Heart disease

DS512/513 Data Analytics
DS514/515 Data Science

Narawit Tharapoompipat 68199160275
Ratanaporn Thaochalee 68199160293
Sakdhithad Chanfeungfu 68199160301

[14 December 2025]

DATA PROJECT CANVAS

Designed by:

Date:

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.
- Risk rises with BMI, but the highest danger occurs when multiple risk factors are combined.

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

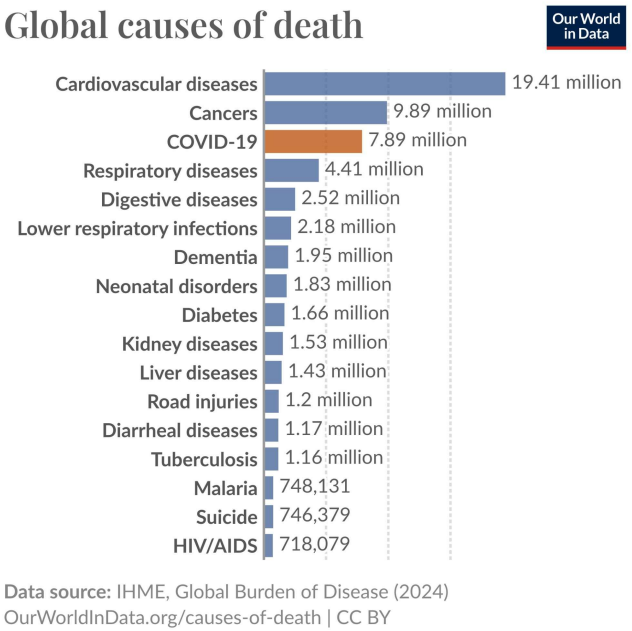
Heart disease

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



- 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

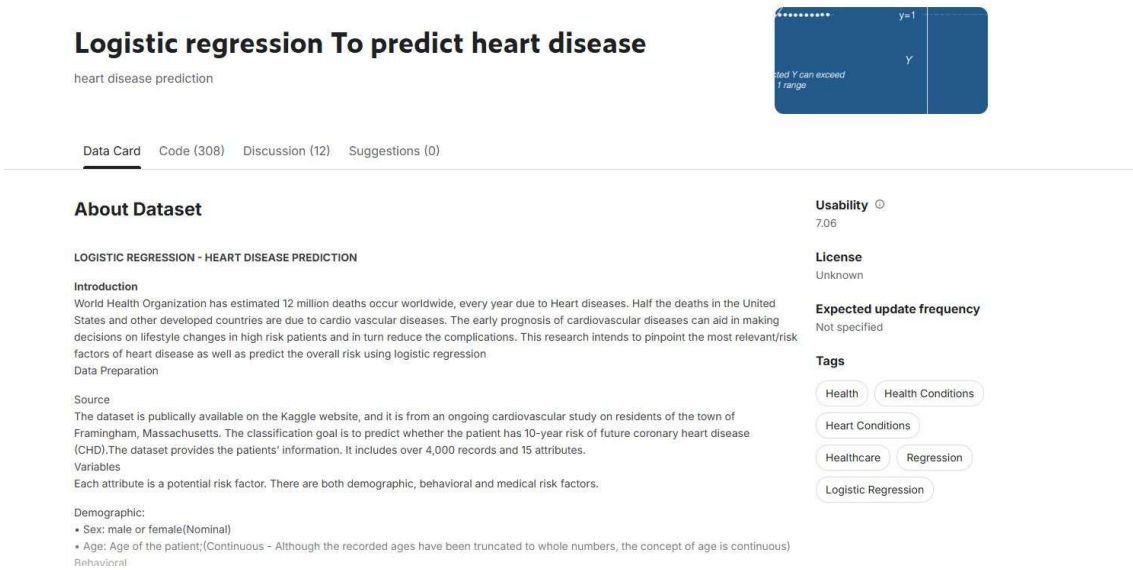
- totChol (Cholesterol), sysBP, diaBP, BMI, heart Rate, glucose

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

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 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 & 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

