Analysis of factors for predicting the 10-year risk of heart disease to support prevention.

DS512/513 Data Analytics
DS514/515 Data Science

Narawit Tharapoompipat 68199160275
Ratanaporn Thaochalee 68199160293
Sakdhithad Chanfeungfu 68199160301

[14th December 2025]

DATA PROJECT CANVAS

Designed by: Narawit, Ratanaporn, Sakdhithad **Date:** 2025-12-14

Title:

1. Problem Statement/Background

- CVD is the Leading Cause of Death globally.
- Accounts for 32% of all deaths (19.8 million annually).
- 85% of CVD deaths are Heart Attacks & Strokes, directly matching our Ten-Year CHD target.
- As WHO confirms CVDs are preventable, data analysis is used for early.

2. Questions/Hypothesis

How do age, gender, and BMI trend to impact CHD risk, and is glucose the dominant factor over cholesterol, considering its link to blood pressure and the cumulative danger of multiple risk factors?

Predict the 10 years CHD patient based on given demographic, behavioral and medical data.

3. Value Propositions

Launch a campaign to reduce the risk level of participants by 5% within 3 months.

4. Data Sources/Attributes

- Data sources & collection
- Data cleaning & preprocessing

Primary Source: Framingham Heart Study Dataset.(kaggle)
Data Volume: 4,238 Patient records with 16 Attributes.

Target: 10-year-CHD risk
Feature: 2 demographic, 2 behavioral, 10 medical features
Scaling strategies : RobustScaler and MinmaxScaler
Imbalanced class handling: class_weight, SMOTE, Undersampling

5. Analysis/Model Development

- Analytics Methodology
  - Descriptive statistics and pivot tables by **Excel**
  - EDA and visualization by **Tableau**

- Modeling Methodology
  - LogisticRegression including ElasticNet
  - KNeighborClassifier
  - RidgeClassifier

Evaluation metric

- Accuracy, Precision, Recall (primary) and F1 score

6. Findings and Insights

- Age is the primary driver of risk for everyone, regardless of gender.
- High glucose is the dominant factor more than cholesterol and is linked to higher blood pressure.
- The Overweight category represents the highest volume of at-risk cases, surpassing the Obese group.

Model performance: LogisticRegression with class re-weighing provided the best recall (0.89), but low precision.

- Sex is the most feature important, consistent with medical literature showing higher cardiovascular disease rates in men.
- Patients on blood pressure medication show substantially elevated CHD risk. This likely indicates underlying hypertension management and pre-existing cardiovascular conditions.

7. Recommendation/Action and Impact

Targeting heart disease as the leading preventable cause of death, this study demonstrates that knowing specific data is effective for prevention.

The Multiplier Effect: Combined risk factors make the danger much higher, requiring us to treat the whole picture instead of just one problem.

Try Advanced Models: Gradient boosting or tree-based models may handle this problem better than linear approaches.
- **Balance the Dataset**: Collect more minority class samples to improve data distribution and model performance.

Modified from Bill Schmarzo's Machine Learning Canvas and Jasmine Vasandani's Data Science Workflow Canvas for SWU MSDS

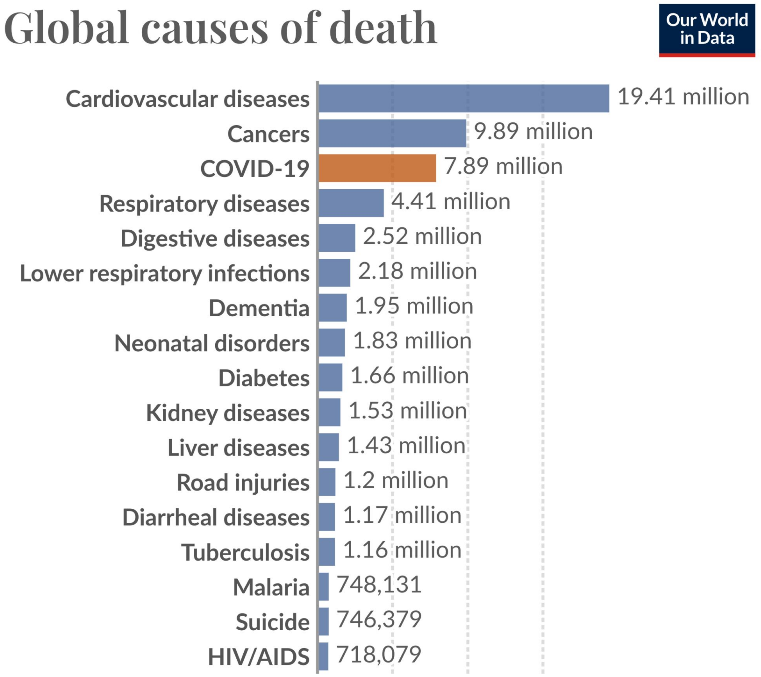
Heart Disease

What is heart disease ?

Heart disease is a broad term for a range of conditions that affect your heart. It is also often called cardiovascular disease**,** which generally refers to conditions that involve narrowed or blocked blood vessels, leading to a risk of heart attack, chest pain (Angina Pectoris), or stroke.

Why Interest ?

The Global Health Crisis



Data source: IHME, Global Burden of Disease (2024)
OurWorldInData.org/causes-of-death | CC BY

- CVD is the Leading Cause of Death globally.
- Nearly 1 in 3 global deaths are from CVD.
- 19 million annually.

https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29

Direct Relevance & Data Validation

- Our initial analysis confirms these global concerns.
- We observed the impact of risk factors (Age, Smoking, BP) on our target: TenYearCHD.
- This validates our dataset as a powerful tool for this study.

The Goal: Insight for Early Awareness

- Identify concrete "Risk Indicators"
- Analyze the combined impact of factors (e.g., BP + Cholesterol + Age).
- Provide insights for early awareness and prevention—detecting risks *before* they become critical.

Project Objective

- **Analyze Risk Factors:** Identify key demographic, behavioral, and clinical drivers (e.g., age, smoking, glucose) of 10-year CHD risk using the Framingham dataset.

- **Develop Guidelines:** Create data-driven preventive guidelines to promote healthier lifestyle behaviors (diet, activity, BP control).

- **Measurable Goal**: Reduce participants' modifiable risk indicators by at least 5% within 3 months.

- **Visualize Data Insights:** Create an interactive dashboard to clearly communicate risk patterns.

- **Build Prediction Model**: Develop a machine learning model using hyperparameter tuning and imbalance handling to predict 10-year CHD risk.

