



Heart disease

DS512/513 Data Analytics

DS514/515 Data Science

Narawit Tharapoompipat 68199160275

Ratanaporn Thaochalee 68199160293

Sakdhithad Chanfeungfu 68199160301

[14 December 2025]

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 17 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-weighting 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.



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.

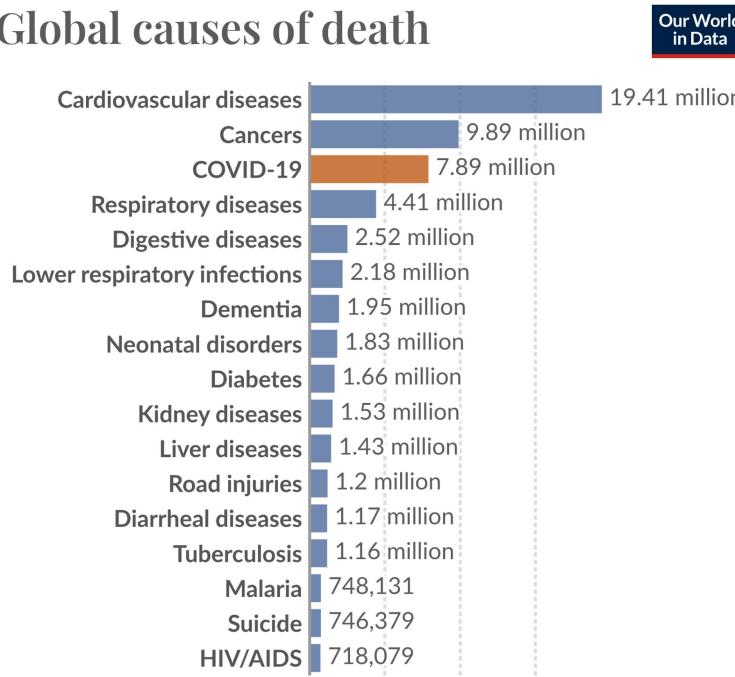




Why interest ?

The Global Health Crisis

Global causes of death



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.

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](https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression)

kaggle

Logistic regression To predict heart disease

heart disease prediction

Data Card Code (308) Discussion (12) Suggestions (0)

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 publicly 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 & 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, 17 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

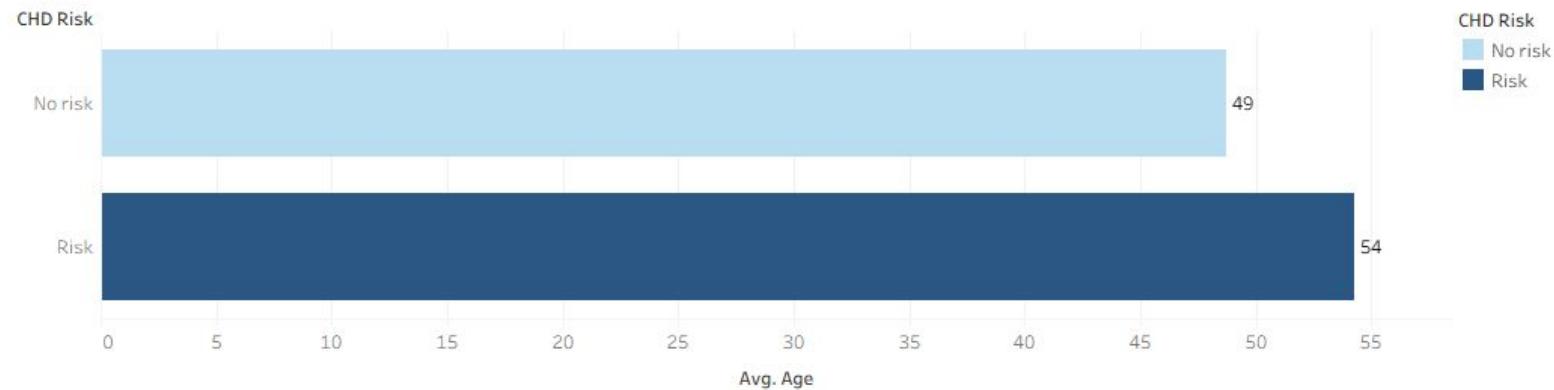
- Does age have a significant impact on ten-year CHD risk?
- Age is a risk factor. But does the impact of age on Ten-Year CHD risk different between genders?

