PREDICTIVE MODELING FOR CARDIOVASCULAR DISEASES

By,

Shafiq Abubacker

Data Analysis and Machine Learning Approach

Introduction

Framingham Cardiovascular Study: Predictive Modeling for CHD Risk

- The Framingham, Massachusetts, cardiovascular study aims to predict the 10-year risk of coronary heart disease (CHD) in residents.

- Dataset comprises 4,000+ records with 15 attributes, rich in demographic, behavioral, and medical risk factors.

- Develop a predictive model for early CHD risk identification, enabling timely intervention and personalized healthcare.

Project Workflow

Data Preprocessing

Handle missing values, encode categorical variables, and normalize/standardize numerical features for dataset quality.

Exploratory Data Analysis

- Gain insights into variable distributions and relationships.
- Identify and address outliers crucial for model performance.

Feature Engineering:

- Enhance predictive power through new feature creation or existing feature transformation.

Model Selection and Training

- Choose common classification algorithms (logistic regression, decision trees, random forests,etc).
- Split dataset into training and testing sets for evaluation.
- Train models on the training set and evaluate on testing set using metrics like accuracy, precision, recall, and F1 score.

Interpretability

- Understand feature contributions for healthcare professionals to grasp key risk factors.
- Significance in dealing with patient outcomes, ensuring alignment with existing medical knowledge

Final Predictive Model

- Thoroughly evaluate and potentially fine-tune hyperparameters.
- Provides a tool for assessing 10-year CHD risk, stratifying patients based on risk levels for prioritized interventions.

Continuous Validation and Updating

- Essential as new data becomes available.
- Collaboration with domain experts and healthcare professionals for continuous refinement and alignment with dynamic cardiovascular health research.

Data Overview

- Individual Identifier:
  - id: Unique number for each person
- Personal Information:
  - age: Age in years
  - sex: Gender (Male or Female)
- Health Behaviors:
  - is_smoking: Smoking status (YES or NO)
  - cigsPerDay: Cigarettes smoked per day (may be missing)
- Medical History:
  - prevalentStroke: Previous stroke (1 for Yes, 0 for No)
  - prevalentHyp: Hypertension (1 for Yes, 0 for No)
  - diabetes: Diabetes (1 for Yes, 0 for No)
- Health Measures:
  - totChol: Total cholesterol (may be missing)
  - sysBP: Systolic blood pressure
  - diaBP: Diastolic blood pressure
  - BMI: Body mass index (may be missing)
  - heartRate: Heart rate (one missing value)
  - glucose: Glucose level (many missing values)
- Target Variable:
  - TenYearCHD: Coronary heart disease in the next 10 years (1 for Yes, 0 for No)

- Missing Values: Several features have missing data, requiring appropriate handling.
- Target Variable: The goal is to predict 10-year CHD risk based on other features.

Data Types:

Integer: id, age, disease indicators

Float: education, health measures

Object: sex, smoking status

Data Pre-processing

```
Missing Values Count
id
age
education                        8
sex
is_smoking
cigsPerDay                       2
BPMeds                           4
prevalentStroke
prevalentHyp
diabetes
totChol                          3
sysBP
```

