Disease Risk Prediction Using Data Science:
Framingham Heart Study (Case Study)

A project done under guidance of Dr. Ganapati Natarajan

CEO and Chief Scientist

Causal Intelligence Ltd

London (UK) / Mysuru (India)

Submitted by: Mahima Shivraj

Abstract:

Cardiovascular disease (CVD) remains a major cause of morbidity and mortality globally, thus, there continues to be an important interest in the development of robust risk prediction algorithms. This study applied an end-to-end data science approach based on the OSEMN framework to generate a 10-year risk prediction model for coronary heart disease (CHD) from the Framingham teaching dataset. Once the dataset was preprocessed to compensate missing values, exploratory data analysis showed for highly observable patterns and interrelations between predictors and CHD occurrences. The low prevalence of CHD (~15%) was considered when a logistic regression model was trained with stratified sampling and with class weighting. The model's performance was evaluated using probability-based metrics like ROC-AUC, precision, recall, F1-score, and threshold analyses. Coefficients and odds ratios were interpreted to assess clinical plausibility and consistency with EDA findings. The results demonstrate that factors such as age, sex, diabetes, blood pressure, and prior stroke are important predictors of 10-year CHD risk. This model provides a transparent and interpretable baseline for risk prediction, while highlighting potential avenues for future research using more complex machine learning approaches and external validation.

Table of Contents:

Chapter 1: Introduction

1.1. Role of Data Science in Disease Risk Prediction

Data science has dramatically changed healthcare with the ability to systematically analyze substantial amounts of complex medical data. Data science involves use of machine learning and statistical modeling to find the association or relationships in the data that would not be able to be found with other types of analyses. The use of data science is particularly useful when it comes to finding and identifying at-risk patients with chronic conditions such as coronary heart disease (CHD). CHD is a disease which develops from multiple influences including age, sex, race/ethnicity, lifestyle and clinical characteristics.
The purpose of developing risk prediction models is to identify the probability that an individual will develop a condition within a specified period of time. By identifying individuals who are at elevated risk of developing a cardiovascular related condition, providers can begin to implement prevention strategies and lifestyle changes and provide closer follow-up care to those identified as being at higher risk. Therefore, the development of data driven risk predictions is a critical component in the support of both population level health management and the practice of personalized medicine.

1.2. Risk prediction vs. Clinical diagnosis:

Predicting disease risk is about how likely an individual will develop a disease in the future using current risk factors (for example, age, blood pressure, cholesterol levels). Diagnosing a disease is finding out whether a patient has a disease or not by conducting physical examinations, laboratory tests and by utilizing a physician's expertise.

The predictive model does not replace a doctor with it being a tool that helps a physician to identify which patients could be considered for additional testing and/or interventions.

Risk prediction models are commonly used as screening tools for the purpose of prioritizing patients to receive additional testing and/or interventions.
The importance of this differentiation is most evident in chronic diseases such as CHD, because identifying a patient's risk early reduces potential complications down the line regardless of if the patient has had symptoms yet.

1.3. Importance of interpretable models in healthcare:

Healthcare providers need to know that their model's decisions are reasonable based on the data they were trained on so that they can have confidence in applying those

results to real patient care. Model interpretation is important for trust in application of artificial intelligence models in healthcare, but it will limit the adoption of black box models by healthcare providers because of lack of transparency into how each factor influences the final prediction.

Interpretable models such as logistic regression provide clear, explainable relationships between predictors and outcomes through coefficients and odds ratios. This allows clinicians and researchers to validate model behavior against medical knowledge and ensures that predictions remain clinically plausible. For this reason, interpretable models are particularly suitable for baseline disease risk prediction and form a solid foundation for more advanced modeling approaches.

Chapter 2: Background

2.1. Cardiovascular disease and Public Health Relevance:

Cardiovascular disease (CVD) is still one of the leading causes of death and morbidity in the world. CHD accounted for a substantial proportion of premature deaths and long-term disability, putting significant pressure on healthcare systems. [1], [14]. The early detection of risk factors and appropriate intervention help to prevent most cardiovascular diseases, and therefore risk prediction is a core element of public health interventions.

Smoking, physical activity, poor diet, and clinical factors like hypertension, diabetes and obesity also increased risk for cardiovascular diseases. Because these factors can interact, data-driven approaches are more suitable for quantifying and understanding how combined effect may affect disease outcome.

2.2. Overview of the Framingham Heart Study:

The Framingham Heart Study is a landmark longitudinal study initiated in 1948 in Framingham, Massachusetts, to assess cardiovascular risk factors for heart disease. For generations of participants, demographic, behavioral, and clinical data has been collected, and many known cardiovascular risk factors such as high blood pressure, smoking, high cholesterol and diabetes have been identified. [1], [14]

The present study has been instrumental in shaping contemporary cardiovascular epidemiology and has generated a rich set of data for research, education and model development. Its findings are still relevant to clinical practice and preventive medicine worldwide.

2.3. Description of the Framingham teaching dataset:

The Framingham teaching dataset, obtained from Kaggle is a curated subset of the original study data commonly used for educational and methodological purposes. [3] It contains 4240 observations (rows) and 16 variables (columns). Each row represents an individual patient, and each column represents a health-related attribute. It has patient-level information on demographic variables, lifestyle behaviors, medical history, and clinical measurements. The dataset includes both numerical variables (such as age, blood pressure, BMI, and cholesterol levels) and binary indicators (such as smoking status, diabetes, and medication use).

The outcome variable of interest is TenYearCHD, which indicates whether an individual developed coronary heart disease within a ten-year follow-up period. The dataset is particularly suitable for supervised learning tasks due to its structured format and well-defined outcome.

2.4. Objective of the study:

The primary objective of this study is to develop a data-driven model to predict the 10-year risk of coronary heart disease (CHD) using the Framingham teaching dataset. Logistic regression is employed to estimate individual risk probabilities based on known cardiovascular risk factors. The aim is not to provide a clinical diagnosis, but rather, a way of showing how the use of analytically valid statistical models is a useful baseline tool for predicting disease risk and to assist with early risk stratification.

Chapter 3: Data Science Methodology

3.1. Overview of the OSEMN Framework:

The OSEMN framework is a commonly used data science workflow that details the key stages involved in transforming raw data into actionable insights. OSEMN stands for **Obtain, Scrub, Explore, Model, and iNterpret**, representing a structured and iterative approach to data analysis. This framework ensures that data quality, exploratory analysis, modeling, and interpretation are systematically addressed.

By following the OSEMN framework, data science projects maintain transparency and reproducibility while reducing the risk of overlooking critical steps. The framework is particularly suitable for healthcare applications, where data integrity, interpretability, and careful evaluation are essential.

3.2. Mapping OSEMN stages to this study:

Each stage of the OSEMN framework was applied in this study as follows:

- Obtain: The Framingham teaching dataset was obtained as a structured dataset containing demographic, lifestyle, and clinical variables relevant to cardiovascular health.
- Scrub: Data cleaning was performed by identifying missing values and applying median imputation for numerical variables and mode imputation for binary or categorical variables. This ensured a complete dataset without removing observations.
- Explore: Exploratory Data Analysis (EDA) was conducted to understand variable distributions, identify class imbalance, and examine relationships between predictors and the CHD outcome using summary statistics and visualizations.
- Model: A logistic regression model was trained using a stratified train–test split. Class weighting was applied to address class imbalance, and model performance was evaluated using probability-based metrics.
- Interpret: Model outputs were interpreted using coefficients and odds ratios. Results were compared with EDA findings to assess consistency and clinical plausibility, and model limitations were discussed.

Chapter 4: Data Preparation (Scrub Stage)

4.1. Description of variables:

The Framingham teaching data set includes demographic, behavioral, and clinical variables commonly used in the assessment of cardiovascular risk. Variables include a combination of constant measures as well as binary measures of medical history and lifestyle behaviors. TenYearCHD is an outcome variable that indicates coronary heart disease in ten years.

Table 1: Description of Variables Used in the Study

Variable	Meaning	Type
age	Age of the patient (years)	Numerical
male	Gender (1 = male, 0 = female)	Binary
education	Education level	Ordinal

currentSmoker	Whether the person is a current smoker (1 = yes, 0 = no)	Binary
cigsPerDay	Number of cigarettes smoked per day	Numerical
BPMeds	Whether the patient is on blood pressure medication (1 = yes, 0 = no)	Binary
prevalentStroke	History of stroke (1 = yes, 0 = no)	Binary
prevalentHyp	Presence of hypertension (1 = yes, 0 = no)	Binary
Diabetes	Presence of diabetes (1 = yes, 0 = no)	Binary
totChol	Total cholesterol level (mg/dL)	Numerical
sysBP	Systolic blood pressure (mmHg)	Numerical
diaBP	Diastolic blood pressure (mmHg)	Numerical
BMI	Body Mass Index	Numerical
heartRate	Heart rate (beats per minute)	Numerical
glucose	Blood glucose level (mg/dL)	Numerical
TenYearCHD	Whether the individual developed coronary heart disease within 10 years (1 = yes, 0 = no)	Binary

4.2. Identification of missing values:

Before proceeding with analysis, we examined the dataset to identify missing values in each feature. For all variables, the number of missing entries and percentage of missing values were calculated. This step provided us with information about which features had missing information and allowed us to choose an appropriate imputation strategy. The analysis revealed that some variables, such as glucose and education, had higher proportions of missing values, while most other features were nearly complete.

4.3. Summary table of missing values by feature:

A summary of missing values for each variable is presented in the table below. This table shows the total number of missing entries and the corresponding percentage of missing values for all features. Most variables are nearly complete, with glucose and

education having the highest proportion of missing data. This overview provides a clear understanding of the dataset's completeness before applying imputation strategies.

Table 2: Summary of Missing Values for Each Feature

Variable	Missing Count	Missing %
male	0	0.00%
age	0	0.00%
education	105	2.48%
currentSmoker	0	0.00%
cigsPerDay	29	0.75%
BPMeds	53	1.33%
prevalentStroke	0	0.00%
prevalentHyp	0	0.00%
diabetes	0	0.00%
totChol	50	1.24%
sysBP	0	0.00%
diaBP	0	0.00%
BMI	19	0.78%
heartRate	1	0.08%
glucose	388	9.08%
TenYearCHD	0	0.00%

4.4. Breakdown of missing values by CHD class:

To examine whether missingness varied by outcome, missing values were summarized separately for participants with and without 10-year CHD events. The results, shown in Table, indicate that some variables (such as glucose and education) have slightly higher missing values in certain CHD classes, while most features have minimal differences. This step ensures that imputation strategies do not inadvertently bias the dataset toward one outcome class.

Table 3: Number and percentage of missing values for each variable, stratified by 10-year CHD Outcome

Variable	Missing count (CHD=0)	Missing % (CHD=0)	Missing count (CHD=1)	Missing % (CHD=1)
male	0	0.00%	0	0.00%
age	0	0.00%	0	0.00%
education	89	2.47%	16	2.48%
currentSmoker	0	0.00%	0	0.00%
cigsPerDay	27	0.75%	2	0.31%
BPMeds	42	1.17%	11	1.71%

prevalentStroke	0	0.00%	0	0.00%
prevalentHyp	0	0.00%	0	0.00%
diabetes	0	0.00%	0	0.00%
totChol	41	1.14%	9	1.40%
sysBP	0	0.00%	0	0.00%
diaBP	0	0.00%	0	0.00%
BMI	9	0.25%	10	1.55%
heartRate	0	0.00%	1	0.16%
glucose	338	9.40%	50	7.76%
TenYearCHD	0	0.00%	0	0.00%

From this table, variables such as male, age, currentSmoker, prevalentStroke, prevalentHyp, diabetes, sysBP, diaBP, heartrate, and TenYearCHD have no missing values, indicating complete observations for both CHD=0 and CHD=1 groups.