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

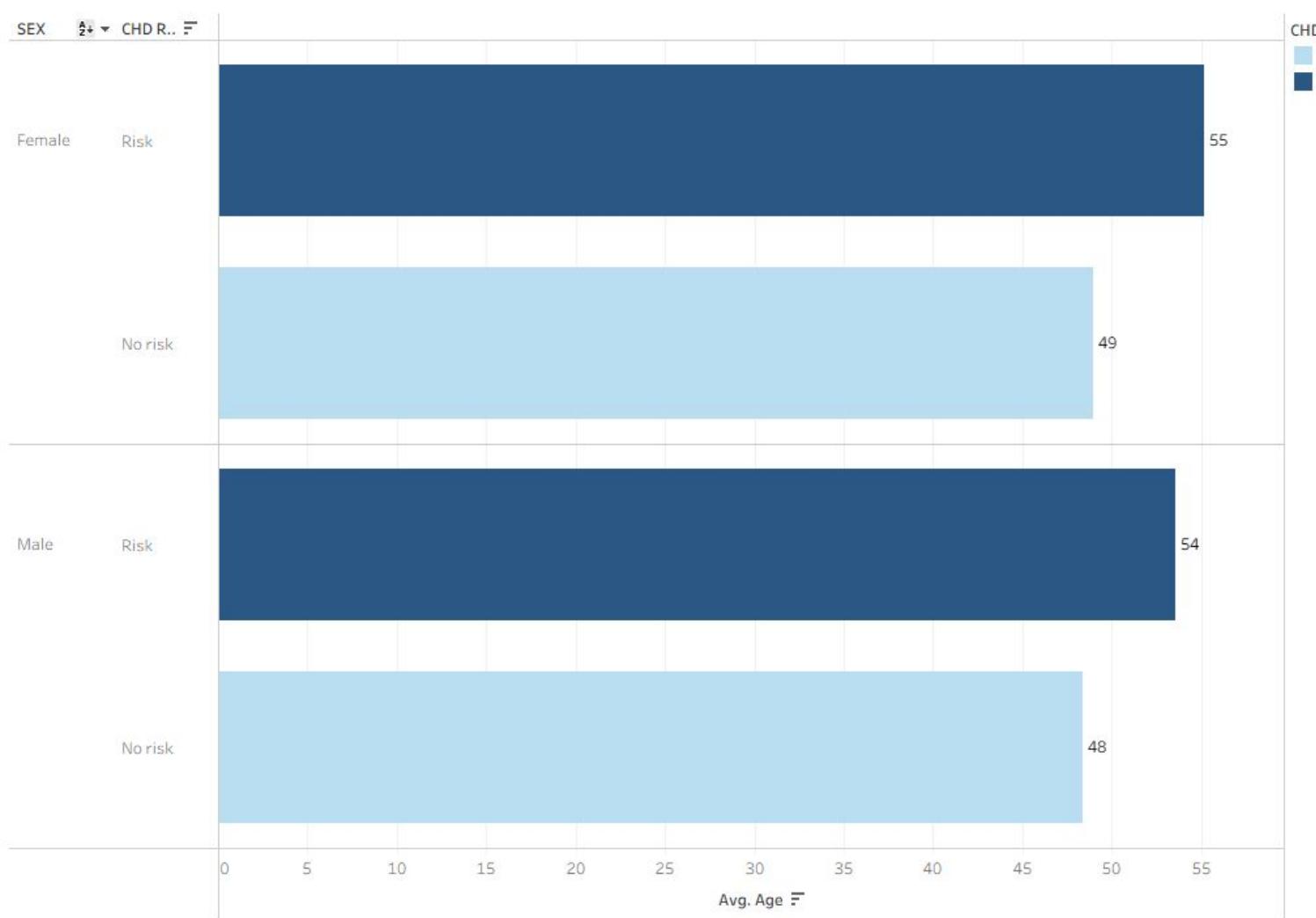


- Does age have a significant impact on ten-year CHD risk?

