption, physical activity.
    - *Target*: cardio (binary indicator of CVD).

**Tools**: Python (pandas, NumPy, scikit-learn, XGBoost), Streamlit, Matplotlib, Seaborn.

## 2. Data Preparation and Cleaning

**Steps**:

1. **Data Loading**:
    - Imported tab-separated data from cleaned_data.csv.
    - Ensured column consistency (e.g., split merged fields).
2. **Outlier Handling**:
    - **Height**: Replaced values <120 cm or >210 cm with median (~165 cm).
    - **Weight**: Dropped non-convertible entries; replaced values <30 kg or >300 kg with median (~75 kg).
    - **Blood Pressure**: Removed records with ap_hi <50 or ap_hi < ap_lo.
3. **Data Type Conversion**:
    - Numerical features cast to int/float; categorical variables (e.g., smoke, alcohol) converted to category.
4. **Feature Extraction**:
    - Derived age_years (converted from days).
    - Calculated pulse_pressure (ap_hi - ap_lo).

# 3. Exploratory Data Analysis (EDA)

**Key Analyses**:

1. **Univariate Analysis**:
    - **Numerical Features**: Skewness and outliers observed in age_years, bmi, and blood pressure.
    - **Categorical Features**:
        - Most individuals were non-smokers (70%), physically active (65%), and had normal cholesterol/glucose levels.
        - Class imbalance: 53% male, 47% female.
2. **Correlation Analysis**:
    - **Strong Positive Correlations**: Systolic vs. diastolic blood pressure (r >0.7).
    - **Moderate Correlations**: BMI with weight (r = 0.8), cardio with age, blood pressure, and cholesterol.
3. **Multivariate Analysis**:
    - Higher CVD prevalence in hypertensive, obese, and older individuals.
    - Lifestyle factors (smoking, alcohol, inactivity) showed weaker correlations but contributed to risk accumulation.

# 4. Feature Engineering

**Engineered Features**:

1. **BMI Categories**:
    - Binned into *Underweight*, *Normal*, *Overweight*, *Obese* (BMI ≥30).
    - Enabled non-linear risk assessment (e.g., obese individuals had 2x higher CVD risk).
2. **Lifestyle Score**:
    - Composite score (0–3) summing smoking, alcohol use, and inactivity.
    - Higher scores correlated with elevated CVD risk.
3. **Blood Pressure Categories**:
    - Classified into *Normal*, *Elevated*, *Hypertension Stage 1/2*.
    - Hypertensive categories had 50% higher CVD incidence.

**Impact**: Improved model interpretability and predictive power by capturing non-linear relationships and clinical thresholds.

# 5. Model Development

| Model | Test Accuracy | Precision | Recall | F1-Score | Overfitting Risk |
|---|---|---|---|---|---|
| **XGBoost** | **73.67%** | **0.74** | **0.74** | **0.74** | Low |
| **Random Forest** | 73.44% | 0.74 | 0.73 | 0.73 | Mild |
| **Logistic Regression** | 72.68% | 0.73 | 0.73 | 0.72 | Low |
| **SVM** | 72.62% | 0.73 | 0.73 | 0.72 | Low |
| **KNN** | 70.35% | 0.70 | 0.70 | 0.70 | **High** |

**Key Insights**:

- **Top Features**: Blood pressure, BMI, cholesterol, glucose, and age dominated predictions.
- **Excluded Features**: Demographics (gender) and lifestyle factors (smoking, alcohol) had lower importance.
- **Class Imbalance**: Models favored predicting "No CVD" (higher recall for Class 0).

**Best Model (XGBoost)**:

- **Hyperparameters**: n_estimators=200, max_depth=3, learning_rate=0.1.
- **Confusion Matrix**:
    - **True Positives**: 72% (CVD cases correctly identified).
    - **False Negatives**: 14% (high-risk individuals missed).

# 6. Deployment

**Streamlit Application**:

1. **Interface**:
    - **Predict Risk Tab**: Users input demographics, health metrics, and lifestyle factors.
    - **Dashboard Tab**: Interactive visualizations (univariate, bivariate, multivariate).
2. **Prediction Workflow**:
    - Real-time calculation of BMI and pulse pressure.
    - Preprocessing, feature selection, and XGBoost prediction.
    - Output: "High Risk" or "Low Risk" with probability.
3. **Technical Setup**:
    - Caching for efficient model/data loading.
    - Error handling for user inputs.

# 7. Key Insights and Conclusions

1. **Risk Factors**:
    - **Clinical**: Age, hypertension, high cholesterol, obesity.
    - **Lifestyle**: Physical inactivity, smoking, alcohol use (cumulative effect).
2. **Model Limitations**:
    - Moderate accuracy (73.67%) due to class imbalance and complex interactions.
    - Limited generalizability to underrepresented demographics.
3. **Future Work**:
    - Collect more data on lifestyle factors.
    - Experiment with deep learning architectures.

**Impact**: The deployed application enables individuals to assess CVD risk proactively, emphasizing preventive care.