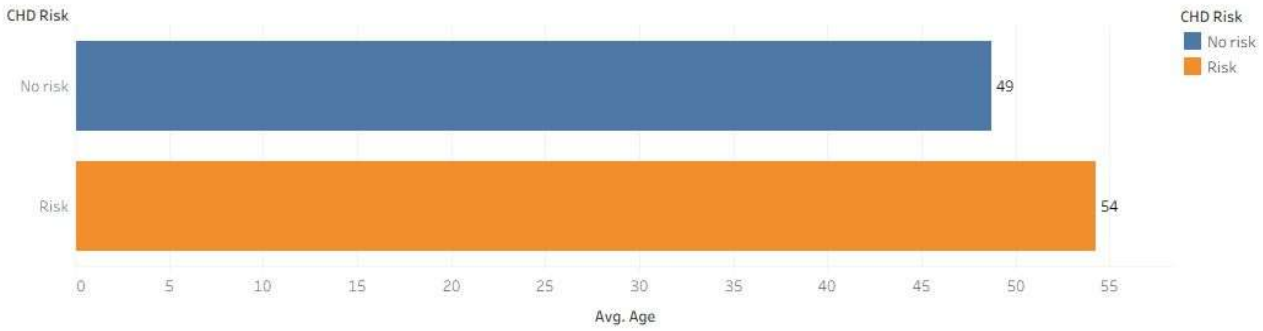
General Information

- 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)?
- Does the BMI category (underweight, normal weight, overweight, obese) show a clear trend in Ten-Year CHD risk?
- Does having multiple risk factors increase TenYearCHD risk compared to just one?

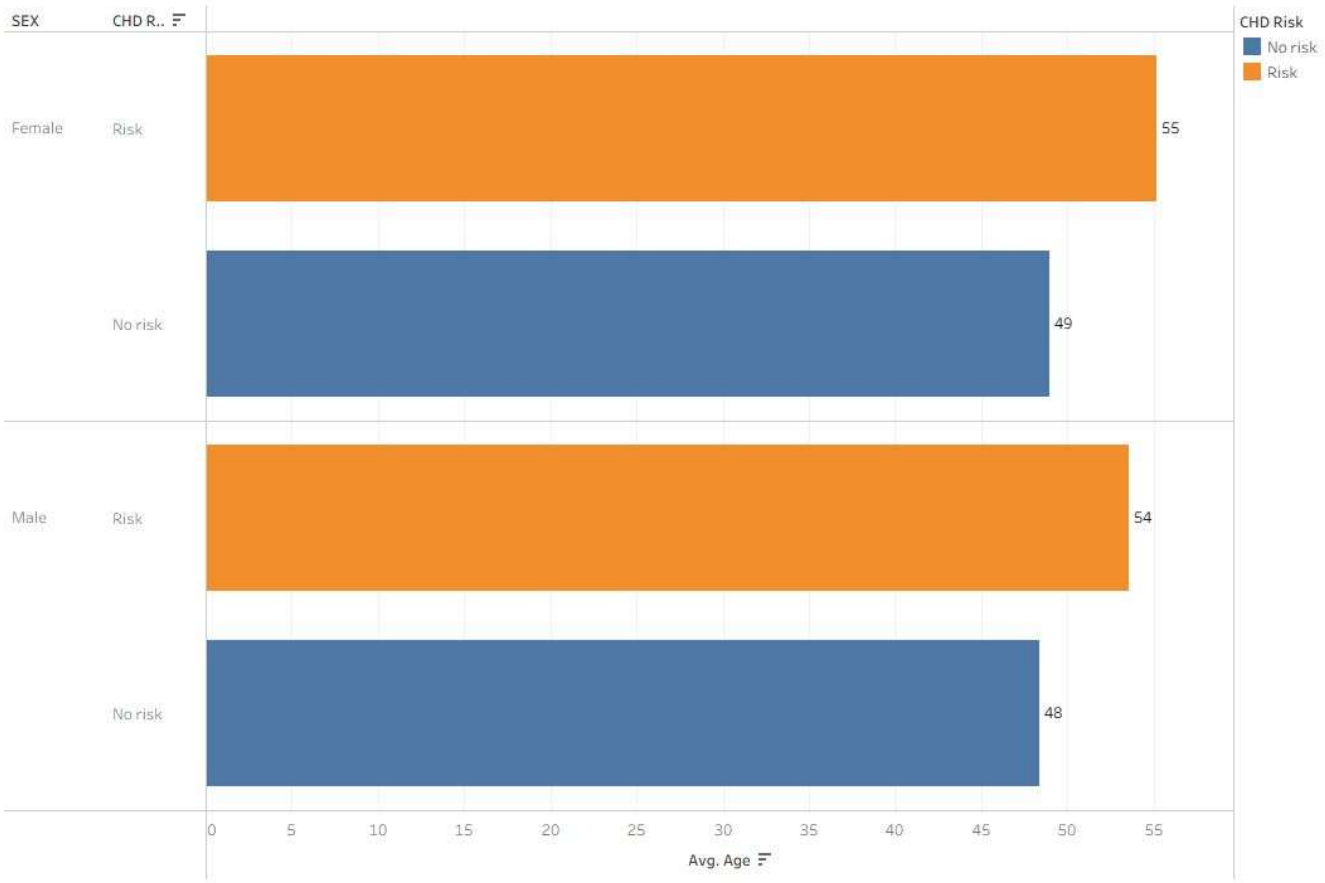
Findings and Insights



- **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.