Data Dictionary

Data Dictionary Overview

Target Variable

- TenYearCHD: 10-year risk of coronary heart disease (1 = Risk, 0 = No Risk)

Demographic

- age, sex (Male/Female), education

Behavioral

- current Smoker, cigs Per Day (cigarettes/day)

Medical History

- BP Meds (Blood pressure meds), prevalent Stroke, prevalent Hyp (Hypertension), diabetes

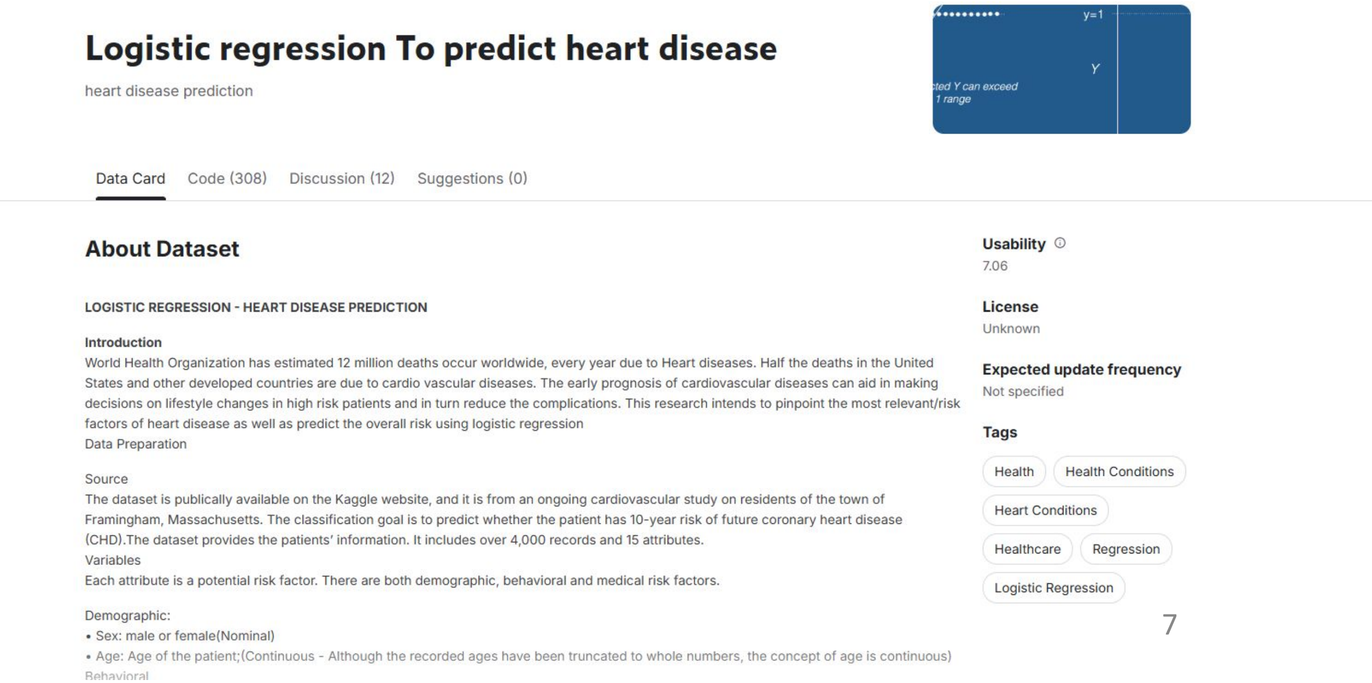
Current Health Stats

- totChol (Cholesterol), sysBP, diaBP, BMI, heart Rate, glucose

Data Collection

Primary Data Source

- Primary Source: Framingham Heart Study Dataset including our target TenYearCHD.
- www.kaggle.com%2Fdatasets%2Fdileep070%2Fheart-disease-prediction-using-logistic-regression%2Fdata



Logistic regression To predict heart disease

heart disease prediction

About Dataset

LOGISTIC REGRESSION - HEART DISEASE PREDICTION

Introduction

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression

Data Preparation

Source

The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

Variables

Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

Demographic:

- Sex: male or female(Nominal)
- Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioral

Usability

7.06

License

Unknown

Expected update frequency

Not specified

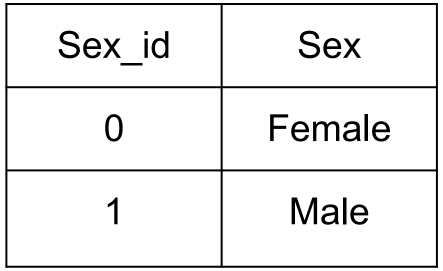
Tags

Health · Health Conditions · Heart Conditions · Healthcare · Regression · Logistic Regression

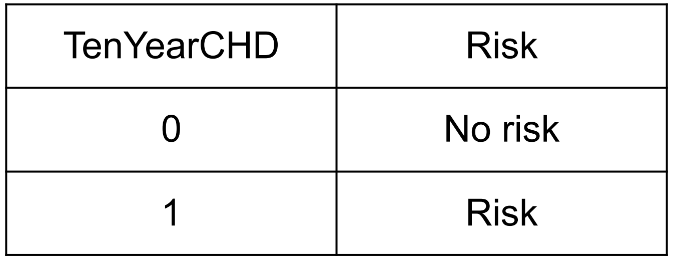
Data Cleaning and Preprocessing

Data Integration: Merged lookup tables to translate unclear numerical codes (e.g., 0, 1) into human-readable labels like Male/Female ,and No Risk/Risk.

Handling Issues: Dropped missing values (NaNs).



Sex_id	Sex
0	Female
1	Male



TenYearCHD	Risk
0	No risk
1	Risk

Understanding Data

- **Data Inspection** Confirmed 4,238 records, 16 attributes, and identified TenYearCHD as the target.

- **Problems Identified** : Identified three critical problems: **Missing Values**, **Extreme Outliers**, and **an Imbalanced Target.**

Setting Questions/ Hypothesis

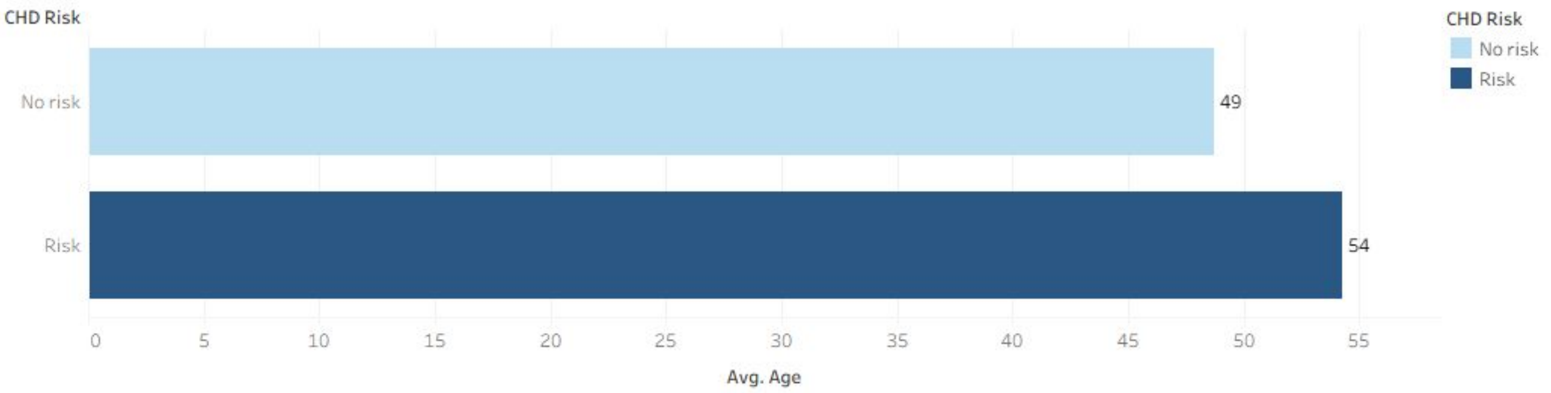
General Information

- Does age have a significant impact on ten-year CHD risk?
- Age is a risk factor. But does the impact of age on Ten-Year CHD risk different between genders?

Medical Data Factors

- Which is the Dominant Risk Factor: Glucose or Cholesterol ?
- Is elevated blood glucose associated with higher systolic blood pressure (sysBP)?
- Which BMI Category Has the Highest Number of Cases?