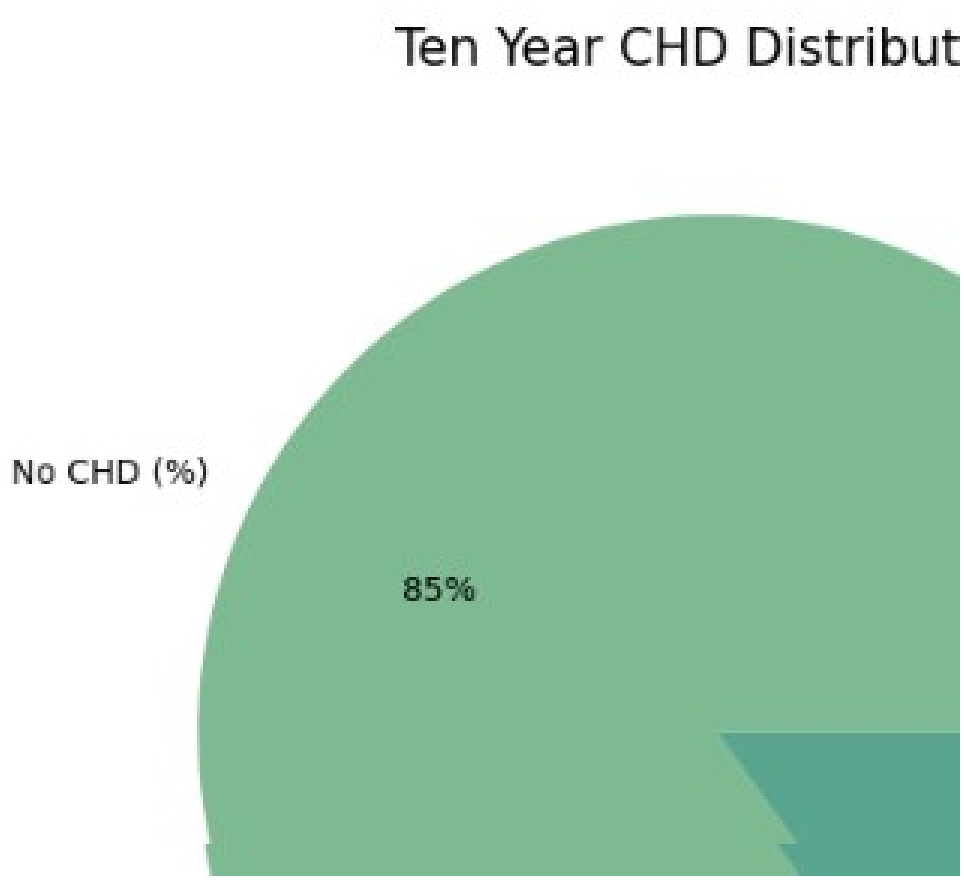
- Replaced missing values in numerical columns with the median.
- Replaced missing values in categorical columns with the mode

```
id
age
education
sex
is_smoking
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
```

Checked for duplicate values in the dataset and found there were no duplicates

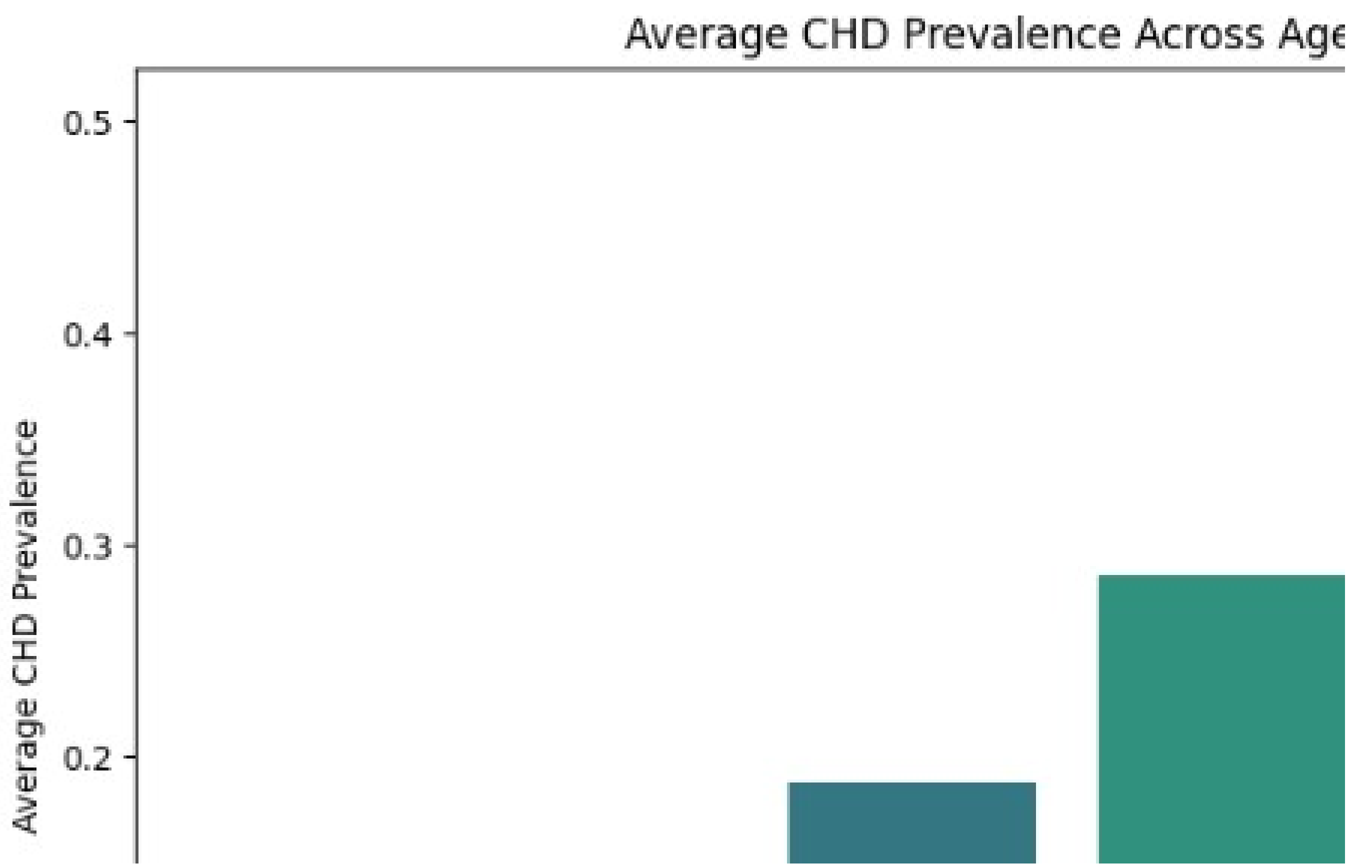
Exploratory Data Analysis (EDA)