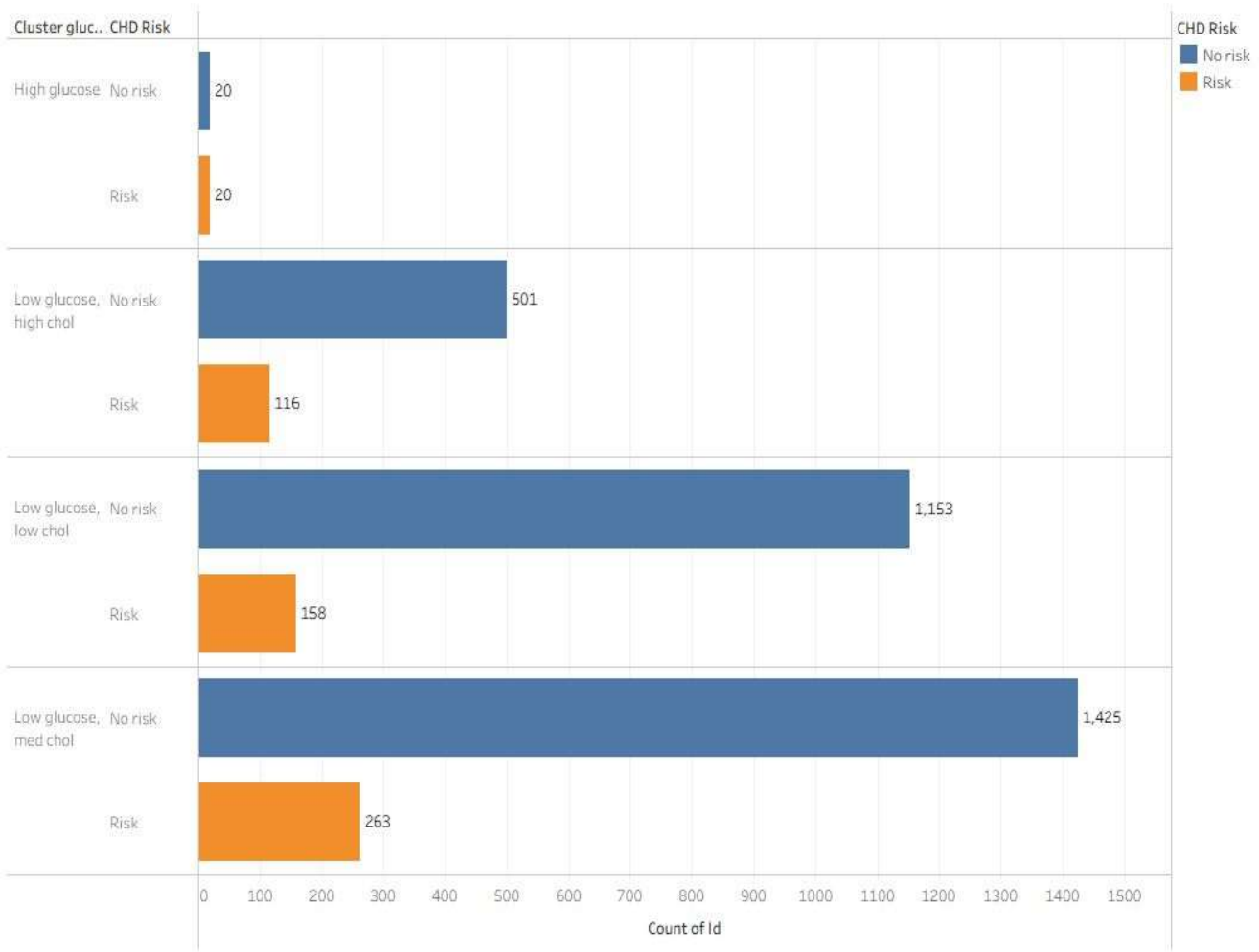
Findings and Insights



- **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.