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



Findings and Insights

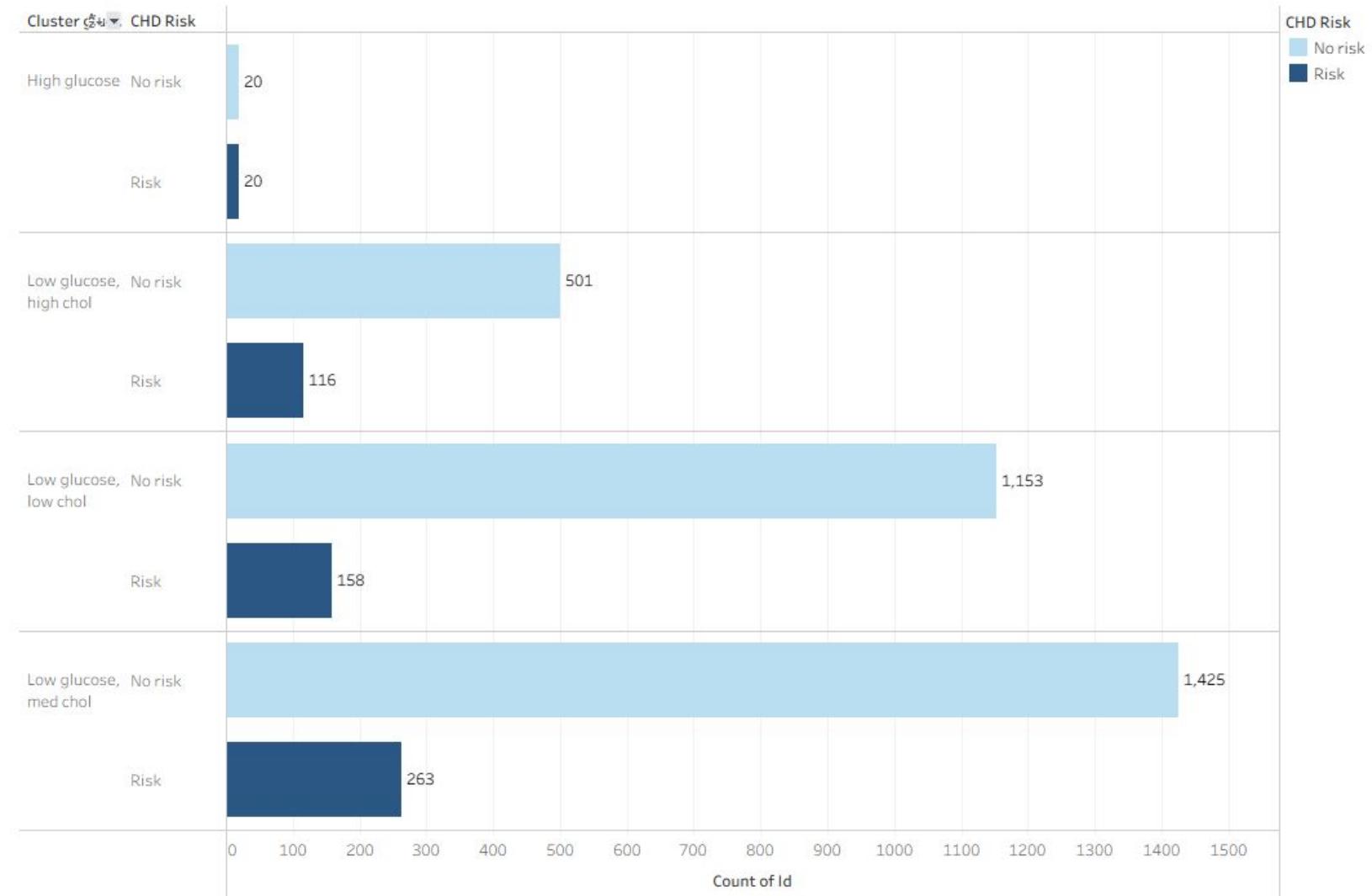


- **Age is a risk factor.** But is the impact of age on ten-year CHD risk different between genders?