- Distribution of Ten-Year Coronary Heart Disease (CHD) Risk in the Dataset



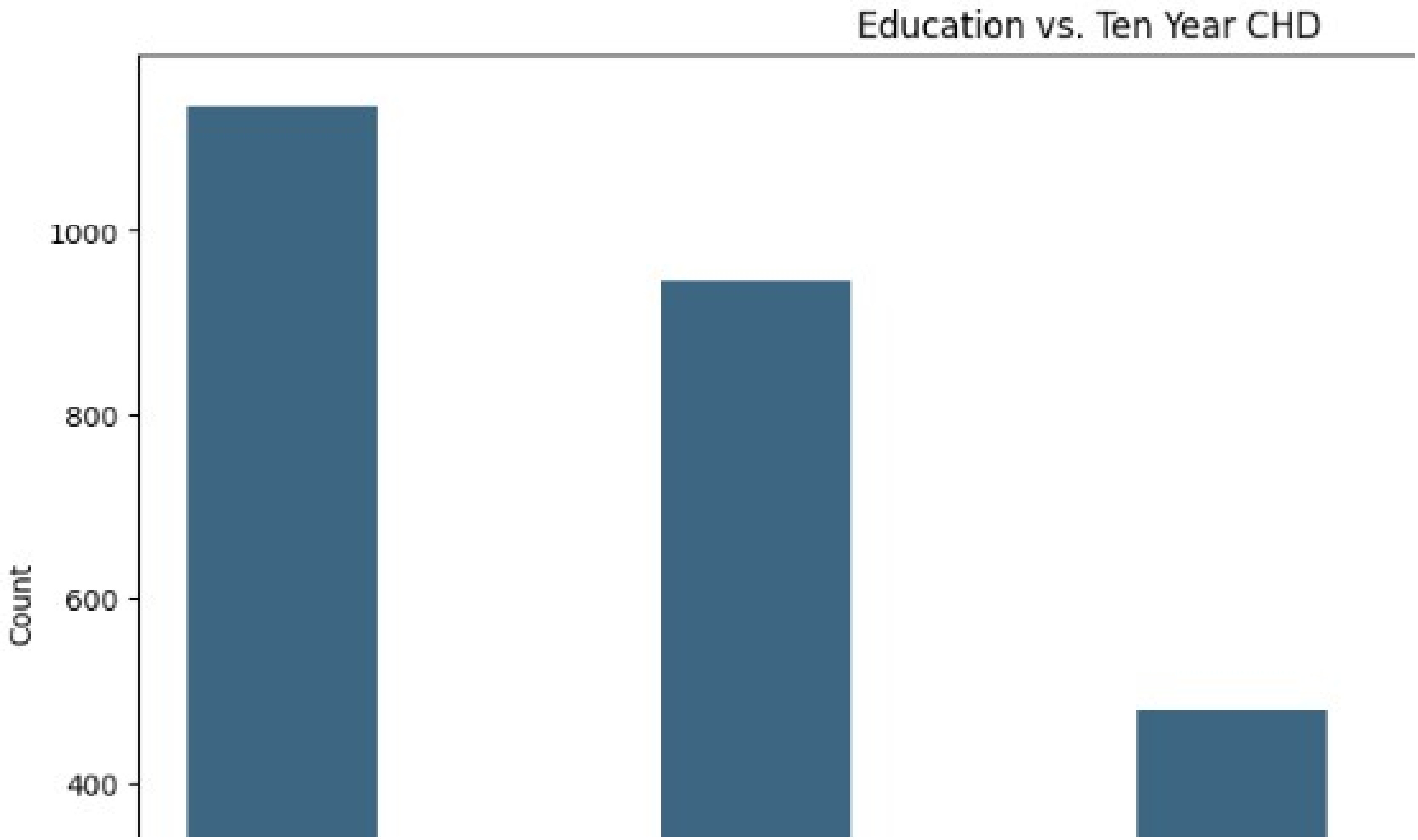
Exploratory Data Analysis (EDA)