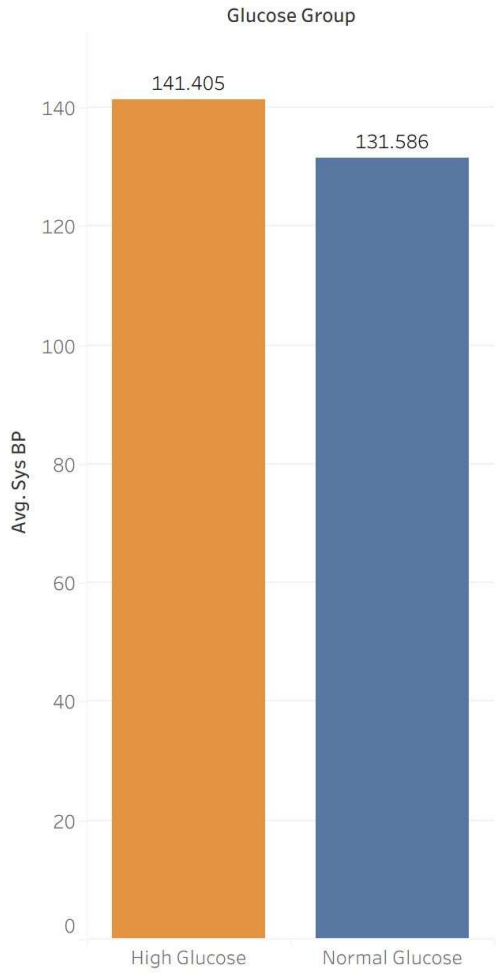
Findings and Insights

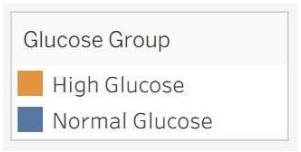


- **Which is the Dominant Risk Factor: Glucose or Cholesterol ?**

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

Findings and Insights

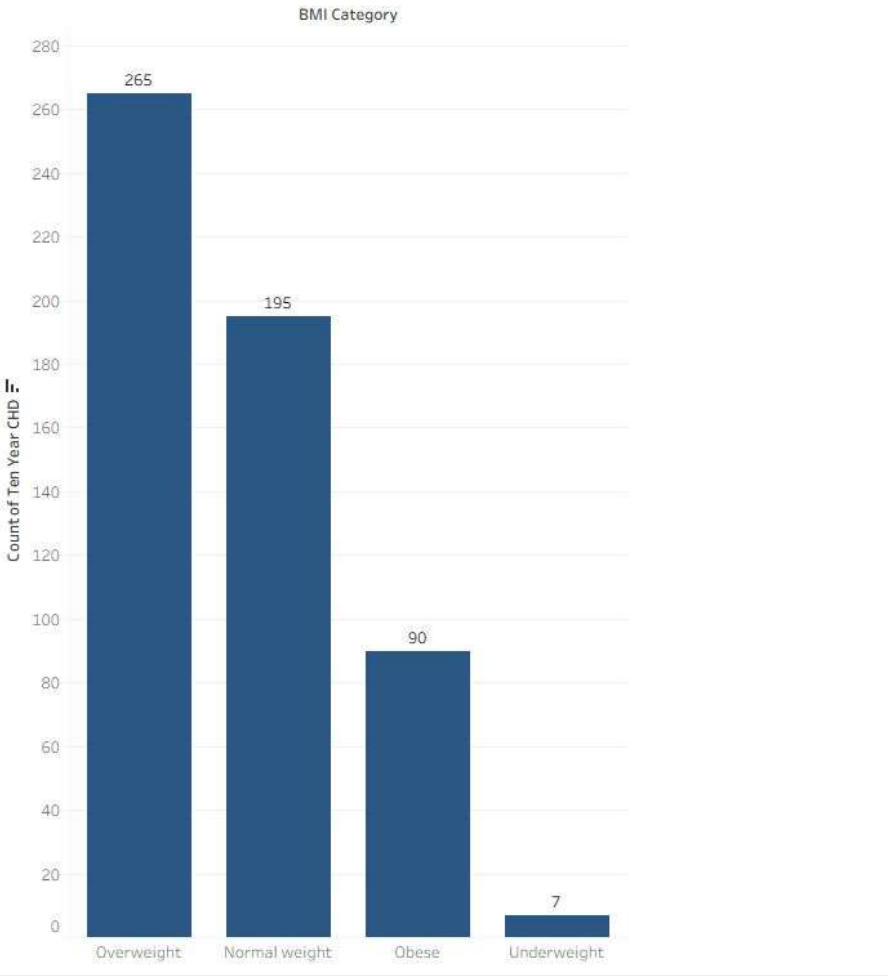




- Is elevated blood glucose associated with higher systolic blood pressure (sysBP)?

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



- **Does the BMI category** show a clear trend in Ten-Year CHD risk?

Heart disease risk rises with BMI, peaking at nearly 20% in the obese group compared to 12% in normal-weight individuals, confirming a strong association with increased CHD risk.

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
- **Does the BMI category show a clear trend in Ten-Year CHD risk?**
  Higher BMI is strongly associated with an increased risk of heart disease.
- **Does having multiple risk factors increase TenYearCHD risk compared to just one?**
  Multiple combined risk factors result in the highest CHD risk