Some variables have a small proportion of missing data. For example, total cholesterol is missing in 41 patients with CHD=0 (1.14%) and 9 patients with CHD=1 (1.55%). CigsPerDay and BPMeds also show minor missing values.

The most notable is glucose, which is missing in 338 patients with CHD=0 (9.40%) and 50 patients with CHD=1 (7.76%). While this is a substantial proportion, it can be handled by median imputation to retain most of the data.

4.5. Imputation strategy:

Missing values in numerical variables were replaced using the median, while missing values in binary or categorical variables were replaced using the mode. Median imputations avoid distortions from outliers and preserve the central tendency of skewed variables, while mode imputation maintains the most common category for discrete variables. No rows were dropped during preprocessing.

4.6. Final dataset construction:

After applying the above imputation strategy:

- All missing values were resolved, and the dataset is now complete.
- The feature matrix x consists of all predictor variables listed below.
- The target variable is TenYearCHD, indicating whether the patient developed CHD within 10 years.

The variables used in the modeling were:

a) Numerical variables

- Age
- Systolic Blood Pressure

- Diastolic Blood Pressure
- BMI
- Total cholesterol
- Cigarettes per day
- Heart rate

b) Binary variables
- Sex (male/female)
- Current smoker
- Blood Pressure medication
- Prevalent hypertension
- Prevalent stroke
- Diabetes

c) Ordinal variable:
- Education level

This completed dataset forms the foundation for subsequent logistic regression modeling, ensuring that the model receives clean, consistent input and that predictions are reliable.

Chapter 5: Exploratory Data Analysis (Explore Stage)

5.1. Final dataset construction:

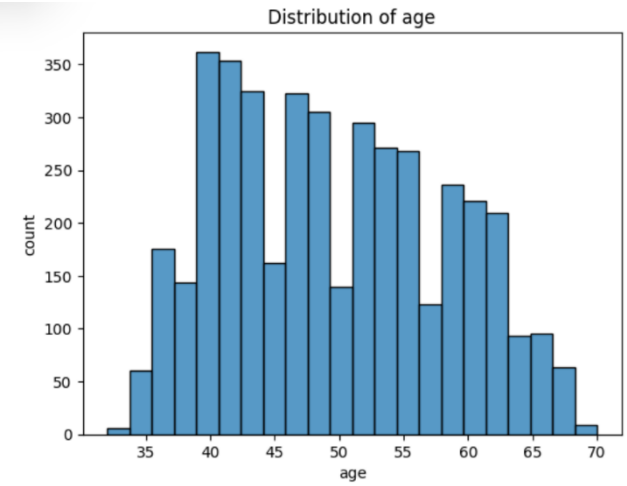
The target variable **TenYearCHD** was examined by counting the number of individuals who developed coronary heart disease within ten years and those who did not. The majority of individuals did **not** develop heart disease during the follow-up period, while a smaller proportion experienced a CHD event, indicating the presence of class imbalance.

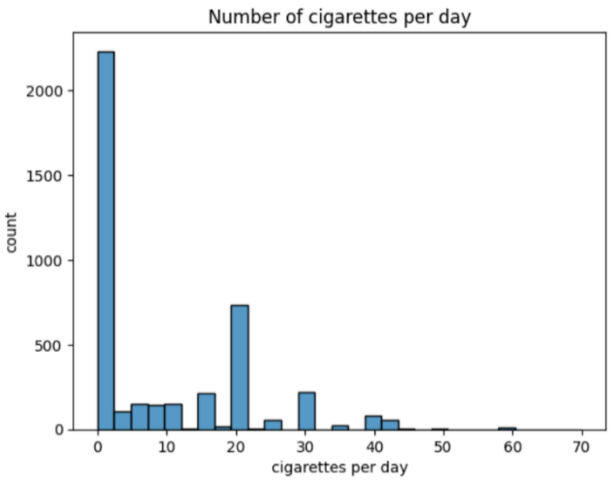
The observed **10-year CHD incidence was 15.2%**, meaning that approximately one in six individuals in the dataset developed coronary heart disease within ten years. This level of incidence is consistent with reported rates from large population-based cohort studies in the United States, supporting the external validity of the dataset.

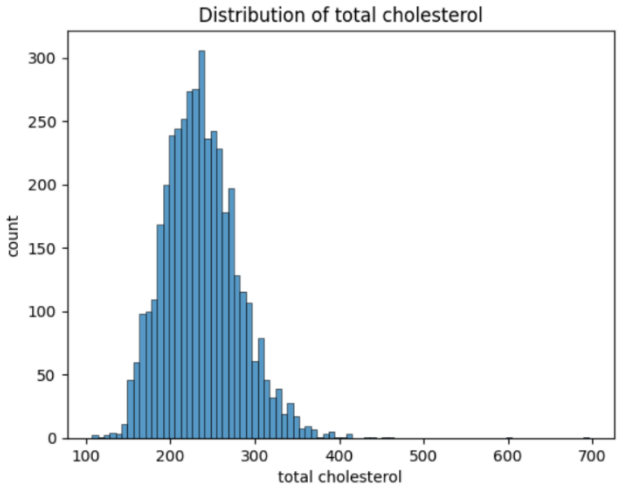
Because the target variable is imbalanced, with substantially more non-CHD cases than CHD cases, this imbalance may reduce the model's ability to correctly identify individuals who develop heart disease. Therefore, class imbalance was considered in later stages of the analysis, particularly during model training and evaluation. [6], [11]

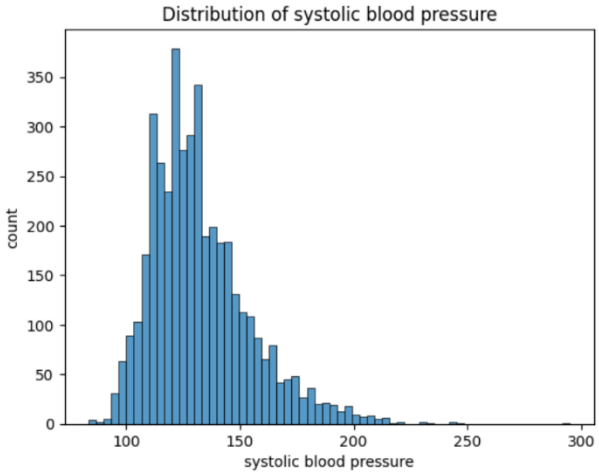
5.2. Univariate analysis of numerical features:

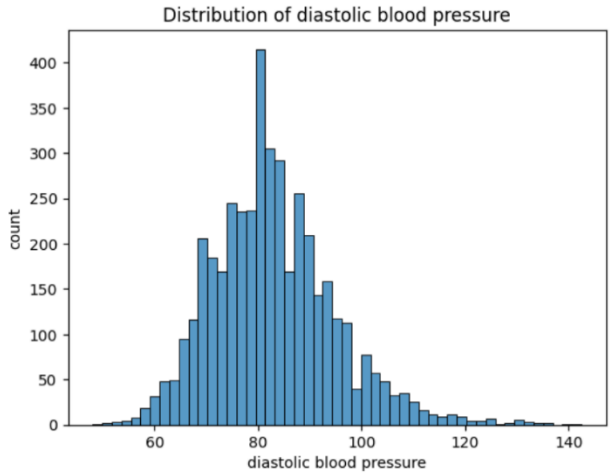
Histograms were also used to show the distribution of each numerical variable in the dataset, such as age, blood pressure, BMI and cholesterol levels. These plots also allow for a visual characterization of the range, central tendency, and distribution of data as well as the presence of skewness or outliers. Understanding these distributions is an important first step in exploratory data analysis, as it can identify potential data problems, guide the selection of transformations or standardization, and provide insight into the way each feature may interact with the outcome variable TenYearCHD.

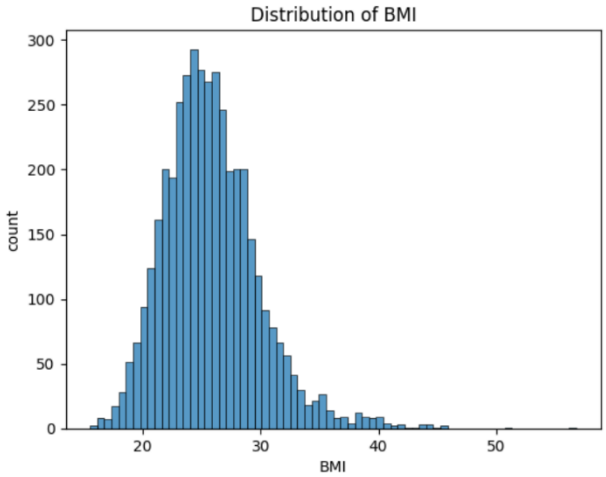


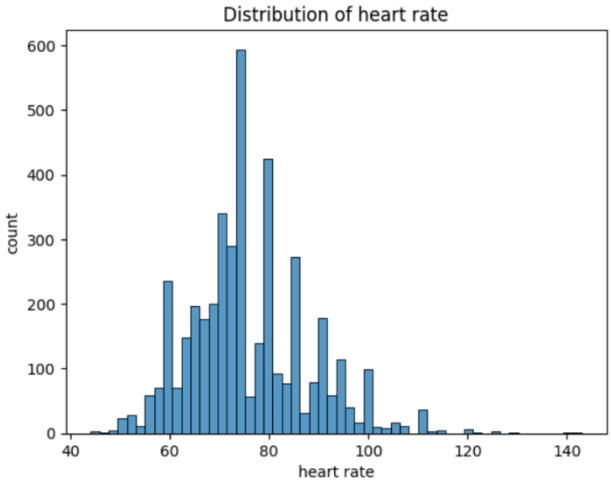


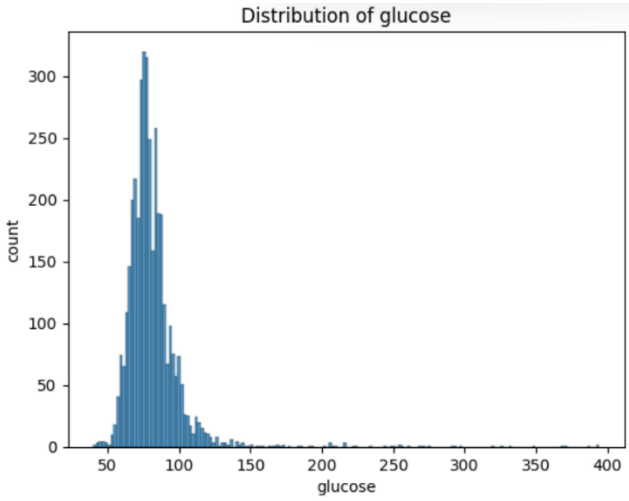












Key observations:

- **Age** shows a minimal right-skewed distribution, with most patients being middle-aged. No extreme values observed and are medically reasonable.
- **Cigarettes per day:** strong right skew with many zero values, medically plausible representing non and heavy smokers.
- **Total cholesterol:** slight skew, some high extreme values, medically plausible.
- **Glucose:** strong right skew, several extreme high values, medically plausible.
- **Systolic blood pressure** shows right-skewed distributions with some extreme values indicative of severe hypertension. Medically reasonable.
- **Diastolic blood pressure** approximately normal with mild right skew and few high extremes, medically reasonable.
- **BMI** values are mostly within a medically reasonable range, though a few high values are present.
- **Heart rate** appears approximately normally distributed with occasionally high values. Medically reasonable.