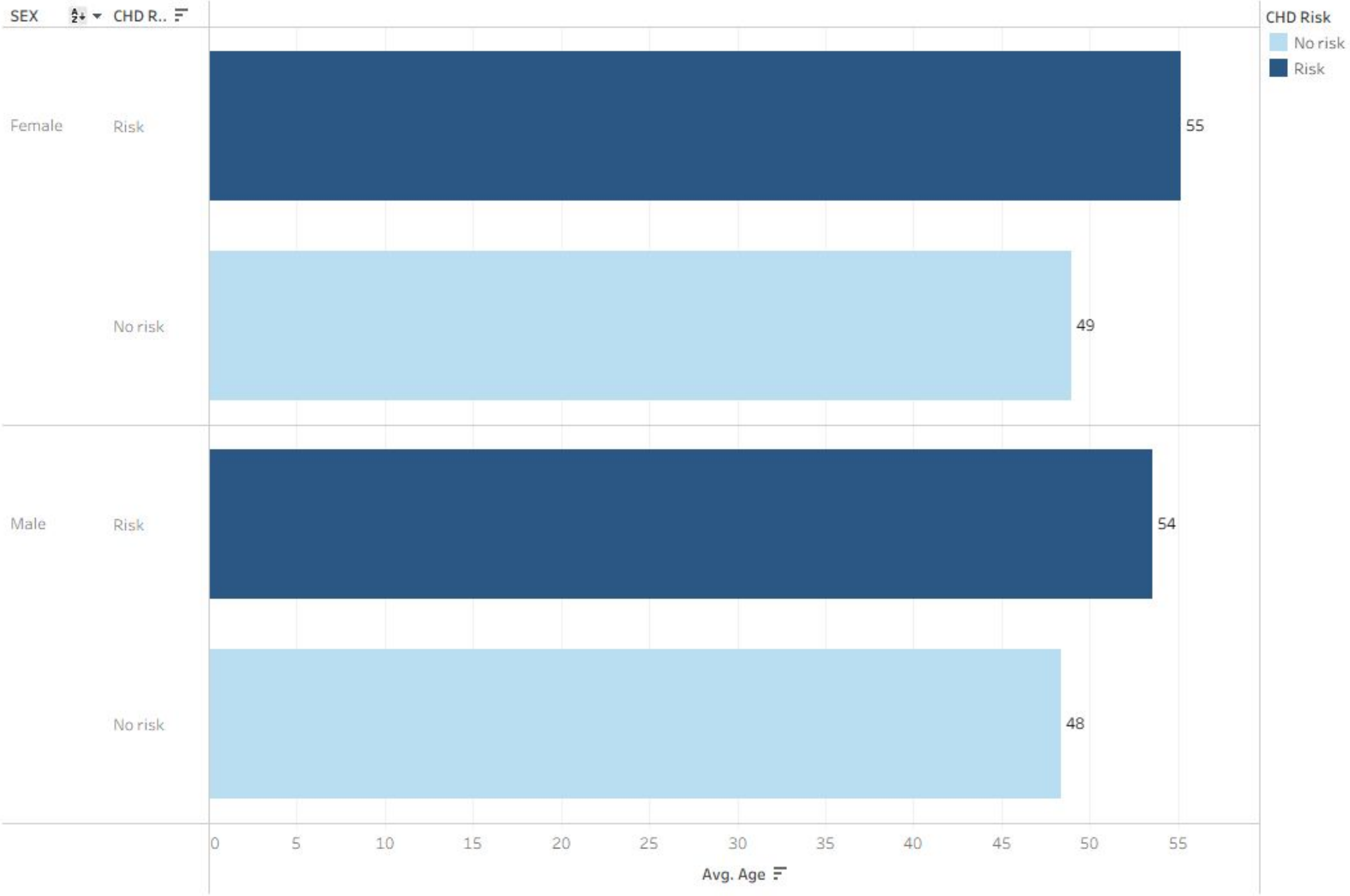
Findings and Insights



- **Does age have a significant impact on ten-year CHD risk?**

Consistent Age Impact: Across both genders, the 'Risk' group is consistently older than the 'No Risk' group.

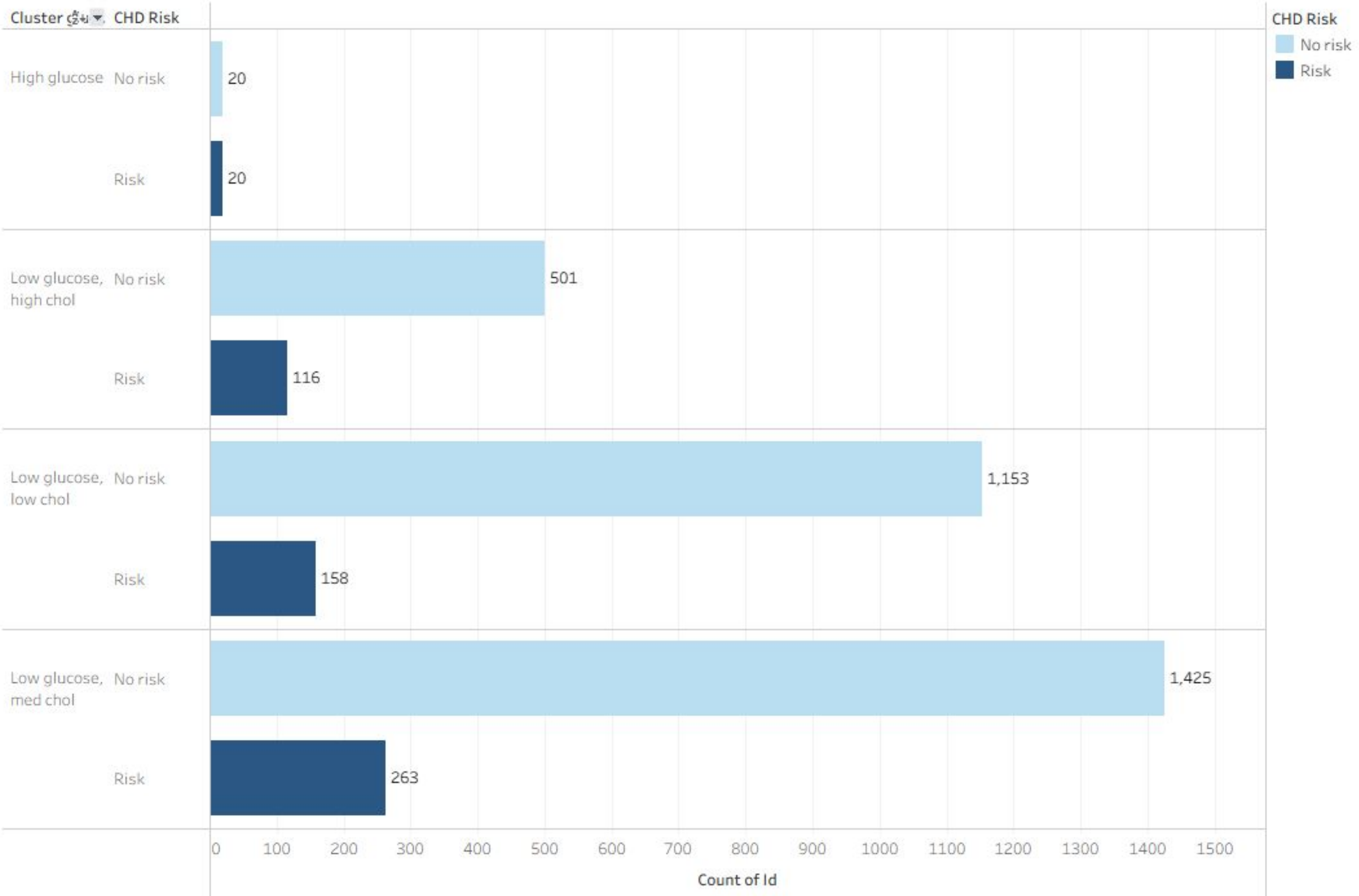
Findings and Insights



- **Age is a risk factor**. But is the impact of age on ten-year CHD risk different between genders?