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



Findings and Insights

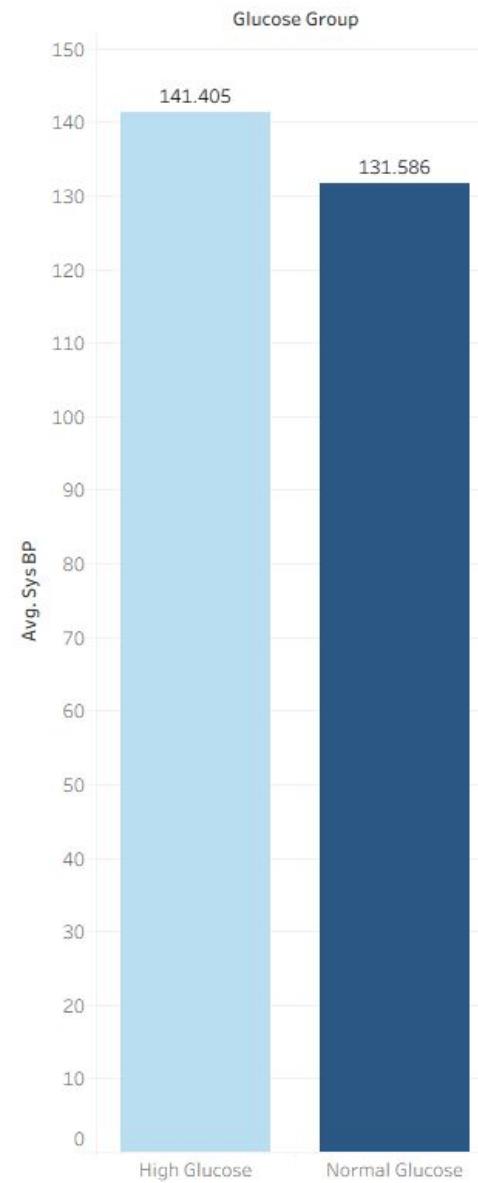


- **Which is the Dominant Risk Factor: Glucose or Cholesterol ?**

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



Findings and Insights

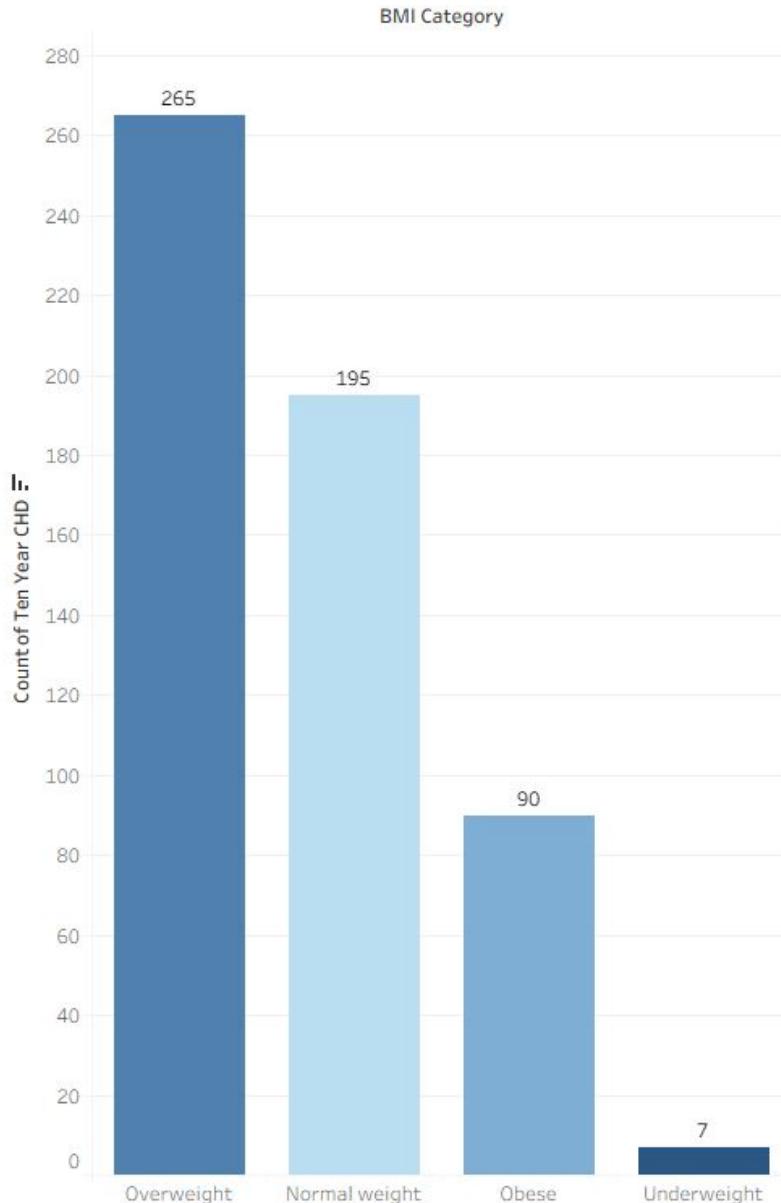


- Is elevated blood glucose associated with higher systolic blood pressure (sysBP)?