Overall, the numerical variables are medically plausible, but skewness and abnormal values are present in several features.

Table 4: Summary statistics table

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
age	4240	50	9	32	42	49	56	70
BMI	4221	26	4	16	23	25	28	57
sysBP	4240	132	22	84	117	128	144	295
diaBP	4240	83	12	48	75	82	90	143
totChol	4190	237	45	107	206	234	263	696
glucose	3852	82	24	40	71	78	87	394
heartRate	4239	76	12	44	68	75	83	143
cigsPerDay	4211	9	12	0	0	0	20	70

5.3. Univariate analysis of categorical features:

Counts and proportions were calculated for each binary and categorical variable, such as smoking status, diabetes, BP medication use, and prevalence of hypertension. This analysis provides insight into the distribution of categorical features and highlights class imbalances that may exist in the dataset. Understanding these distributions is important for interpreting model performance, selecting appropriate evaluation metrics, and informing any necessary preprocessing steps, such as encoding or balancing techniques.

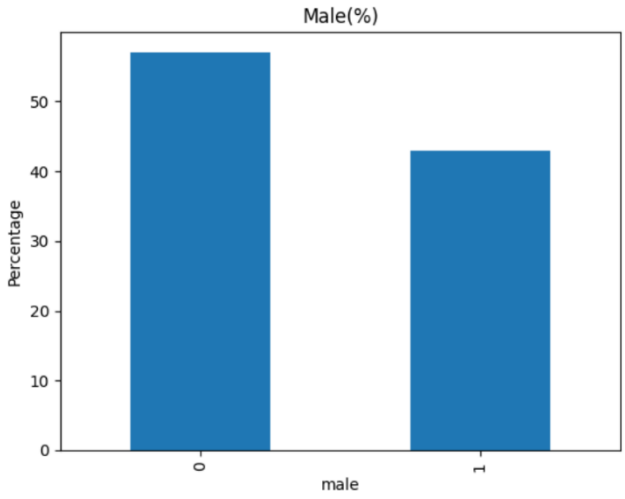
Table 5: Distribution of categorical variables in the Framingham dataset

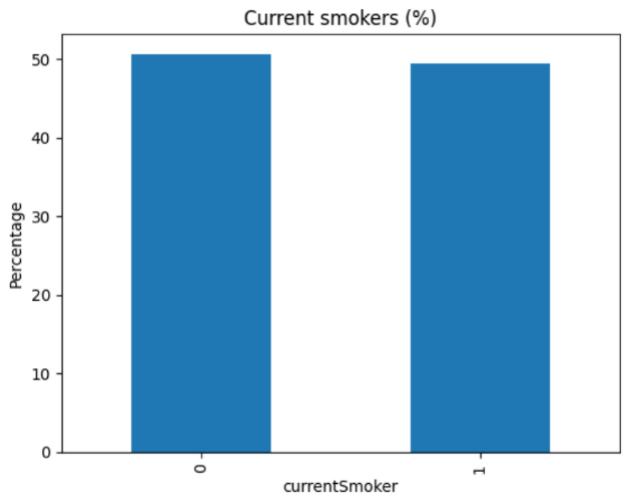
Variable	Category	Count	Percentage (%) (Round off)
male	0 (Female)	2420	57.1
	1 (Male)	1820	42.9
currentSmoker	0 (No)	2145	50.6
	1 (Yes)	2095	49.4
BPMeds	0 (No)	4063	97
	1 (Yes)	124	3
prevalentStroke	0 (No)	4215	99.4
	1 (Yes)	25	0.6
prevalentHyp	0 (No)	2923	68.9
	1 (Yes)	1317	31.1
diabetes	0 (No)	4131	97.4
	1 (Yes)	109	2.6
TenYearCHD	0 (No)	3596	84.8
	1 (Yes)	644	15.2

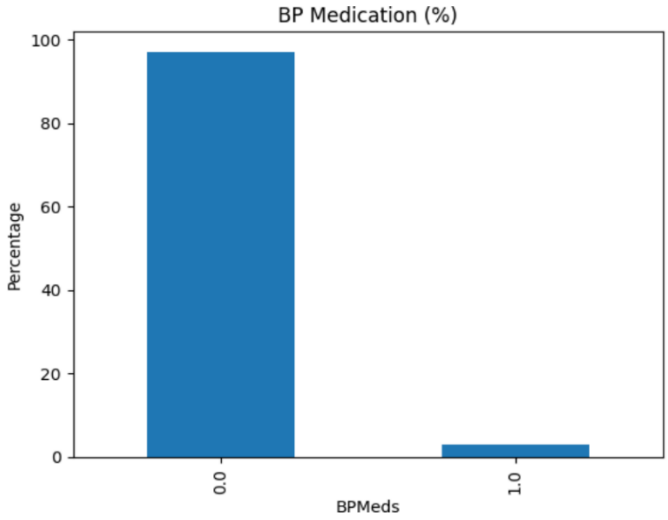
Variable	1 (high school or less)	2 (high school graduate)	3 (college)	4 (college graduate or more)
education	1720	1253	689	473

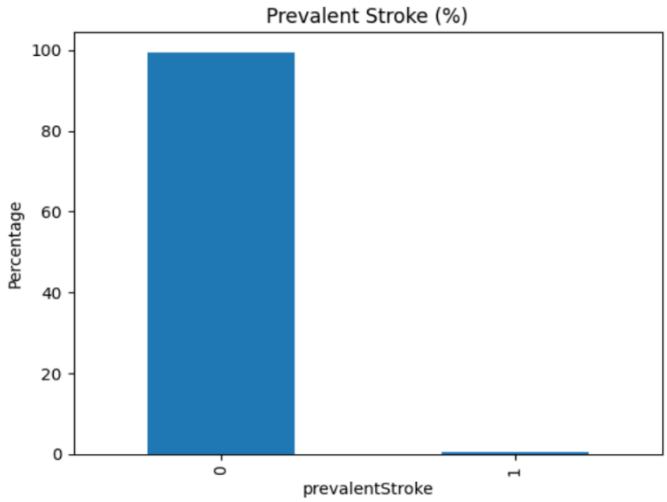
Visualization:

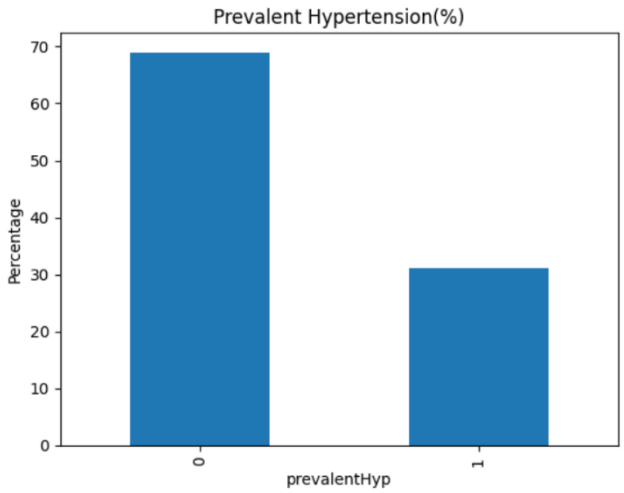
For binary variables, simple bar charts were used to compare the relative frequency of each category. In addition, both a bar chart and a box-and-whisker plot were used for the education variable, which is ordinal in nature, to compare the distribution of educational levels across CHD outcome groups.

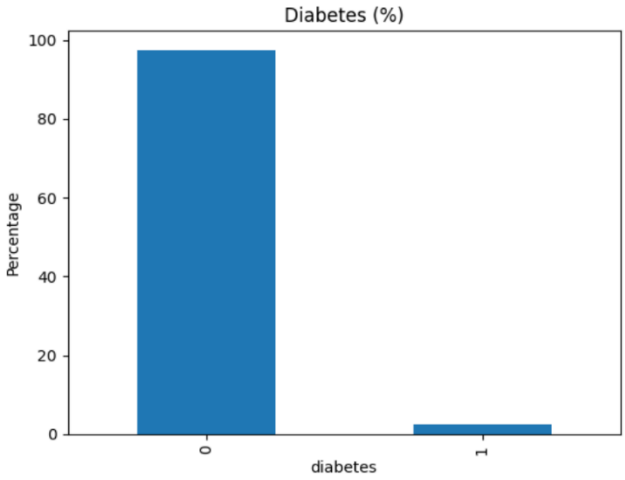


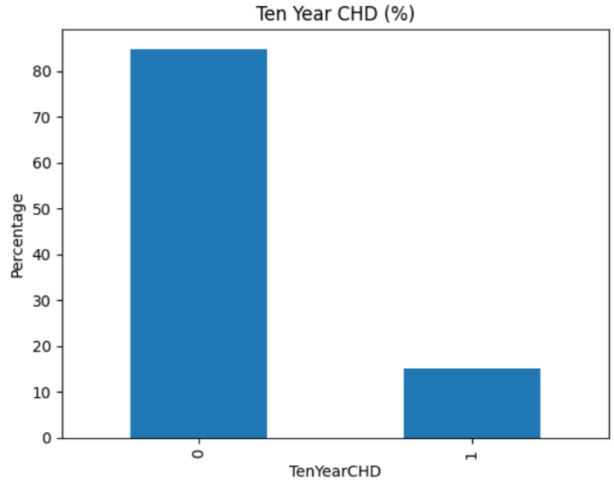




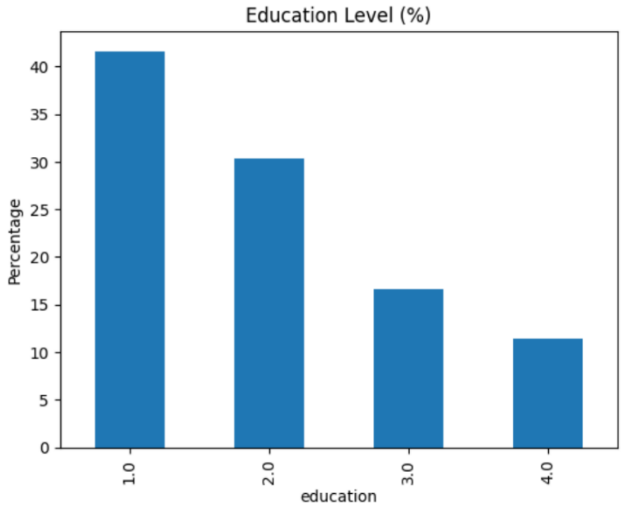


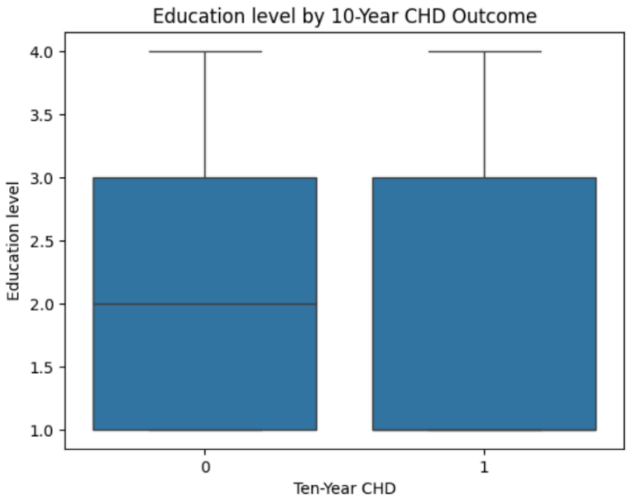






Special case: (ordinal)