Gender shows minimal impact on heart disease risk, whereas age proves to be the dominant factor for both groups.

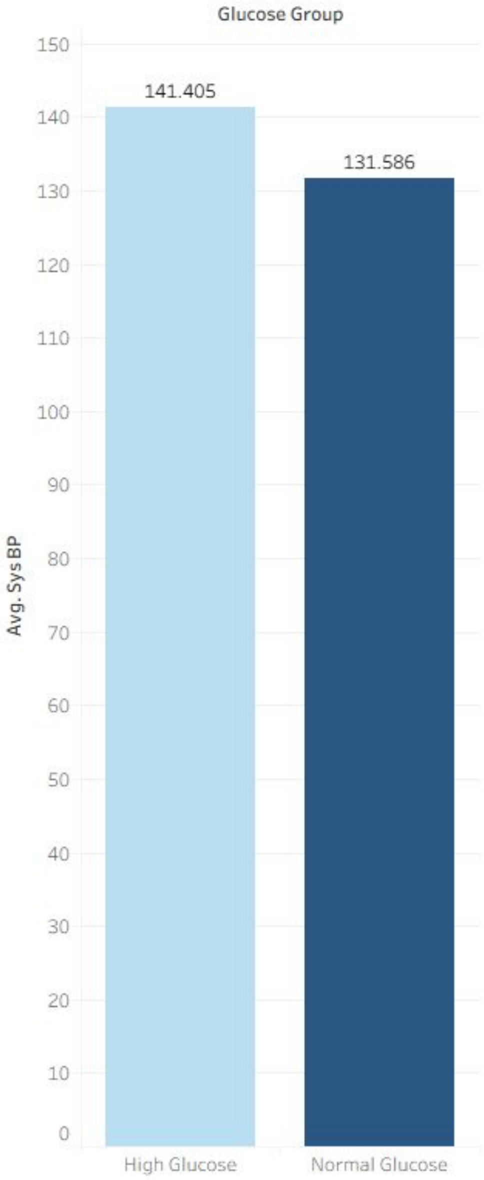
Findings and Insights



- **Which is the Dominant Risk Factor: Glucose or Cholesterol ?**

High glucose is identified as the dominant risk factor, significantly outweighing the impact of high cholesterol

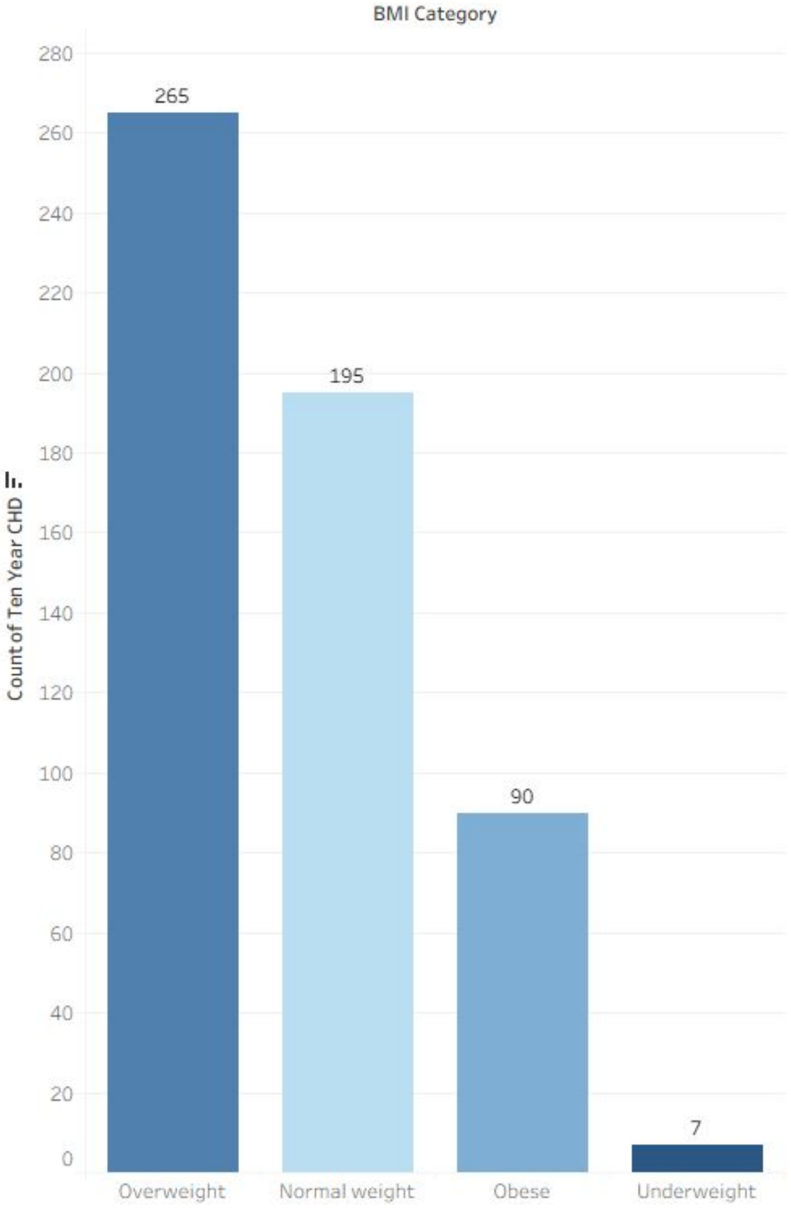
Findings and Insights



- **Is elevated blood glucose associated with higher systolic blood pressure (sysBP)?**

Elevated blood glucose is associated with higher systolic blood pressure 141.4 compare to 131.6 mmHg, confirming a link to increased cardiovascular stress

Findings and Insights



- **Which BMI Category Has the Highest Number of Cases?**

The Overweight category represents the highest volume of at-risk patients with 265 cases, significantly surpassing both the Normal weight 195 and Obese 9 groups, identifying it as the primary demographic for intervention.

This standard is based on the World Health Organization (WHO), a globally recognized medical benchmark.

BMI < 18.5 indicates underweight

BMI 18.5–24.9 indicates normal weight

BMI ≥ 25.0 indicates overweight

BMI ≥ 30.0 indicates obesity

https://apps.who.int/nutrition/landscape/help.aspx?menu=0&helpid=420

Findings and Insights

- **Does age have a significant impact on ten-year CHD risk?**

  Consistent Age Impact: Across both genders, the 'Risk' group is consistently older than the 'No Risk' group.

- **Age is a risk factor. But is the impact of age on ten-year CHD risk different between genders?**

  Gender shows minimal impact on heart disease risk, whereas age proves to be the dominant factor for both groups.