- Relationship Between Age and Ten-Year Coronary Heart Disease (CHD) Risk



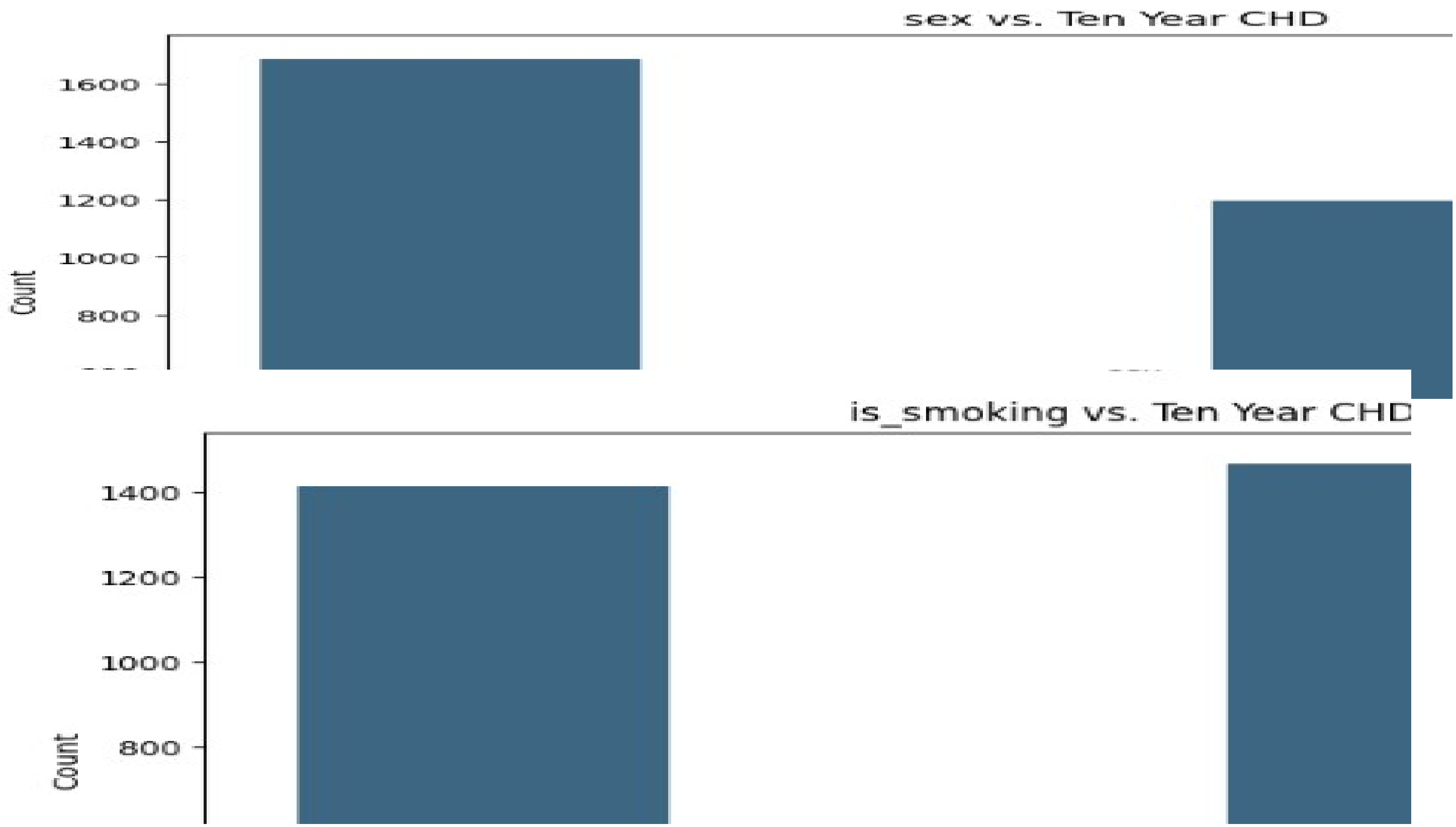
Exploratory Data Analysis (EDA)