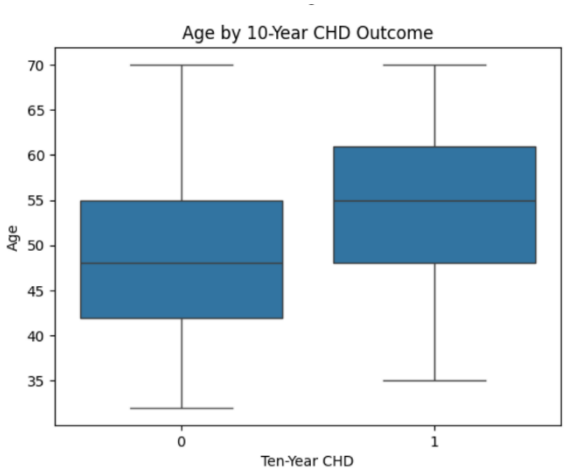


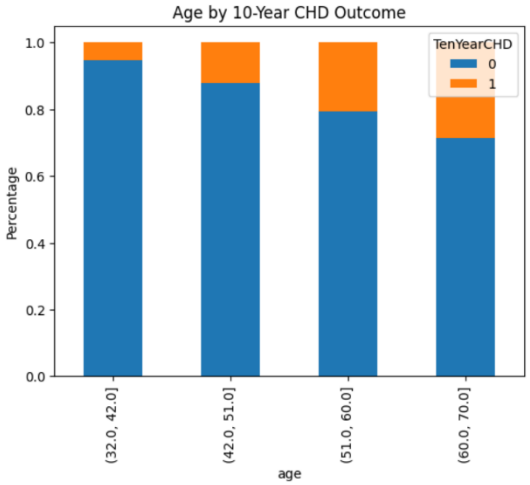
Key observations:

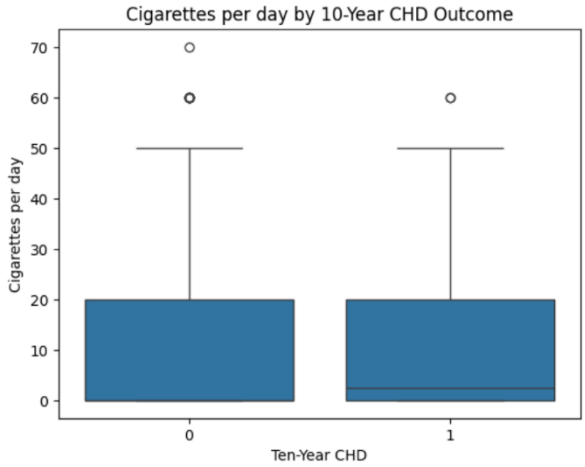
- **Male**:
  There are more females (0 = 2420) than males (1 = 1820) in the dataset. The distribution is moderately unbalanced, with females forming the majority.
- **Current Smoker**:
  The number of non-smokers (2145) and current smokers (2095) is almost equal. This variable is well balanced, indicating similar representation of both categories.
- **BPMeds**:
  A very large majority of individuals are not on blood pressure medication (4063)**,** while only a small number are on medication (124). This variable is highly imbalanced**.**
- **Prevalent Stroke**:
  Most individuals have no history of stroke (4215)**,** while only 25 individuals report a previous stroke. This variable is extremely imbalanced**.**
- **Prevalent Hypertension**:
  Most individuals do not have hypertension (2923)**,** but a substantial number do have hypertension (1317). Although imbalanced, both categories are reasonably represented.
- **Diabetes**:
  Diabetes is relatively rare in the dataset, with 4131 non-diabetic individuals and only 109 diabetic individuals. This variable is highly imbalanced.
- **TenYearCHD**:
  Most individuals did not develop heart disease (3596) within 10 years, while 644 individuals did.
- **Education**:
  Lower education levels are more prevalent with education level 1 forming the largest group (1720). This indicates a moderately imbalanced distribution. From the box plot, both CHD-negative and CHD-positive groups show a similar median education level (around level 2) with largely overlapping distributions.
  Overall, education level shows limited variation between CHD groups, showing that it is a weak indicator of 10-year CHD risk.

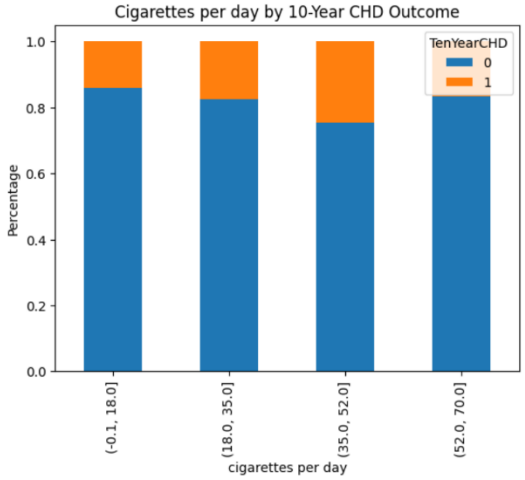
5.4. Stratified analysis by CHD outcome:

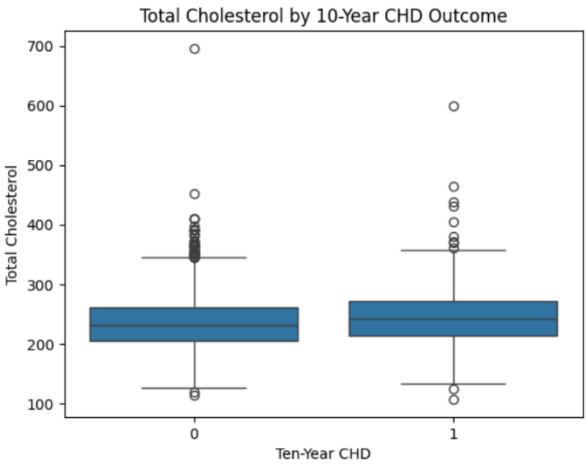
Stratified analysis was used to examine differences between those who did and did not develop 10-year CHD. An analysis of the distributions across CHD groups using box-and-whisker plots showed differences in median values, variability and outliers. The ranges of numbers were also divided into bins and bar charts showed the number or percentage of participants with and without CHD in stacked bar charts for each bin. It allows a visualization of the changes in CHD risk across the range of each numerical variable.

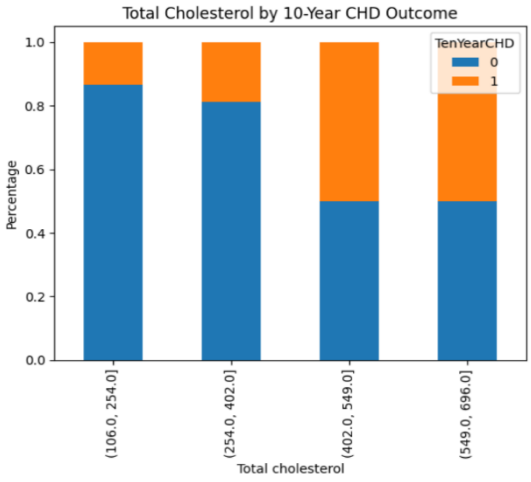

Age by 10-Year CHD Outcome
Age
Ten-Year CHD
0
1
35
40
45
50
55
60
65
70

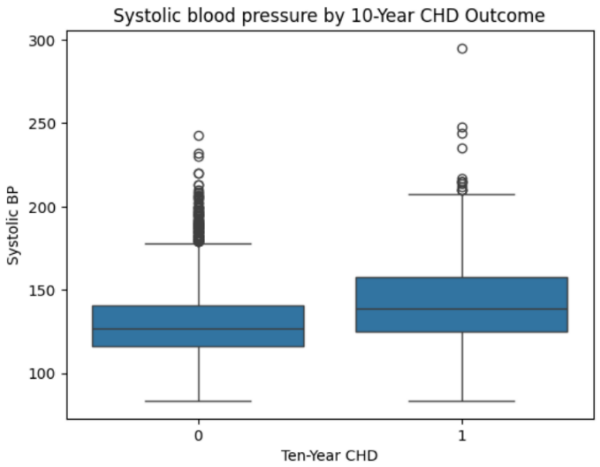

Age by 10-Year CHD Outcome
TenYearCHD
0
1
Percentage
age
(32.0, 42.0]
(42.0, 51.0]
(51.0, 60.0]
(60.0, 70.0]
0.0
0.2
0.4
0.6
0.8
1.0

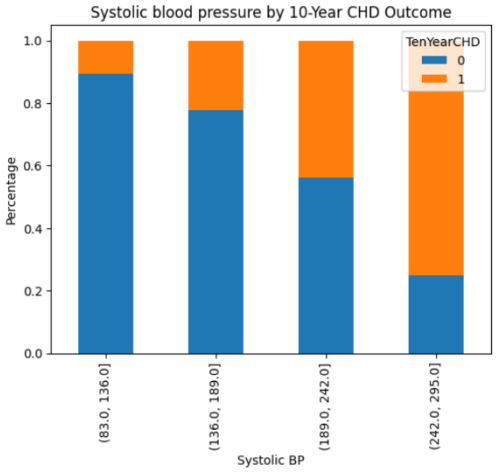

Cigarettes per day by 10-Year CHD Outcome
Cigarettes per day
Ten-Year CHD
0
1
0
10
20
30
40
50
60
70

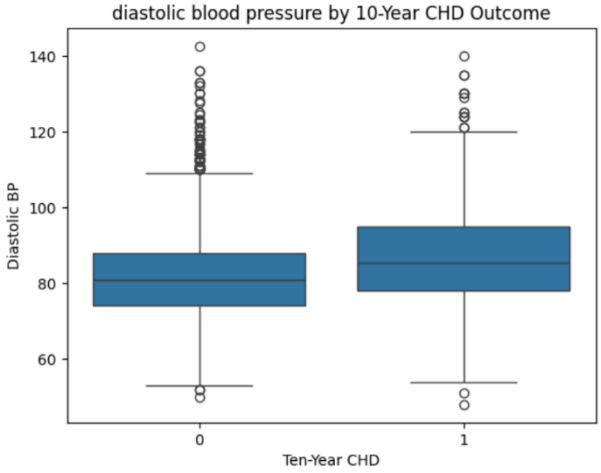

Cigarettes per day by 10-Year CHD Outcome
TenYearCHD
0
1
Percentage
cigarettes per day
(-0.1, 18.0]
(18.0, 35.0]
(35.0, 52.0]
(52.0, 70.0]
0.0
0.2
0.4
0.6
0.8
1.0

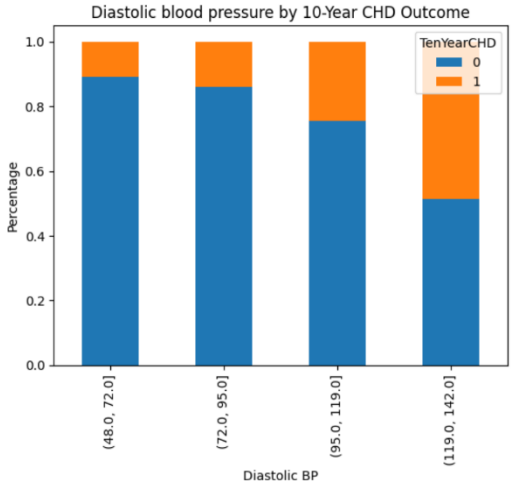

Total Cholesterol by 10-Year CHD Outcome
Total Cholesterol
Ten-Year CHD
0
1
100
200
300
400
500
600
700