- **Which is the Dominant Risk Factor: Glucose or Cholesterol ?**

  High glucose is identified as the dominant risk factor, significantly outweighing the impact of high cholesterol.

- **Is elevated blood glucose associated with higher systolic blood pressure (sysBP)?**

  Elevated blood glucose is associated with higher blood pressure and increased cardiovascular stress.

- **Which BMI Category Has the Highest Number of Cases?**

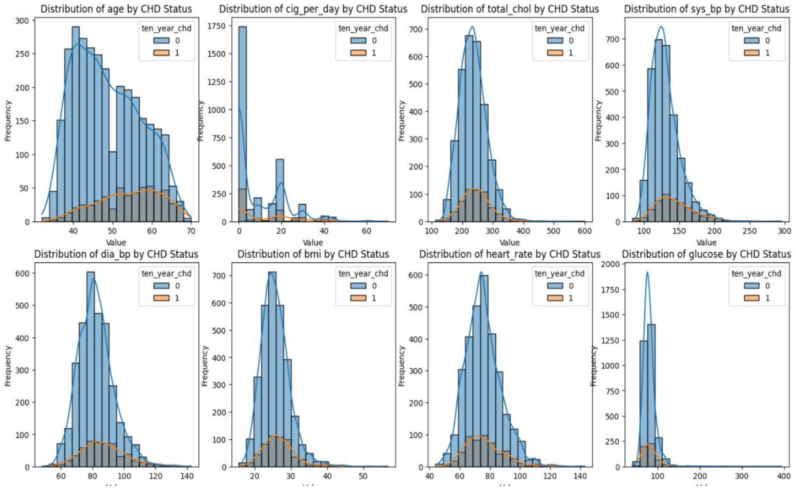
  The Overweight category accounts for the highest volume of at-risk cases, surpassing both the Normal weight and Obese groups.

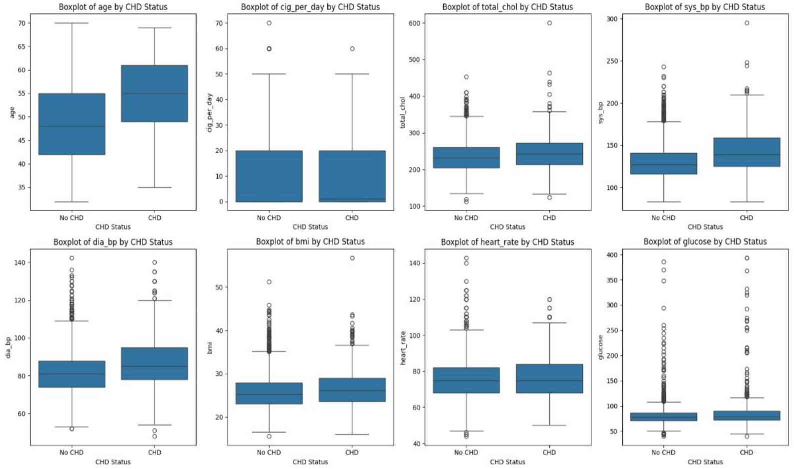
Modelling Approach

Classification problem

~~15~~ 13 Features to train model

Cat_feature	Num_feature
sex	tot_chol
prevalent_storke	sys_bp
bp_meds	dia_bp
prevalent_hyp	heart_rate
diabetes	age
~~current_smoke~~	cig_per_day
~~education_level~~	bmi
	glucose