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.



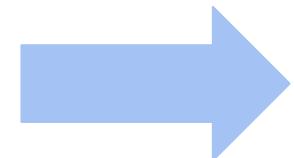
Modelling approach

Classification problem

13 Features to train model

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

Predict

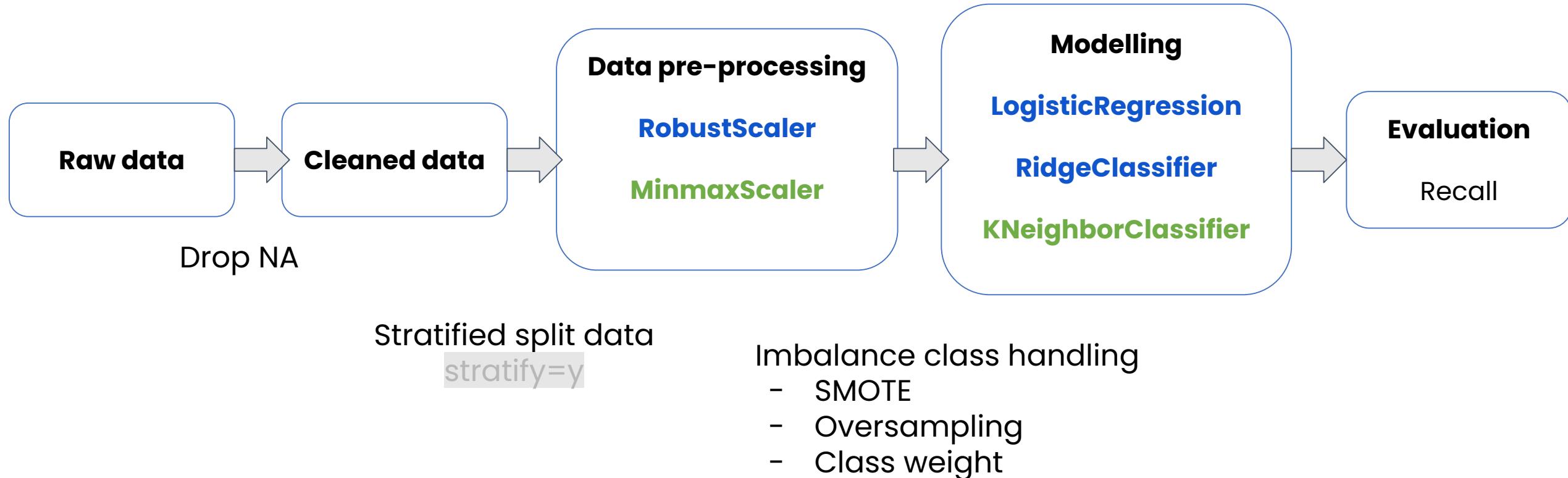


Target class

ten_year_chd
0,1



Modelling approach





Hyperparameter tuning

SMOTE

- k_neighbor

LogisticRegression

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

KNeighborsClassifier

- n_neighbor
- weight

RidgeClassifier

- alpha
- **class_weight**



The best model: LogisticRegression

	model_name	model_parameters	accuracy	precision	recall	f1_score
0	Logistic Regression (Class Reweighting)	{'log_C': 0.001, 'log_class_weight': {1: 10}, 'log_solver': 'sag'} {'log_C': 0.001, 'log_class_weight': {1: 10}, 'log_solver': 'sag'}	0.375570	0.181818	0.886228	0.301733
3	ElasticNet (Class Reweighting)	{'elas_C': 0.001, 'elas_class_weight': {1: 6}, 'elas_l1_ratio': 0.25} {'elas_C': 0.001, 'elas_class_weight': {1: 6}, 'elas_l1_ratio': 0.25}	0.273473	0.159827	0.886228	0.270814
8	Ridge (Class Reweighting)	{'ridge_alpha': 1000.0, 'ridge_class_weight': {1: 10}, 'ridge_l1_ratio': 0.25} {'ridge_alpha': 1000.0, 'ridge_class_weight': {1: 10}, 'ridge_l1_ratio': 0.25}	0.449407	0.196106	0.844311	0.318284
2	Logistic Regression (Undersampling)	{'log_C': 0.001, 'log_penalty': 'l2'} {'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_k_neighbors': 5} {'log_C': 0.001, 'log_penalty': 'l2', 'smote_k_neighbors': 5}	0.577940	0.221805	0.706587	0.337625