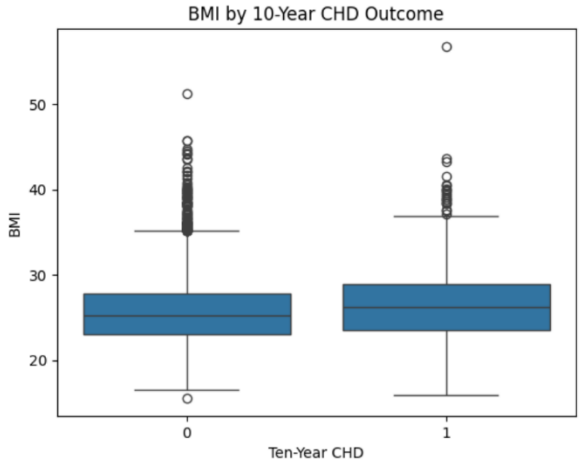

Total Cholesterol by 10-Year CHD Outcome
TenYearCHD
0
1
Percentage
Total cholesterol
(106.0, 254.0]
(254.0, 402.0]
(402.0, 549.0]
(549.0, 696.0]
0.0
0.2
0.4
0.6
0.8
1.0

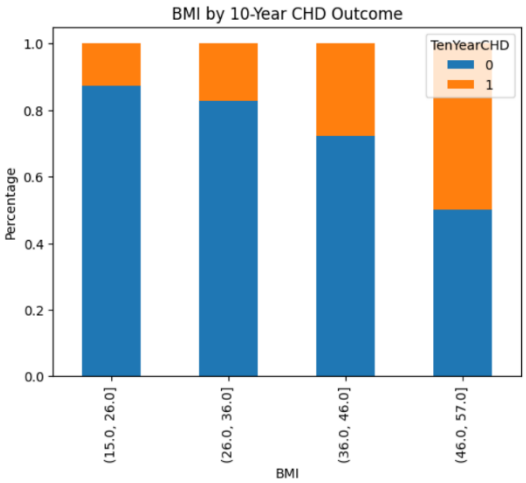

Systolic blood pressure by 10-Year CHD Outcome
Systolic BP
300
250
200
150
100
0
1
Ten-Year CHD

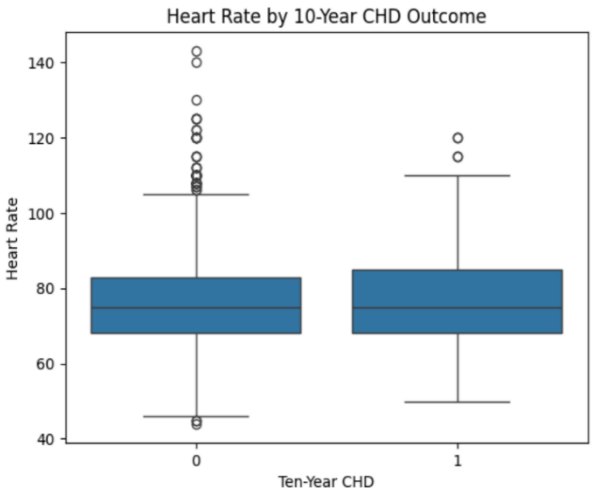

Systolic blood pressure by 10-Year CHD Outcome
TenYearCHD
0
1
Percentage
1.0
0.8
0.6
0.4
0.2
0.0
(83.0, 136.0]
(136.0, 189.0]
(189.0, 242.0]
(242.0, 295.0]
Systolic BP

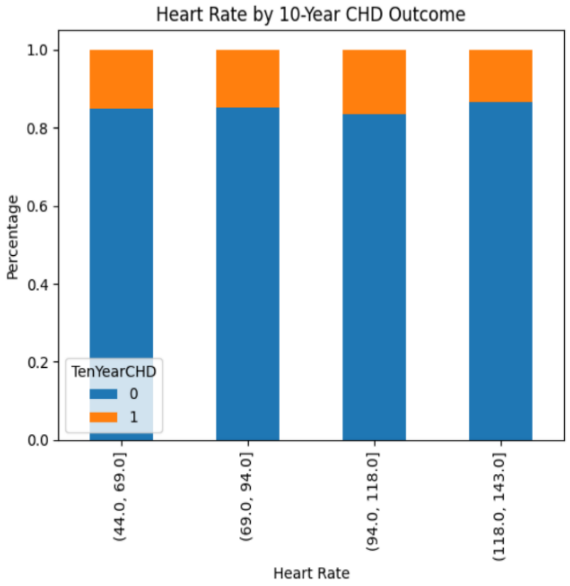

diastolic blood pressure by 10-Year CHD Outcome
Diastolic BP
140
120
100
80
60
0
1
Ten-Year CHD

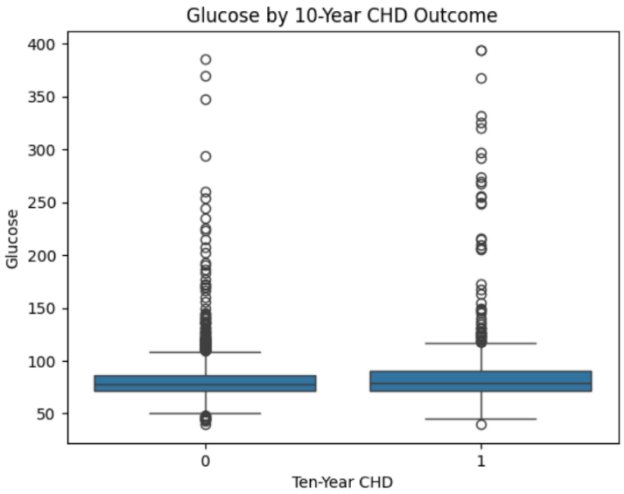

Diastolic blood pressure by 10-Year CHD Outcome
TenYearCHD
0
1
Percentage
1.0
0.8
0.6
0.4
0.2
0.0
(48.0, 72.0]
(72.0, 95.0]
(95.0, 119.0]
(119.0, 142.0]
Diastolic BP

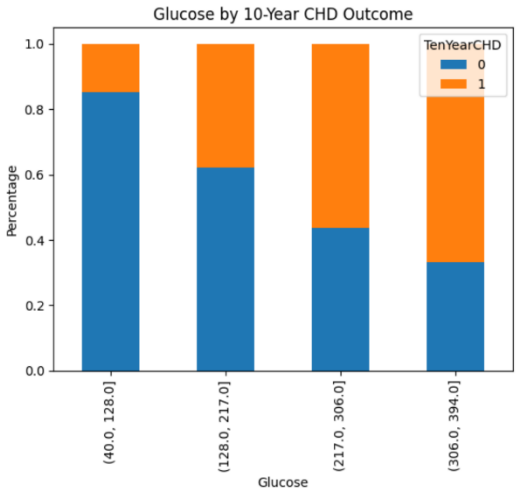

BMI by 10-Year CHD Outcome
BMI
50
40
30
20
0
1
Ten-Year CHD


BMI by 10-Year CHD Outcome
TenYearCHD
0
1
Percentage
1.0
0.8
0.6
0.4
0.2
0.0
(15.0, 26.0]
(26.0, 36.0]
(36.0, 46.0]
(46.0, 57.0]
BMI









Key observations:

1. **Age:**
   - Participants who developed CHD (TenYearChd = 1) are generally older than those who did not.
   - Median age is higher in the group of participants that developed CHD and the interquartile range is slightly shifted right.
   - Older age bins show higher proportions of CHD.
2. **Cigarettes per day:**
   - Median cigarettes per day are slightly higher in the CHD-positive group, but distributions overlap.
   - Most participants in both groups smoke very few or no cigarettes, but a higher consumption shows a larger proportion of CHD-positive participants.

- Presence of a few outliers visible in both groups.

3. **Total cholesterol:**
   - Median total cholesterol is slightly higher in the CHD-positive group.
   - There are many extreme values in both, but the proportion of CHD-positive participants is higher in higher cholesterol bins.
   - Lower and middle cholesterol bins have mostly CHD-negative participants.
4. **Systolic blood pressure:**
   - Median systolic BP is higher in the CHD-positive group.
   - Lower systolic BP categories are dominated by CHD-negative participants, while higher categories show an increasing proportion of CHD-positive cases.
   - Several outliers are present, particularly in the CHD-positive group.
5. **Diastolic blood pressure:**
   - Median diastolic blood pressure is slightly higher in the CHD-positive group, though distributions overlap.
   - The proportion of CHD-positive individuals rise with increasing diastolic BP.
   - A few outliers are present in both groups, with greater variability in CHD-positive group.
6. **Body Mass Index (BMI):**
   - Lower BMI categories are predominantly CHD-negative, while higher BMI categories show a larger proportion of CHD-positive participants.
   - The highest BMI range has nearly equal representation of CHD-positive and CHD-negative individuals, suggesting increased CHD risk at higher BMI levels.
   - Several high BMI outliers in both groups.
7. **Heart Rate:**
   - Median heart rate is slightly higher in CHD-positive group, though distributions largely overlap.
   - The categories show only minor variation in the proportion of CHD-positive individuals across heart rate levels.
   - Several outliers with high heart rates are present in both groups suggesting that heart rate alone is a weak indicator of CHD risk.
8. **Glucose:**
   - Median glucose levels are higher in the CHD-positive group, with a wider distribution.
   - Lower glucose categories are dominated by CHD-negative participants.
   - Numerous outliers present particularly among CHD-positive participants, indicating a strong association between high glucose levels and CHD risk.

Grouped summary:

Mean and median values of numerical variables were computed for participants with and without 10-year CHD to quantify differences in the plot.

Table 6: Mean values of numerical features stratified by 10-year CHD Outcome

TenYearCHD	0	1
age	49	54
cigsPerDay	9	11
totChol	235	245
sysBP	130	144
diaBP	82	87
BMI	26	27
heartRate	76	77
glucose	81	89

The summary table organized by **mean values** illustrates distinct differences between individuals who developed CHD and those who did not. Typically, people in the CHD group are older and exhibit elevated systolic and diastolic blood pressure, increased BMI, and greater total cholesterol levels relative to the non-CHD group. These trends align with known cardiovascular risk factors and indicate that elevated levels of these variables correlate with an increased chance of developing CHD within 10 years.

Table 7: Medan values of numerical features stratified by 10-year CHD Outcome

TenYearCHD	0	1
age	48	55
cigsPerDay	0	3
totChol	232	241
sysBP	127	139
diaBP	81	86
BMI	25	26
heartRate	75	75
glucose	78	79

The grouped summary table based on **median values** confirms the trends observed in the mean-based analysis. Median age, blood pressure, BMI, and cholesterol levels are generally higher in the CHD group than in the non-CHD group. The similarity between mean and median patterns indicates that the observed differences are not driven solely by extreme values or outliers but represent consistent differences between the two groups.

5.5. Correlation analysis among numerical variables:

A correlation matrix was computed for all numerical variables to examine relationships between variables such as age, blood pressure, cholesterol, etc. The following heatmap highlights the strength of these associations.

The correlation analysis reveals several moderates to strong positive relationships among the variables.