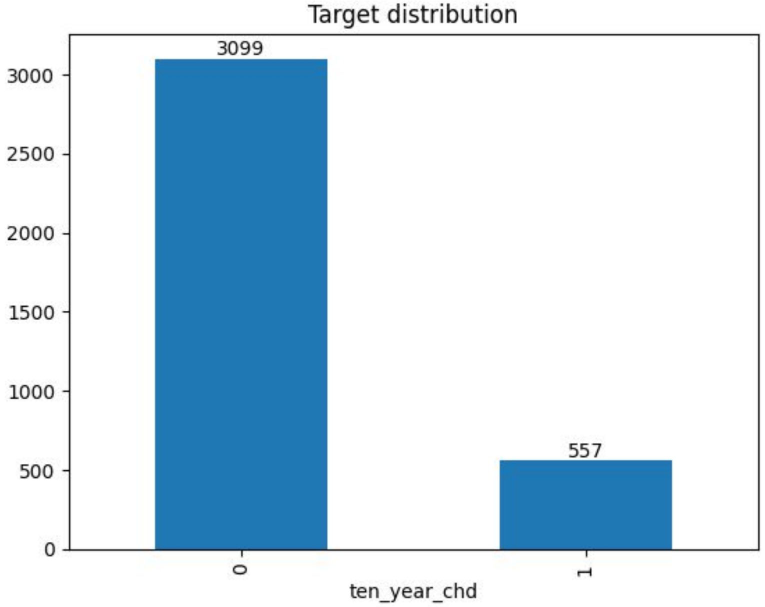




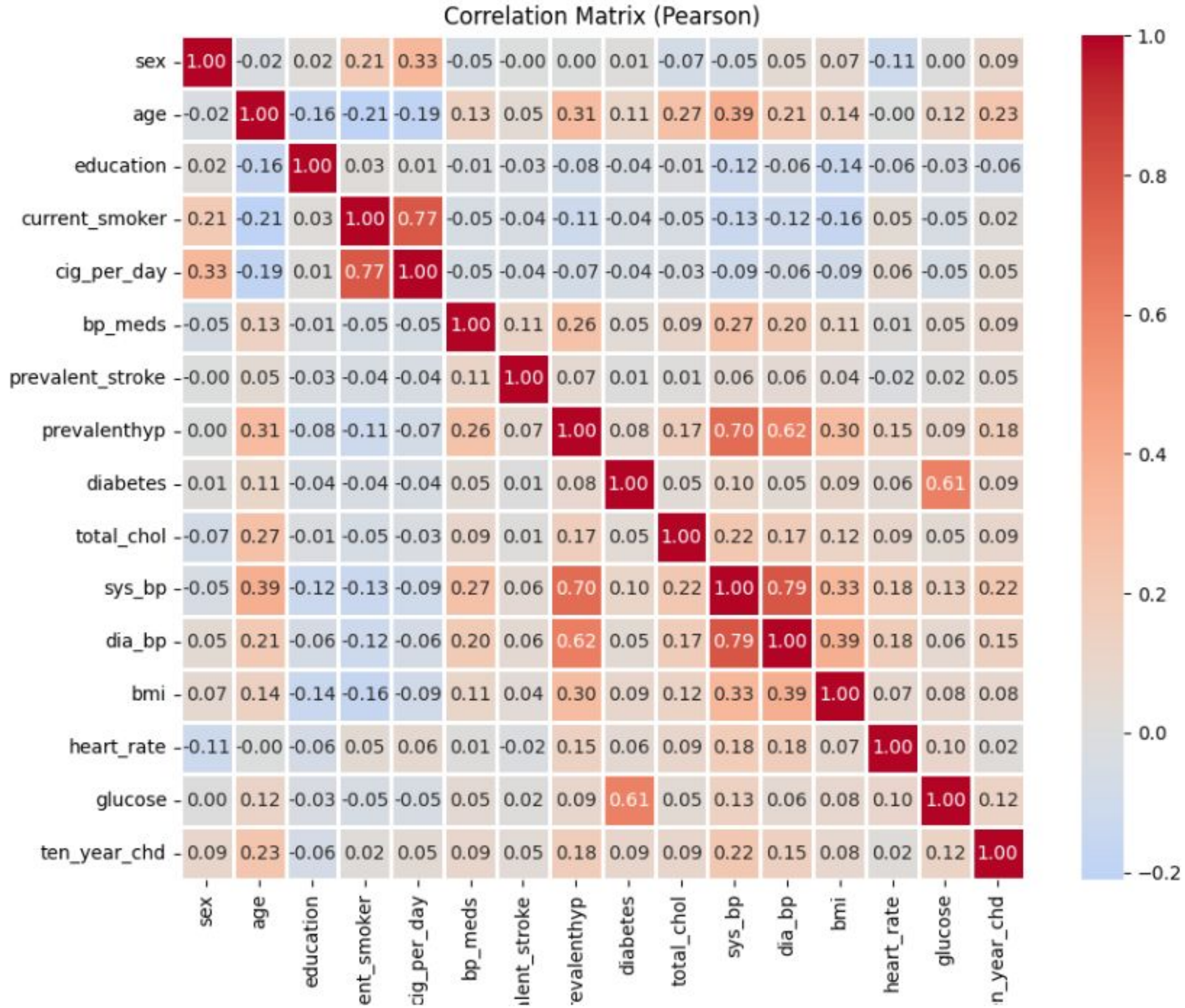
Predict

Target class

ten_year_chd
0,1



Pearson correlation

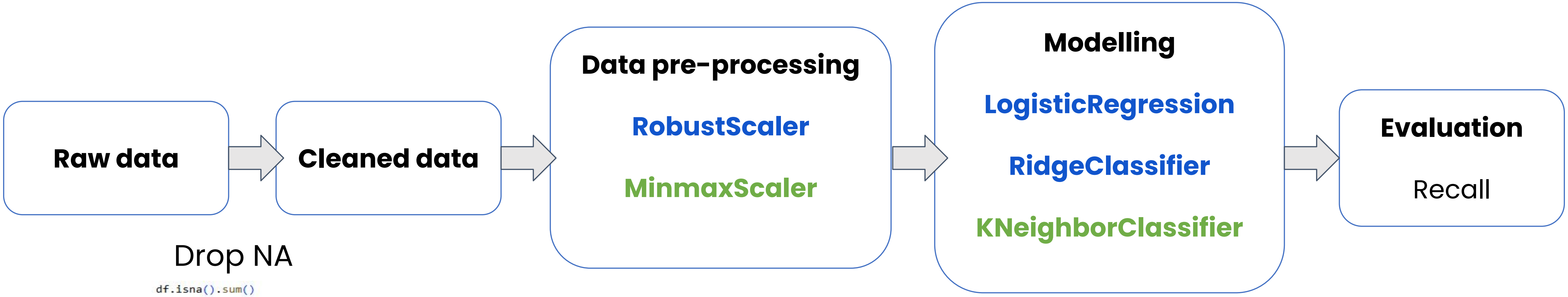


Correlation Matrix (Pearson)

	sex	age	education	ent_smoker	cig_per_day	bp_meds	lent_stroke	evalenthyp	diabetes	total_chol	sys_bp	dia_bp	bmi	heart_rate	glucose	n_year_chd
sex	1.00	-0.02	0.02	0.21	0.33	-0.05	-0.00	0.00	0.01	-0.07	-0.05	0.05	0.07	-0.11	0.00	0.09
age	-0.02	1.00	-0.16	-0.21	-0.19	0.13	0.05	0.31	0.11	0.27	0.39	0.21	0.14	-0.00	0.12	0.23
education	0.02	-0.16	1.00	0.03	0.01	-0.01	-0.03	-0.08	-0.04	-0.01	-0.12	-0.06	-0.14	-0.06	-0.03	-0.06
current_smoker	0.21	-0.21	0.03	1.00	0.77	-0.05	-0.04	-0.11	-0.04	-0.05	-0.13	-0.12	-0.16	0.05	-0.05	0.02
cig_per_day	0.33	-0.19	0.01	0.77	1.00	-0.05	-0.04	-0.07	-0.04	-0.03	-0.09	-0.06	-0.09	0.06	-0.05	0.05
bp_meds	-0.05	0.13	-0.01	-0.05	-0.05	1.00	0.11	0.26	0.05	0.09	0.27	0.20	0.11	0.01	0.05	0.09
prevalent_stroke	-0.00	0.05	-0.03	-0.04	-0.04	0.11	1.00	0.07	0.01	0.01	0.06	0.06	0.04	-0.02	0.02	0.05
prevalenthyp	0.00	0.31	-0.08	-0.11	-0.07	0.26	0.07	1.00	0.08	0.17	0.70	0.62	0.30	0.15	0.09	0.18
diabetes	0.01	0.11	-0.04	-0.04	-0.04	0.05	0.01	0.08	1.00	0.05	0.10	0.05	0.09	0.06	0.61	0.09
total_chol	-0.07	0.27	-0.01	-0.05	-0.03	0.09	0.01	0.17	0.05	1.00	0.22	0.17	0.12	0.09	0.05	0.09
sys_bp	-0.05	0.39	-0.12	-0.13	-0.09	0.27	0.06	0.70	0.10	0.22	1.00	0.79	0.33	0.18	0.13	0.22
dia_bp	0.05	0.21	-0.06	-0.12	-0.06	0.20	0.06	0.62	0.05	0.17	0.79	1.00	0.39	0.18	0.06	0.15
bmi	0.07	0.14	-0.14	-0.16	-0.09	0.11	0.04	0.30	0.09	0.12	0.33	0.39	1.00	0.07	0.08	0.08
heart_rate	-0.11	-0.00	-0.06	0.05	0.06	0.01	-0.02	0.15	0.06	0.09	0.18	0.18	0.07	1.00	0.10	0.02
glucose	0.00	0.12	-0.03	-0.05	-0.05	0.05	0.02	0.09	0.61	0.05	0.13	0.06	0.08	0.10	1.00	0.12
ten_year_chd	0.09	0.23	-0.06	0.02	0.05	0.09	0.05	0.18	0.09	0.09	0.22	0.15	0.08	0.02	0.12	1.00

All features show weak correlations with the target variable (ten_year_chd).

Modelling Approach



Raw data

Cleaned data

Data pre-processing

RobustScaler

MinmaxScaler

Modelling

LogisticRegression

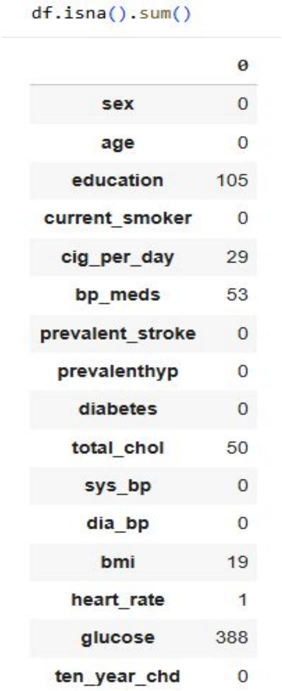
RidgeClassifier

KNeighborClassifier

Evaluation

Recall

Drop NA



```
df.isna().sum()
```

	0
sex	0
age	0
education	105
current_smoker	0
cig_per_day	29
bp_meds	53
prevalent_stroke	0
prevalenthyp	0
diabetes	0
total_chol	50
sys_bp	0
dia_bp	0
bmi	19
heart_rate	1
glucose	388
ten_year_chd	0