- Exploring the Relationship Between Education and



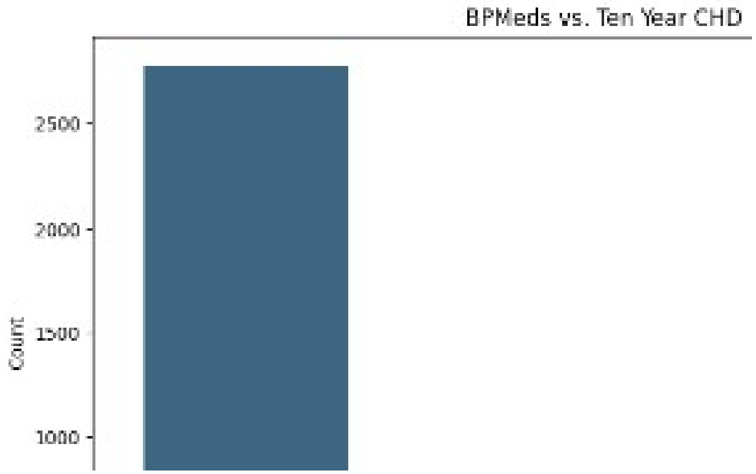
Exploratory Data Analysis (EDA)

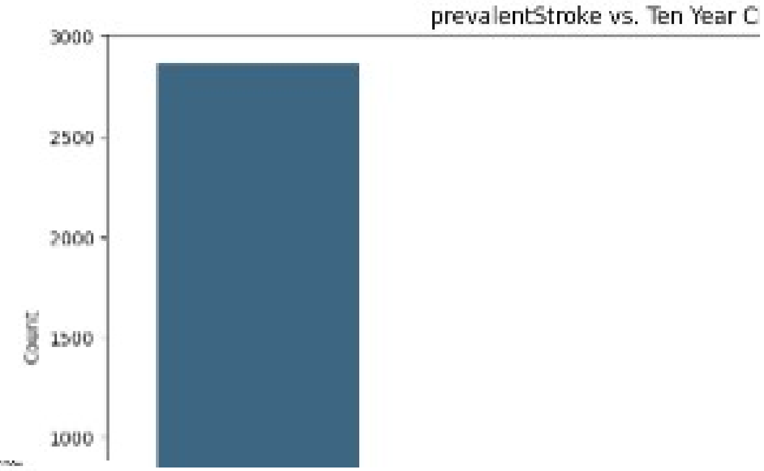
- Categorical Variables Distribution

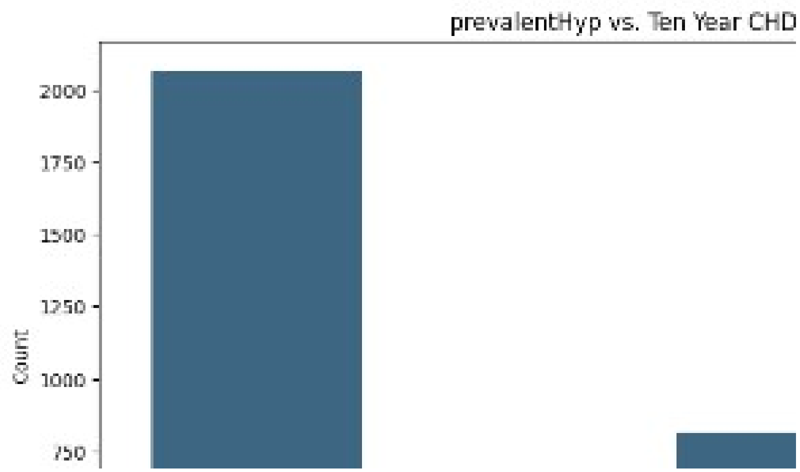


Exploratory Data Analysis (EDA)