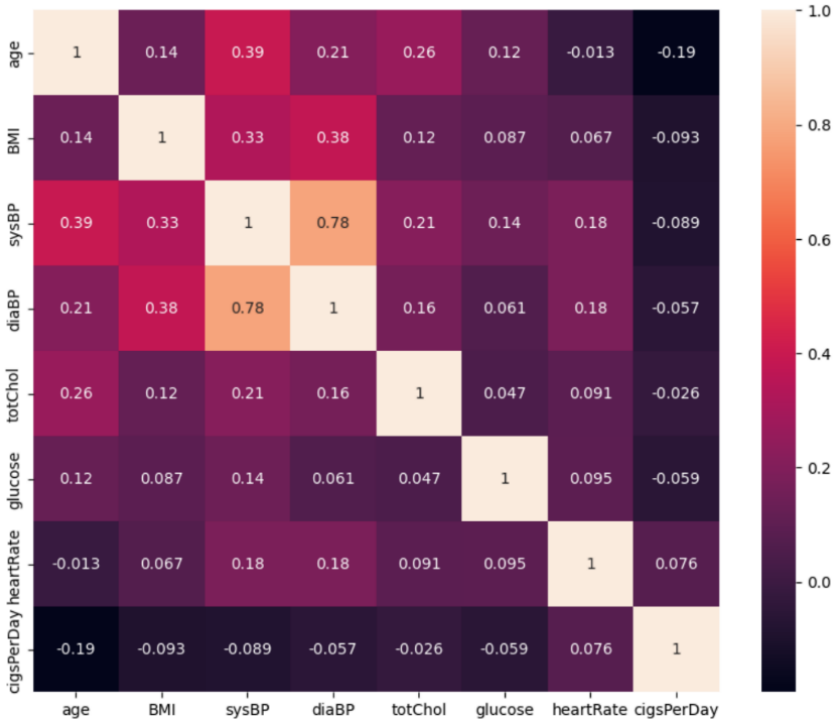
The strongest association is observed between systolic blood pressure and diastolic blood pressure, indicating that individuals with higher systolic blood pressure also tend to have higher diastolic blood pressure. This is expected, as both measures represent closely related components of blood pressure.

Age shows a moderate positive correlation with systolic blood pressure, suggesting that systolic blood pressure tends to increase with advancing age, consistent with known age-related cardiovascular changes.

Body Mass Index (BMI) exhibits moderate positive correlations with both systolic and diastolic blood pressure, indicating that higher BMI is associated with elevated blood pressure levels.

Overall, these correlations are clinically meaningful and align with established cardiovascular risk patterns. The strong relationship between systolic and diastolic blood pressure reflects the close physiological link between the two measurements.

Figure: Correlation matrix



Chapter 6: Model selection (Model Stage)

6.1. Binary classification problem formulation:

The modeling task in this study is formulated as a binary classification problem, where the objective is to predict whether an individual will develop coronary heart disease (CHD) within a ten-year period. The outcome variable, TenYearCHD, takes a value of 1 if CHD occurs within ten years and 0 otherwise. Predictor variables include demographic characteristics, lifestyle behaviors, and clinical measurements such as age, sex, smoking status, blood pressure, cholesterol levels, BMI, diabetes, and medication use. This formulation enables the use of supervised learning techniques to estimate individual-level CHD risk probabilities.

6.2. Justification for logistic regression:

Logistic regression was selected as the primary modeling approach because it is a well-established and widely adopted statistical method for predicting binary outcomes, particularly in epidemiological and medical research [2], [6], [11]. It has a long history of use in cardiovascular risk modeling, including in analyses derived from the Framingham Heart Study, making it a natural and methodologically consistent choice for this task.

The model predicts the likelihood of an event happening by representing the log-odds of the result as a linear function of the input variables. This formulation enables logistic regression to identify the connection between established cardiovascular risk factors and the probability of developing coronary heart disease over a designated period. In contrast to models that generate solely class labels, logistic regression offers probability estimates, which are especially important in healthcare settings where risk stratification and decision thresholds might differ based on clinical or public health objectives.

Logistic regression is appropriate for the Framingham teaching dataset's structure, as it handles both continuous and binary predictors effectively. The model makes fewer assumptions than more complex machine learning techniques, which decreases the likelihood of overfitting in medium-sized datasets and improves the reliability of estimates. Moreover, class imbalance frequently found in disease outcome data can be tackled directly with methods like class weighting, which was utilized in this research.

Significantly, logistic regression provides a strong level of interpretability, which is crucial in medical risk assessment. Model coefficients can be converted into odds ratios, enabling a clear quantification and comparison of both the direction and strength of each predictor's relationship with CHD risk. This transparency allows results to be assessed in relation to recognized epidemiological data and fosters trust and understanding for clinical or public health representatives.

Due to these factors, logistic regression offers a robust, interpretable baseline model for predicting 10-year CHD risk. Although more sophisticated machine learning models can provide slight improvements in performance, logistic regression strikes a balance between predictive capability and transparency, interpretability, and methodological robustness, making it particularly suitable for analyses centered on education and health.

6.3. Advantages for medical risk modeling:

Logistic regression provides multiple benefits when applied to medical risk modeling. Firstly, the model offers a high level of interpretability, enabling coefficients to be readily converted into odds ratios, which are well-known and significant to clinicians and public health researchers. Additionally, it facilitates transparent decision-making, which is crucial in healthcare, where clarity is vital. Third, logistic regression shows consistent performance on moderately sized datasets and is resilient to correlated predictors if multicollinearity is not intense. Finally, its simplicity and statistical basis makes it well suited for estimating baseline risk and supporting early risk stratification instead of replacing clinical diagnosis.

Chapter 7: Logistic regression background

7.1. Logistic (sigmoid) function:

Logistic regression models the probability of a binary outcome using the logistic (sigmoid) function, which maps any real-valued input to a value between 0 and 1, making it suitable for probability estimation. The sigmoid function is defined as:

$$\frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

This function converts the linear combination of the independent variables into a probability, ensuring that the model's predictions remain within the $[0, 1]$range required for binary outcome modeling.

By estimating the coefficients (β values) associated with each predictor, logistic regression quantifies how changes in the independent variables affect the probability of the outcome variable taking value 1, while holding all other variables constant. [12], [13]

7.2. Probability interpretation:

The result of logistic regression indicates the estimated likelihood that a person is part of the positive class (for instance, experiencing CHD in a decade). Expected probabilities lie between 0 and 1 and can be transformed into class labels by applying a selected decision threshold, typically set at 0.5. This probabilistic approach is especially useful in healthcare contexts, where grasping the degree of risk usually matters more than simply making a yes-or-no choice.

7.3. Log-odds formulation:

In logistic regression, the relationship between the predictors and the outcome is modeled through the odds of the event occurring. If p denotes the probability that the outcome variable equals 1 (e.g., development of CHD within 10 years), the odds of success are defined as:

$$\text{Odds} = \frac{p}{1-p}$$

Logistic regression assumes that the logarithm of these odds, known as the log-odds, is a linear function of the independent variables:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Here, β_0 represents the intercept, β_i are the model coefficients, and x_i are the predictor variables. This formulation allows logistic regression to model nonlinear relationships between predictors and probabilities while maintaining linearity in the model parameters.

By exponentiating both sides of the equation, the odds can be expressed as:

$$\frac{p}{1-p} = e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}$$

This indicates that every coefficient signifies the multiplicative change in the odds of the outcome for a one-unit increment in the related predictor, while keeping all other variables fixed. Log-odds formulation is crucial for understanding logistic regression and serves as the foundation for odds ratio analysis in medical and epidemiological research. [10], [13]

7.4. Coefficients and odds ratios:

Each coefficient in a logistic regression model represents the change in the log-odds of the outcome associated with a one-unit increase in the predictor, holding all other variables constant. Exponentiating the coefficients yields odds ratios:

$$\text{Odds Ratio} = e^{\beta}$$

Odds ratios provide an interpretable measure of effect size, commonly used in epidemiological studies. Values greater than one indicate increased odds of the outcome, while values less than one indicate reduced odds. This interpretability makes logistic regression particularly suitable for medical risk modeling.

7.5. Maximum likelihood estimation and loss function:

The parameters in logistic regression are generally determined through maximum likelihood estimation (MLE), which selects coefficient values that optimize the likelihood of witnessing the observed outcomes according to the model. For independent observations, this results in a log-likelihood function that aggregates contributions from every subject, and maximizing this log-likelihood corresponds to minimizing the negative log-likelihood, commonly referred to as logistic loss or cross-entropy loss. Due to the absence of a closed-form solution, numerical optimization methods (like gradient descent or iteratively reweighted least squares) are employed to determine the coefficient estimates. [6], [13]

Chapter 8: Model training

8.1. Train-test split strategy:

The dataset was split into training and test sets using a 75%/25% split. To preserve the proportion of individuals who developed CHD in both sets, stratified sampling was applied using the target variable. Because CHD cases make up only 15% of the aggregate, stratification is especially crucial in this dataset. By preserving this distribution, it is ensured that the training and test sets appropriately represent the underlying class imbalance.

Imputation of missing values was performed after the train–test split. Since the mode (for categorical variables) or median (for numerical variables) used to fill in the missing values in the training set is calculated solely from the training data, this method avoids data leakage. To ensure consistency, the test set is then given the same values.

8.2. Stratified sampling:

In order to maintain the same percentage of people who developed CHD in both sets, stratified sampling was used during the train–test split. Simple random sampling might have produced a test set that under- or over-represents the minority class due to the imbalanced target variable TenYearCHD, which accounts for 15% of cases of CHD. This could have skewed model evaluation and degraded predictive performance.

Each subset maintains the same class distribution as the original dataset by stratifying on the outcome variable. This method guarantees that the test set is suitable for objective assessment of model performance, while the training set offers a representative sample for model fitting. Stratification is a common procedure in both supervised machine learning workflows and epidemiological studies, and it is especially crucial in medical datasets where rare events are frequently of primary interest.

8.3. Handling class imbalance (class weighting):

As identified during exploratory data analysis, the target variable TenYearCHD is With only 15.2% of people developing CHD over the ten-year follow-up, the target variable TenYearCHD is unbalanced, as revealed by exploratory data analysis. This indicates that about one in six people had a CHD event, whereas most people did not. The model may be less able to accurately identify members of the minority class as a result of this imbalance, which could skew predictions in favour of the majority class.

Class weighting was used in the logistic regression model to address this. This approach effectively penalises misclassification of CHD events more severely by giving the minority class (CHD cases) higher weights during training and the majority class (non-CHD cases) lower weights. Class_weight='balanced' is used in scikit-learn to implement this, automatically modifying the weights based on class frequencies.

By incorporating class weighting, the model pays proportionate attention to both classes, improving sensitivity for detecting CHD cases without adversely affecting overall performance. This approach is particularly important for medical datasets where correctly identifying rare but clinically significant events is critical.

8.4. Model fitting procedure:

The logistic regression model was trained on the training dataset using L2 regularization (default) to prevent overfitting. Regularization constrains the magnitude of the coefficients, helping the model generalize better to unseen data and reducing the risk of relying excessively on any single predictor.

During training, class weighting was applied (class_weight='balanced') to account for the imbalance in CHD cases. This ensures that the minority class receives

proportionally greater emphasis during model fitting, improving the model's ability to detect individuals who develop CHD.

Once trained, the model outputs probabilities between 0 and 1 for everyone, representing their estimated 10-year CHD risk. These probability scores can subsequently be transformed into class predictions using different thresholds, allowing flexibility to prioritize either minimizing false positives or maximizing detection of CHD cases, depending on clinical or public health objectives.

Chapter 9: Model evaluation

9.1. Evaluation metrics:

The model was evaluated using multiple metrics at the standard threshold of 0.5.