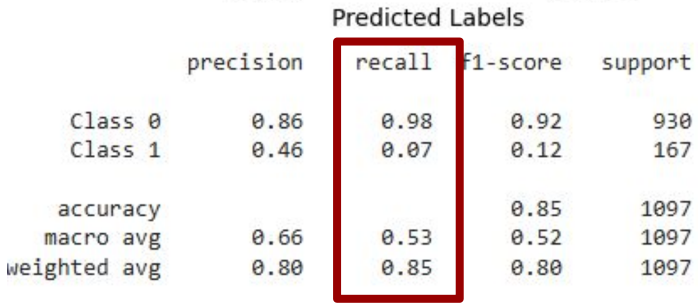
Stratified split data

stratify=y

Imbalance class handling

- SMOTE
- Oversampling
- Class weight

Model without imbalance handling



Predicted Labels

```
              precision    recall  f1-score   support

     Class 0       0.86      0.98      0.92       930
     Class 1       0.46      0.07      0.12       167

    accuracy                           0.85      1097
   macro avg       0.66      0.53      0.52      1097
weighted avg       0.80      0.85      0.80      1097
```

Hyperparameter Tuning

SMOTE

- k_neighbor

LogisticRegression

- C
- penalty
- l1_ratio
- **class_weight**

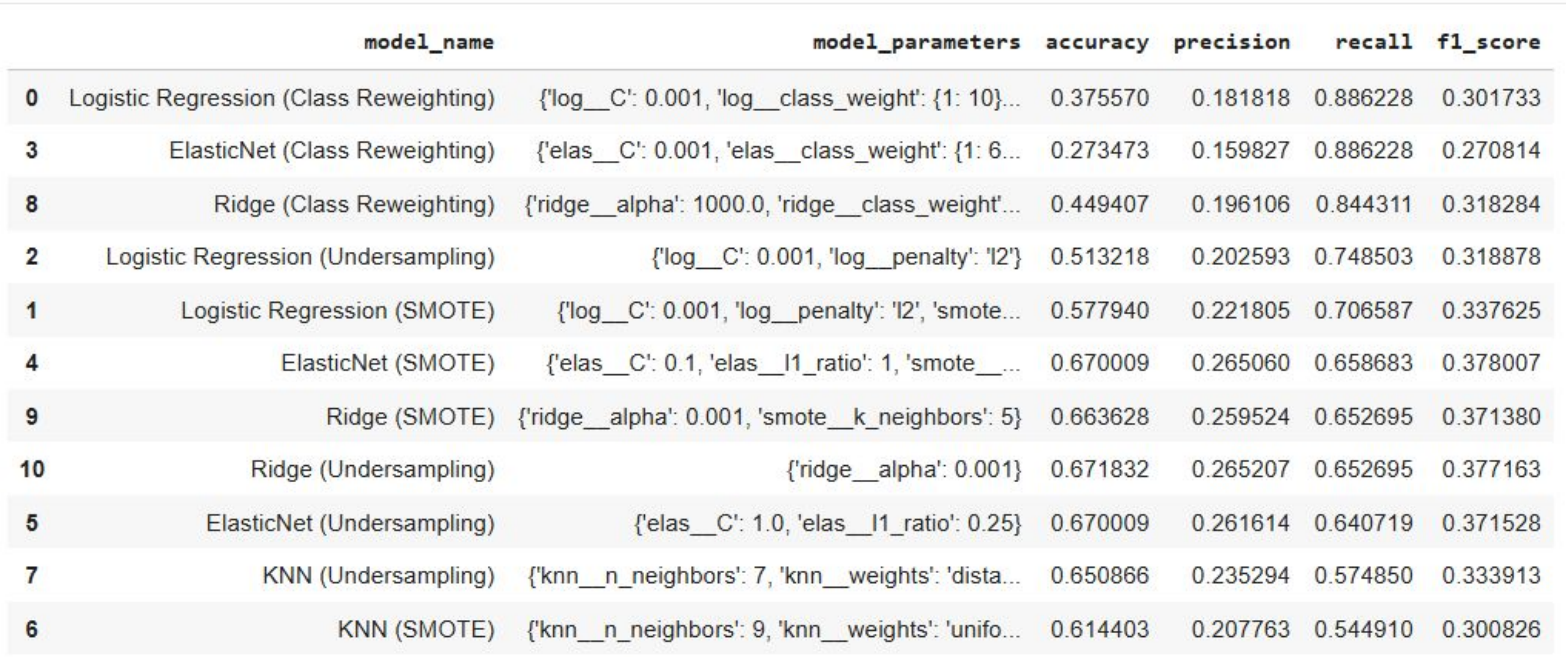
KNeighborsClassifier

- n_neighbor
- weight

RidgeClassifier

- alpha
- **class_weight**

The Best Model: Logistic Regression

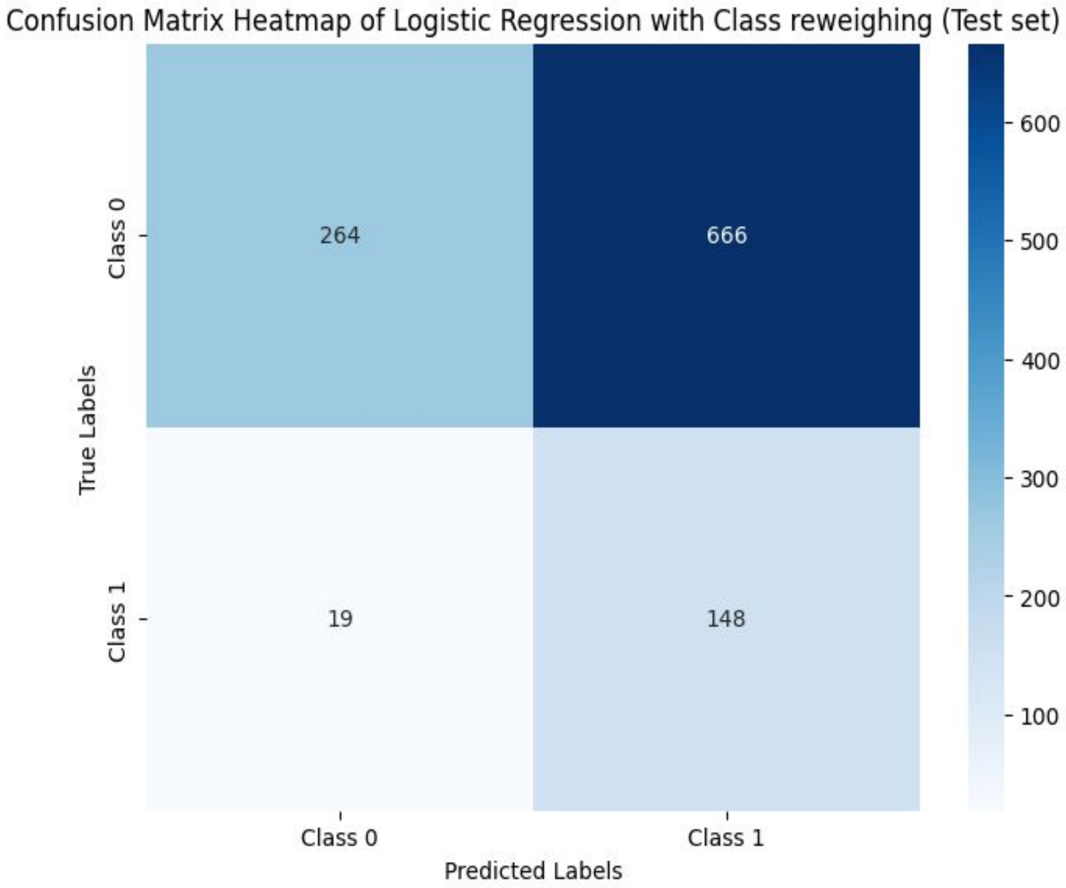


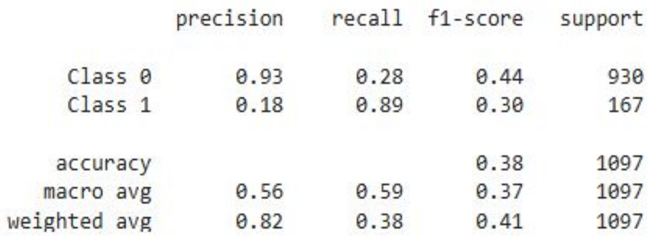
	model_name	model_parameters	accuracy	precision	recall	f1_score
0	Logistic Regression (Class Reweighting)	{'log__C': 0.001, 'log__class_weight': {1: 10}...	0.375570	0.181818	0.886228	0.301733
3	ElasticNet (Class Reweighting)	{'elas__C': 0.001, 'elas__class_weight': {1: 6...	0.273473	0.159827	0.886228	0.270814
8	Ridge (Class Reweighting)	{'ridge__alpha': 1000.0, 'ridge__class_weight'...	0.449407	0.196106	0.844311	0.318284
2	Logistic Regression (Undersampling)	{'log__C': 0.001, 'log__penalty': 'l2'}	0.513218	0.202593	0.748503	0.318878
1	Logistic Regression (SMOTE)	{'log__C': 0.001, 'log__penalty': 'l2', 'smote...	0.577940	0.221805	0.706587	0.337625
4	ElasticNet (SMOTE)	{'elas__C': 0.1, 'elas__l1_ratio': 1, 'smote__...	0.670009	0.265060	0.658683	0.378007
9	Ridge (SMOTE)	{'ridge__alpha': 0.001, 'smote__k_neighbors': 5}	0.663628	0.259524	0.652695	0.371380
10	Ridge (Undersampling)	{'ridge__alpha': 0.001}	0.671832	0.265207	0.652695	0.377163
5	ElasticNet (Undersampling)	{'elas__C': 1.0, 'elas__l1_ratio': 0.25}	0.670009	0.261614	0.640719	0.371528
7	KNN (Undersampling)	{'knn__n_neighbors': 7, 'knn__weights': 'dista...	0.650866	0.235294	0.574850	0.333913
6	KNN (SMOTE)	{'knn__n_neighbors': 9, 'knn__weights': 'unifo...	0.614403	0.207763	0.544910	0.300826

Best Model Performance

```
# The best logistic regression model
preprocessor = ColumnTransformer(
   transformers=[
      ('num', RobustScaler(), num_var),
      ('cat', 'passthrough', cat_var) ],
   remainder='passthrough')
pipe_step = [
   ('scaler', preprocessor),
   ('log', LogisticRegression(max_iter=10000,solver='liblinear', random_state=42))]