4	ElasticNet (SMOTE)	{'elas_C': 0.1, 'elas_l1_ratio': 1, 'smote_k_neighbors': 5} {'elas_C': 0.1, 'elas_l1_ratio': 1, 'smote_k_neighbors': 5}	0.670009	0.265060	0.658683	0.378007
9	Ridge (SMOTE)	{'ridge_alpha': 0.001, 'smote_k_neighbors': 5} {'ridge_alpha': 0.001, 'smote_k_neighbors': 5}	0.663628	0.259524	0.652695	0.371380
10	Ridge (Undersampling)	{'ridge_alpha': 0.001} {'ridge_alpha': 0.001}	0.671832	0.265207	0.652695	0.377163
5	ElasticNet (Undersampling)	{'elas_C': 1.0, 'elas_l1_ratio': 0.25} {'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': 'distance'} {'knn_n_neighbors': 7, 'knn_weights': 'distance'}	0.650866	0.235294	0.574850	0.333913
6	KNN (SMOTE)	{'knn_n_neighbors': 9, 'knn_weights': 'uniform'} {'knn_n_neighbors': 9, 'knn_weights': 'uniform'}	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', 'l2']  
    'log__class_weight': [{1: 10}]} #['balanced', None, {1: 2},{1: 3},{1: 4},{1: 5}, {1: 6}, {1: 7}, {1: 8}  
{1: 9}, {1: 10}]
```

```
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)
```

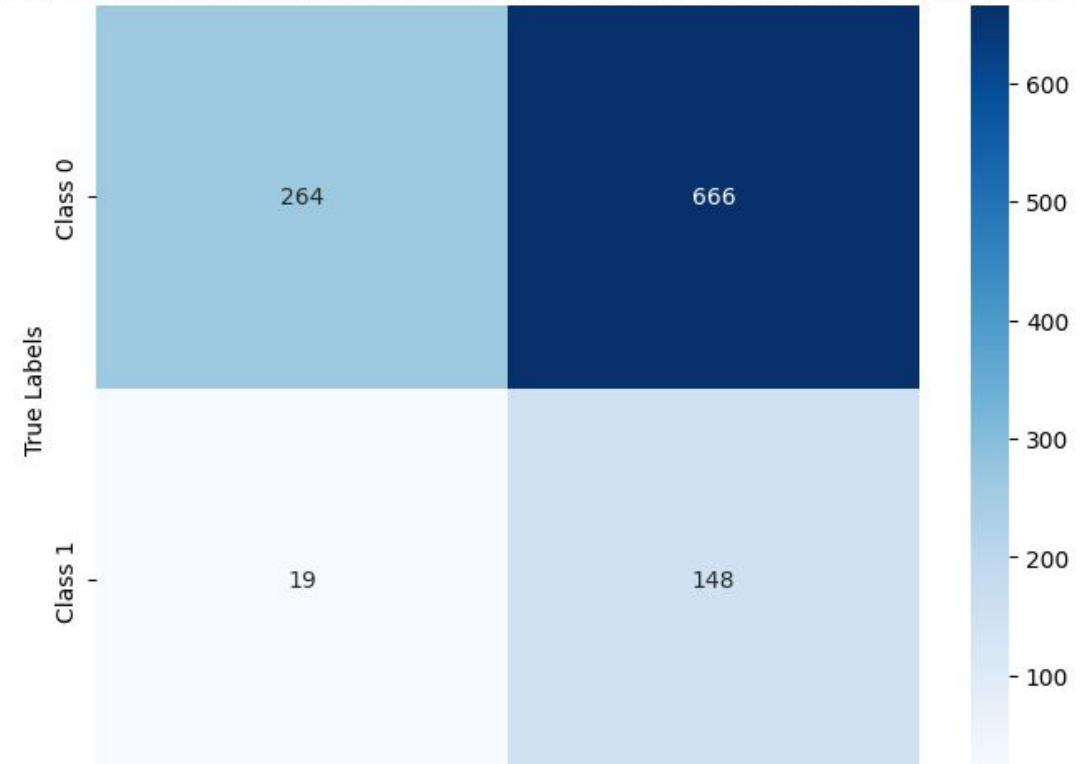
```
best_grid = grid.best_estimator_
```

```
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'])
```

Confusion Matrix Heatmap of Logistic Regression with Class reweighing (Test set)



	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

- **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