Table 8: Model performance metrics at Threshold=0.5

Metric	**Value**
ROC-AUC	0.70
Precision	0.25
Recall	0.60
F1-score	0.36
Accuracy	0.67

Table 9: Confusion Matrix of the Logistic Regression Model at Threshold=0.5

	CHD = 0 (predicted)	**CHD = 1 (predicted)**
CHD = 0 (actual)	493	226
CHD = 1 (actual)	52	77

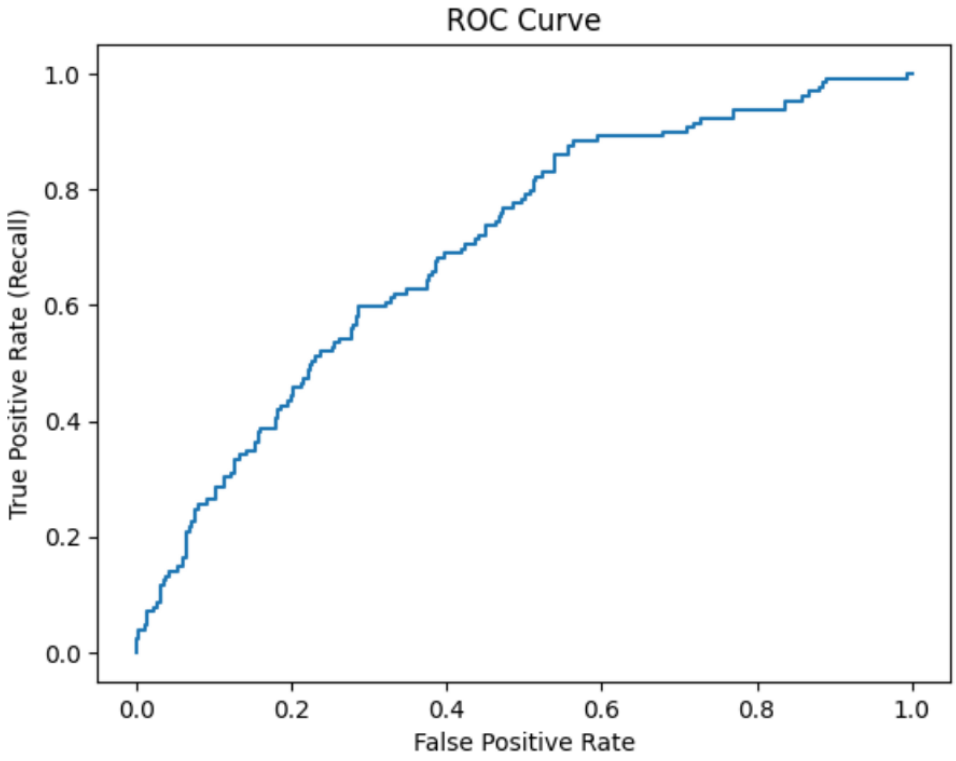
Explanation of metrics:

- **ROC–AUC** measures how well the model separates CHD from non-CHD cases.
- **Precision** indicates how many of the predicted CHD cases were actually CHD.
- **Recall** shows how many actual CHD cases were correctly identified.
- **F1-score** is the harmonic mean of precision and recall, balancing both metrics.
- **Accuracy** gives the overall proportion of correct predictions.

Accuracy alone is not sufficient because CHD is less frequent (15%). A model could predict most patients as non-CHD and appear highly accurate while missing many true CHD cases. Metrics like precision, recall, and F1-score provide a more meaningful assessment.

9.2. ROC curve and AUC:

The effectiveness of the logistic regression model was additionally assessed through the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) metric. The ROC curve displays the true positive rate (recall) versus the false positive rate for different classification thresholds, offering an extensive perspective on model performance across all potential decision boundaries.

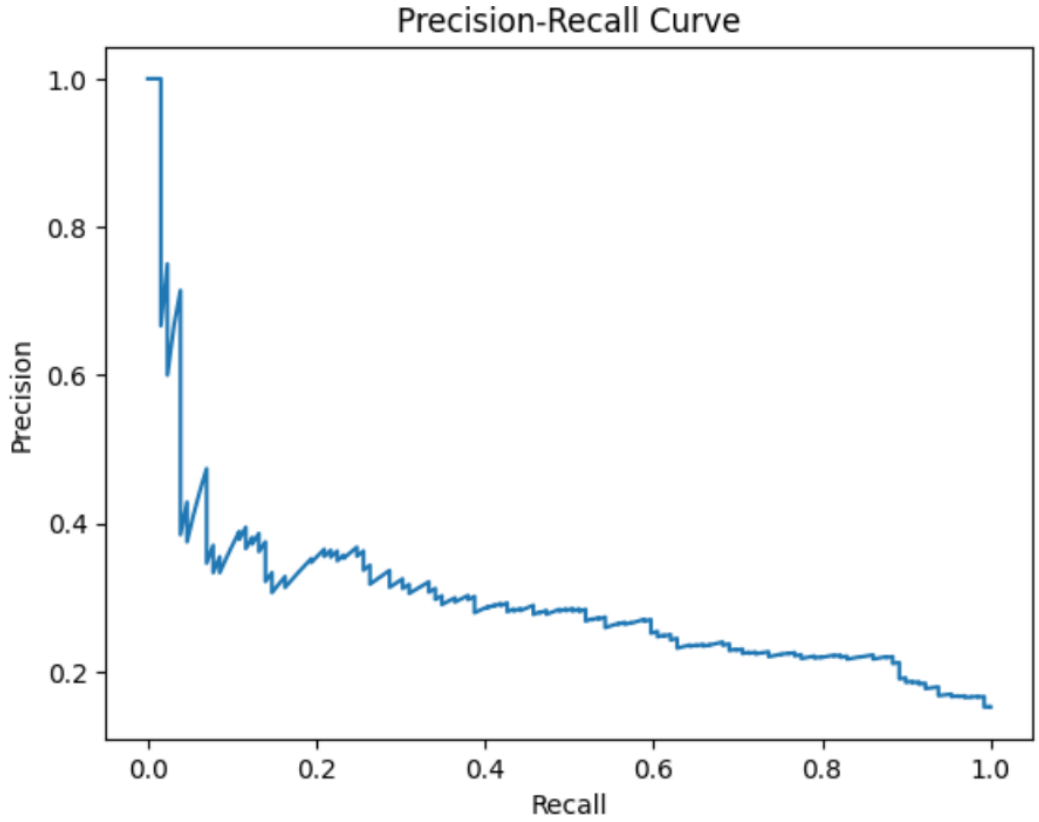


Through cross-validation on the training dataset, the model obtained a mean AUC of 0.724, accompanied by a standard deviation of 0.012. This shows that the model can differentiate between individuals who develop CHD and those who do not, significantly better than random chance, while the low standard deviation indicates consistent performance across various folds of the data.

A ROC AUC greater than 0.7 is typically seen as indicative of satisfactory discriminative power in clinical prediction models. A value of 0.5 represents random guessing, whereas 1.0 signifies flawless discrimination. Consequently, the ROC curve offers a strong assessment of the model's capability to rank individuals based on their 10-year CHD risk, without relying on any particular threshold.

9.3. Precision-Recall curve:

The precision–recall (PR) curve evaluates model performance across all possible classification thresholds, highlighting the trade-off between precision (the proportion of predicted CHD cases that are true positives) and recall (the proportion of actual CHD cases correctly identified). This curve is particularly informative for imbalanced datasets, such as the Framingham teaching dataset, where CHD cases account for only ~15% of the observations.



The average precision of the logistic regression model was 0.30, while the baseline CHD prevalence was 0.15. This suggests that when it comes to identifying people at risk for CHD, the model outperforms random guessing by a significant margin.

The inherent trade-off in medical risk prediction is reflected in the expected decline in precision as recall rises: capturing more true CHD cases (higher recall) inevitably introduces additional false positives, lowering precision. In clinical and public health settings, this trade-off is acceptable because emphasizing high recall guarantees that high-risk individuals are identified for additional assessment, even if some false positives are produced.

As a result, the PR curve offers a useful illustration of how the model strikes a balance between the risk of false positives and sensitivity to CHD cases across various probability thresholds. The model's usefulness as a risk prediction tool is validated by

the average precision score above baseline, which shows that it consistently separates CHD cases from non-cases.

9.4. Threshold analysis:

The logistic regression model outputs probabilities of CHD for each individual, which can be converted into class predictions by applying a decision threshold. To examine the effect of threshold selection, three thresholds were tested, and the resulting precision, recall, accuracy, and ROC-AUC were recorded:

Table 10: Performance metrics at different probability thresholds

Threshold	Precision	Recall	Accuracy	ROC-AUC
0.4	0.22	0.76	0.56	0.70
0.5	0.25	0.60	0.67	0.70
0.6	0.28	0.43	0.75	0.70

As observed, lowering the threshold results in more false positives while increasing recall and decreasing precision, enabling the model to identify a higher percentage of CHD cases. On the other hand, raising the threshold prioritizes accurate identification of positive cases while overlooking some actual CHD events, improving precision but decreasing recall. The model's capacity to rank people by risk regardless of any cutoff is demonstrated by the ROC-AUC, which stays constant across thresholds. Clinical or public health priorities should therefore be taken into consideration when choosing thresholds. While higher thresholds can minimize needless follow-ups or interventions, lower thresholds may be preferred to identify more high-risk individuals for early intervention.

Chapter 10: Model interpretation (interpret stage)

10.1. Model coefficients:

The coefficients of the logistic regression model were extracted to quantify the relationship between each variable and the 10-year CHD risk. Positive coefficients indicate that an increase in the predictor is associated with higher odds of CHD, while negative coefficients indicate a protective effect. To facilitate interpretation, the coefficients were converted to odds ratios by exponentiation, where values greater than 1 represent increased risk and values less than 1 represent decreased risk.

The most influential variables, ranked by the magnitude of their coefficients, are summarized in the following table:

Table 11: Coefficients and odds ratios for predictors of CHD

Feature	Coefficient	Odds Ratio
prevalentStroke	0.93	2.54
BPMeds	0.53	1.69
male	0.40	1.49
diabetes	0.32	1.37
prevalentHyp	0.15	1.16
age	0.07	1.07
currentSmoker	0.06	1.07
cigsPerDay	0.02	1.02
sysBP	0.01	1.01
BMI	0.01	1.01
glucose	0.01	1.01
diaBP	0.00	1.00
totChol	0.00	1.00
heartRate	-0.00	1.00
education	-0.02	0.98

10.2. Odds Ratios:

Odds ratios (ORs), which show the multiplicative change in the likelihood of developing CHD for a one-unit increase in the predictor variable while keeping all other variables constant, were computed by exponentiating the coefficients from the logistic regression model. A predictor is linked to a higher risk of CHD if the OR is greater than 1, while a protective effect exists if the OR is less than 1.

Key observations from the odds ratios include:

- **prevalentStroke (OR = 2.54)**: The strongest risk factor, more than doubling the odds of CHD.
- **BPMeds (OR = 1.69)**: Associated with increased risk, reflecting the impact of hypertension.
- **male (OR = 1.49)**: Being male increases CHD odds.
- **diabetes (OR = 1.37)**: Presence of diabetes elevates risk.
- **age (OR = 1.07 per year)**: Incremental increase in CHD odds with each additional year of age.
- **education (OR = 0.98)**: Slightly protective, reducing CHD odds

Other continuous variables, such as cigsPerDay, sysBP, BMI, and glucose, showed modest increases in risk (OR slightly above 1), whereas variables like diaBP, totChol, heartRate had ORs around 1, indicating negligible effect in the presence of other predictors.