param_grid = {
   'log__C': [0.001],  # np.logspace(-3, 3, 7)
   'log__penalty': ['l2', 'l1'],
   'log__class_weight' : [{1: 10}, 'balanced', None, {1: 2},{1: 3},{1: 4},{1: 5}, {1: 6}, {1: 7}, {1: 8}, {1: 9}]}

scoring = {'accuracy': 'accuracy', 'f1': 'f1','precision': 'precision',  'recall': 'recall'}
pipe = Pipeline(pipe_step)
grid = GridSearchCV(pipe, param_grid, cv=5,scoring = scoring, refit="recall",n_jobs=-1)
grid.fit(X_train, y_train)
best_grid = grid.best_estimator_
y_pred_test = best_grid.predict(X_test)
conf_matrix = confusion_matrix(y_test, y_pred_test)
report = classification_report(y_test, y_pred_test, target_names=['Class 0', 'Class 1'])
```



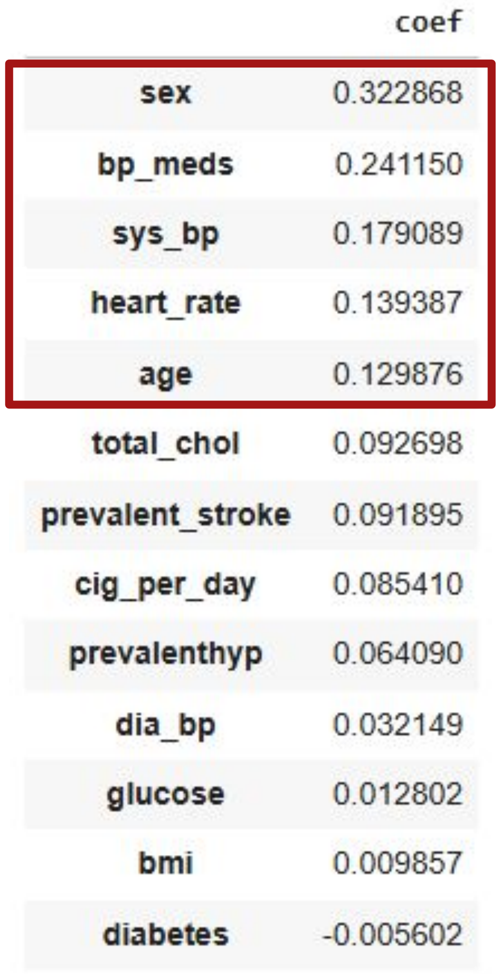


```
              precision    recall  f1-score   support

     Class 0       0.93      0.28      0.44       930
     Class 1       0.18      0.89      0.30       167

    accuracy                           0.38      1097
   macro avg       0.56      0.59      0.37      1097
weighted avg       0.82      0.38      0.41      1097
```

Model Interpretation



	coef
sex	0.322868
bp_meds	0.241150
sys_bp	0.179089
heart_rate	0.139387
age	0.129876
total_chol	0.092698
prevalent_stroke	0.091895
cig_per_day	0.085410
prevalenthyp	0.064090
dia_bp	0.032149
glucose	0.012802
bmi	0.009857
diabetes	-0.005602

Coefficient

- Significantly increases CHD risk in **males**.
- Patients on **blood pressure medication** show substantially elevated CHD risk.
- Higher **systolic blood pressure** is a major risk factor.
- Elevated resting **heart rate** shows moderate positive association with CHD risk.
- **Older patients** face higher CHD risk.

Conclusion and recommendations

- **Targeting heart disease** as the leading preventable cause of death, this study demonstrates that knowing specific data is effective for prevention.

- **The Multiplier Effect:** Combined risk factors make the danger much higher, requiring us to treat the whole picture instead of just one problem.

- **Model Selected:** Logistic Regression (Adjusted with Class Weights)
- **Performance Metric:** Highest Recall at 0.86
- **Performance Trade-off:** Accepted a lower Precision of 0.19
- **Design Justification:** Prioritized Recall because the cost of a False Negative (missed diagnosis) is significantly higher than the cost of a False Positive (unnecessary follow-up).
- **Top Contributing Factors:** Male gender, Blood Pressure and Medication, Heart rate, Age
- **Further improvement:** Advanced model, balance dataset