The odds ratios provide a quantitative measure of the size effect for each predictor and serve as a foundation for identifying the most important risk factors in subsequent analysis.

10.3. Important risk factors:

Based on the logistic regression model, the most important predictors of 10-year CHD risk were identified by examining the magnitude of the model coefficients and their clinical relevance. These predictors are consistent with known cardiovascular risk factors and align with patterns observed in the exploratory data analysis (EDA).

Key risk factors include:

1. **Prevalent stroke**: Individuals with a history of stroke had the highest influence on CHD risk, highlighting the link between cerebrovascular and cardiovascular disease.
2. **Hypertension (BPMeds and prevalentHyp)**: Both current use of blood pressure medication and a history of hypertension increased CHD risk, confirming the established role of blood pressure in cardiovascular events.
3. **Age**: Older age was associated with increased risk, reflecting the non-modifiable nature of aging as a primary CHD determinant.
4. **Diabetes**: Presence of diabetes significantly elevated CHD risk, consistent with metabolic contributions to atherosclerosis.
5. **Sex (male)**: According to epidemiological trends, men were more vulnerable than women.
6. **Smoking (currentSmoker and cigsPerDay)**: Smoking behavior contributed to elevated risk, highlighting the impact of lifestyle factors on CHD development.
7. **BMI and other metabolic measures (glucose, cholesterol)**: These variables had modest effects individually but contribute cumulatively to overall risk.

Protective factors were also identified, notably higher education, which was associated with slightly reduced risk, possibly reflecting socio-economic or behavioral influences.

In summary, the model reinforces that traditional clinical and lifestyle risk factors, including age, sex, smoking, diabetes, hypertension, and prior stroke, remain the most influential in predicting 10-year CHD risk. This is consistent with trends observed in the EDA, confirming the internal validity of the model and highlighting variables that could be prioritized in preventive strategies.

10.4. Consistency with EDA:

Comparisons between the EDA plots, including stacked histograms and binned plots, and the logistic regression results show strong consistency. Variables like age, blood pressure, diabetes, and smoking were slightly higher among individuals who developed CHD during the EDA. These same variables have positive coefficients and odds ratios over 1 in the logistic regression model, confirming their link to increased CHD risk. While the distributions overlap significantly and the median differences are modest, logistic regression can combine multiple predictors at once. This helps small individual effects make a meaningful contribution to overall risk prediction when viewed together. It reinforces the model's internal validity and matches known clinical risk factors.

Chapter 11: Results and Discussion

11.1. Summary of model performance:

The logistic regression model had a mean cross, validated AUC of 0.724 (with a standard deviation of 0.012), which suggested it had moderate ability to discriminate between people who developed CHD and those who did not. The precision recall analysis showed that the average precision was 0.30, which is above the baseline CHD prevalence of 0.15, thus indicating that the model is a better performer than random guessing.

Threshold analysis disclosed the predicted tradeoff between precision and recall: lowering thresholds results in higher recall but lower precision, whereas increasing threshold leads to better precision, but some CHD cases are missed. ROC, AUC scores did not vary much at different threshold levels, which means the model's ability to rank is solid.

11.2. Interpretation of Key Predictors:

The most important feature for predicting 10 year risk of CHD were in fact, the prevalent stroke and hypertension (BPMeds and prevalentHyp), age, diabetes, male sex, and smoking behavior, which really makes sense with clinical knowledge and the trends in the EDA. These factors all had positive coefficients, and their odds ratios were above 1. This means that the more of these elements one has, or the more they are present, the more likely it is for a person to develop CHD. On the other hand, a protective factor, for instance, a higher level of education, was associated with an odds ratio slightly less than 1, which can imply the advantage of socio-economic and behavioral changes. Thus, putting together several predictors in the regression model not only allowed

individual modest effects, such as cigsPerDay or BMI, to make a meaningful contribution to risk estimation, but also helped to greater the overall prediction power.

11.3. Comparison with Epidemiological Studies:

The observed 10-year CHD incidence in this dataset was 15.2%, which is generally in line with reported rates for the US population-based cohort studies, thus supporting the external validity of the dataset. The main risk factors found are consistent with the prior epidemiological evidence: age, male sex, hypertension, diabetes, smoking, and history of stroke are among the factors most commonly associated with the risk of coronary heart disease.

This suggests that the model interprets the data in a way that is clinically relevant and quite reasonably provides understandable insights which can be applied in a real-world setting.

11.4. Impact of Missing Data Handling:

After the train test split, missing values in numerical variables were replaced with median values while missing values in categorical variables were imputed using mode. This method safeguards against data leakage and ensures that the imputation parameters are not defined based on the test set. Proper handling of missing data allows the model to be unbiased in coefficient estimation and predictions can be considered valid as model integrity is maintained. Although missing data may influence the size of coefficients, the general patterns and the ranking of predictors were still the same, hence showing a stable model behaviour.

Chapter 12: Limitations

12.1. Model assumptions:

Logistic regression assumes a linear relationship between predictors and the log-odds of the outcome, independence of observations, and reduced multicollinearity among features. These assumptions, however, may not entirely hold for real, world medical data, consequently, the model's capacity to get complex or nonlinear relationships may be limited.

Features with overlapping distributions of many variables indicate that a single variable cannot clearly distinguish CHD from non, CHD cases. This is consistent with the multifactorial nature of CHD, where risk results from the interaction of several variables rather than the influence of one.

12.2. Missing data effects:

Although missing values were imputed using median (numerical) or mode (categorical) values from the training set, imputation cannot perfectly recreate true values. Some bias in coefficient estimates may remain, and subtle patterns could be lost, particularly for variables with more missing data.

12.3. Generalizability:

The dataset is heavily imbalanced in terms of classes since the number of CHD cases is far less than the number of non-CHD cases. To some extent, weighting classes can mitigate this issue during training, but the model can still lean towards non-CHD predictions.

Logistic regression outputs interpretable risk scores based on probabilities and can be considered as a baseline model, however, it should not be used as a final clinical diagnostic tool. The results obtained from it might not apply to other populations with different age and gender distributions, disease prevalence, or healthcare settings unless externally validated.

Chapter 13: Conclusions

13.1. Key findings:

The logistic regression model was able to pinpoint the major predictors that determine a person's risks for 10-year CHD. These predictors included stroke history, hypertension (BPMeds and prevalentHyp), age, diabetes, sex and smoking. This outcome agrees with an exploratory data analysis revealing that CHD cases had higher median values of these variables. It also accords with the epidemiological evidence from population, based studies that are well known in the field. The models effectiveness, expressed by a mean cross, validated AUC of 0.724 and average precision of 0.30 both above the baseline CHD prevalence of 0.15 indicates that the model is able to capture the target-feature dependencies in the dataset well enough to identify the risk factors and is therefore capable of differentiating between individuals that are at a higher risk and those that are at a lower risk.

Stratified threshold analysis also revealed the tradeoffs between precision and recall that can be seen in the clinical challenge of how to identify the truly high-risk individuals and at the same time having as few false positives as possible. Age, diabetes, and smoking were among the variables that had positive coefficients and odds ratios greater

than 1 and thus contributed most strongly to the predicted risk, while higher education seemed to be a protective factor with the odds ratio being slightly less than 1. Overall, the model is consistent with the concept that CHD is a multifactorial disease, where several modest predictors together can be used for a meaningful risk stratification and this does not depend on a single variable.

13.2. Value of data science in CHD risk prediction:

This research illustrates how data science methods can be useful in measuring and forecasting cardiovascular risk. Through the integration of various clinical and lifestyle factors, computer models based on data not only give understandable explanations but also help identify the highest risk individuals early on and guide preventive measures. Thus, they can be used to supplement conventional epidemiological methods.

13.3. Suitability of logistic regression:

Logistic regression proved to be a suitable baseline model for CHD risk prediction. Its interpretability, probabilistic output, and ability to handle both continuous and categorical predictors make it a practical tool for understanding and communicating risk. While it cannot capture highly complex or nonlinear relationships, it provides a transparent foundation for further modeling.

13.4. Future work:

To see whether predictive performance increases above the baseline logistic regression, future research should expand on this work by trying alternative machine learning models like Random Forest or XGBoost. Furthermore, the model's generalizability to populations other than the Framingham teaching cohort would be evaluated by validating it using external datasets. Model accuracy and interpretability may be further improved by investigating various preprocessing techniques, such as different imputation techniques or adding more pertinent risk factors. Lastly, initiatives to enhance clinical usability, like calibrating risk ratings or showing feature importance, may help make predictions more useful for preventive cardiology.

Chapter 14: References

[1] R. B. D'Agostino Sr., S. Grundy, L. M. Sullivan, and P. Wilson, "Validation of the Framingham coronary heart disease prediction scores: Results of a multiple ethnic groups investigation," JAMA, vol. 286, no. 1, pp. 180–187, 2001, doi: 10.1001/jama.286.2.180.

[2] T. R. Dawber, G. F. Meadors, and F. E. Moore Jr., "Epidemiological approaches to heart disease: The Framingham Study," *American Journal of Public Health*, vol. 41, no. 3, pp. 279–286, 1951. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC1525365/

[3] H. Encord, "What is Logistic Regression?," Encord Blog, 2023. [Online]. Available: https://encord.com/blog/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20a%20statistical,data%20analysis%20for%20classification%20tasks.

[4] A. Sheesh, "Framingham Heart Study Dataset," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

[5] W. McKinney, *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*, 3rd ed., O'Reilly Media, 2017.

[6] C. Schutt and R. O'Neil, *Doing Data Science: Straight Talk from the Frontline*, Sebastopol, CA: O'Reilly Media, 2013.

[7] P. Bruce and A. Bruce, *Practical Statistics for Data Scientists*, 2nd ed., Sebastopol, CA: O'Reilly Media, 2020.

[8] H. Held, "Logistic Regression by Framingham Heart Study," *Kaggle Notebook*, 2020. [Online]. Available: https://www.kaggle.com/code/helddata/logistic-regression-by-framingham-heart-study

[9] H. Joo, "ML Series: Logistic Regression," Medium, 2020. [Online]. Available: https://medium.com/%40hyojoo1531/ml-series-logistic-regression-a50a0ff32cb0

[10] J. Gerald, "Framingham Heart Study Logistic Regression," *RPubs*, 2019. [Online]. Available: https://rpubs.com/jansgerald/framingham

[11] S. Siddiq, *Logistic Regression*, *Bookdown*, 2021. [Online]. Available: https://bookdown.org/siddiq/dnd/logistic-regression.html

[12] GeeksforGeeks, "Understanding Logistic Regression," 2023. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/

[13] Google Developers, “The Sigmoid Function,” *Machine Learning Crash Course*, Google, 2023. [Online]. Available: https://developers.google.com/machine-learning/crash-course/logistic-regression/sigmoid-function

[14] J. Brownlee, “Understanding Logistic Regression: Odds Ratio, Sigmoid, MLE,” *Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/understanding-logistic-regression-the-odds-ratio-sigmoid-mle-et-al-740cebf349a3/

[15] W. B. Kannel, T. R. Dawber, G. D. Friedman, W. E. Glennon, and P. M. McNamara, “Risk factors in coronary heart disease: An evaluation of several serum lipids as predictors of coronary heart disease,” *American Journal of Public Health*, vol. 51, no. 7, pp. 1109–1124, 1961. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/13751193/

[16] S. H. Walker and D. B. Duncan, “Estimation of the Probability of an Event as a Function of Several Independent Variables,” *American Journal of Public Health*, vol. 57, no. 8, pp. 1523–1533, 1967. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC1065119/