- Categorical Variables Distribution







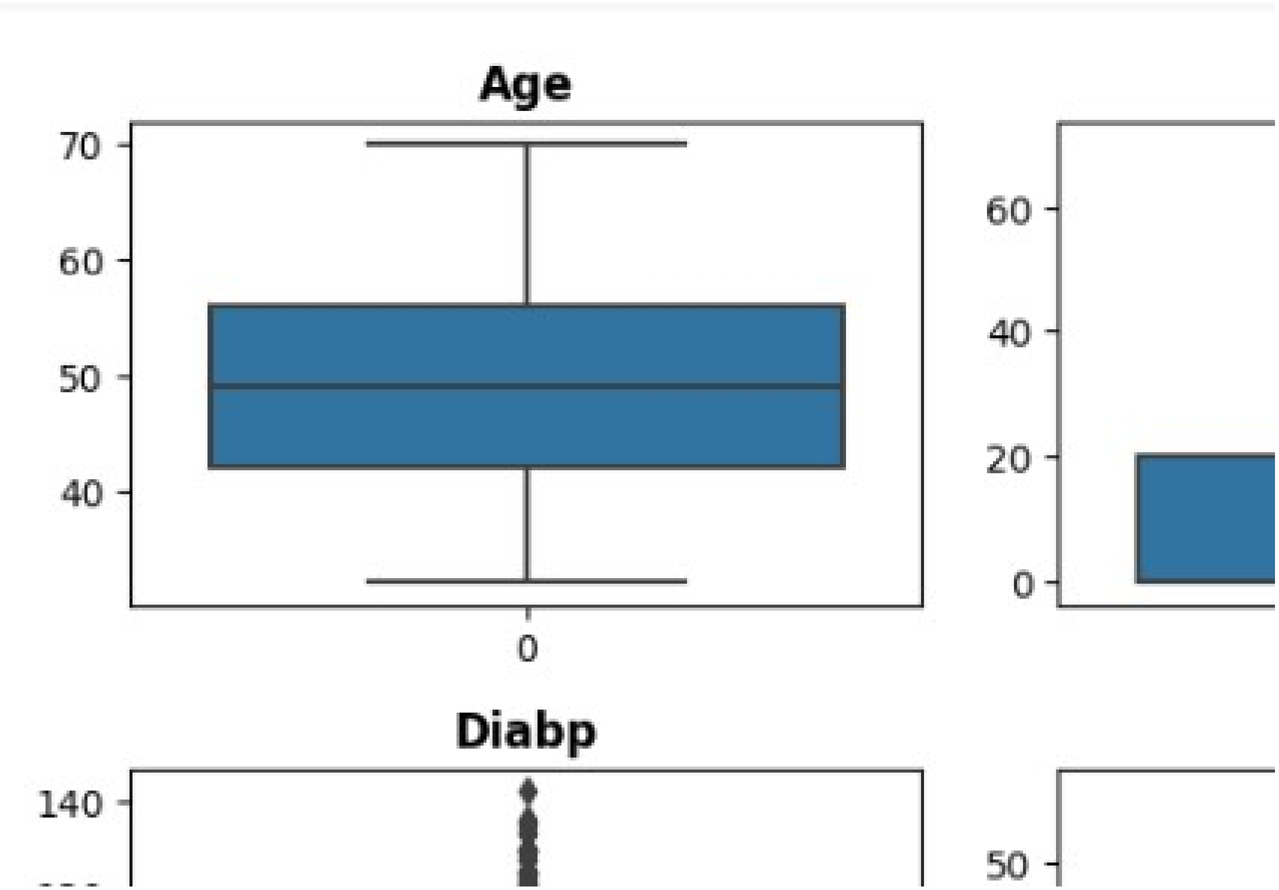
Exploratory Data Analysis (EDA)

Categorical Variables Distribution

- In the 'is_smoking' column, the distribution is relatively even. However, other health-related columns like 'BPMeds,' 'prevalentStroke,' 'prevalentHyp,' and 'diabetes' exhibit imbalances, with fewer positive cases.
- The 'TenYearCHD' column also shows an imbalance, indicating a lower count for positive cases compared to negative cases.
- Total cholesterol and BMI distributions are similar, suggesting a potential linear relationship. This could be helpful for understanding how these two health metrics influence each other.
- Glucose distribution is highly right-skewed with many outliers. These outliers represent individuals with significantly higher glucose levels than the majority.

Exploratory Data Analysis (EDA)

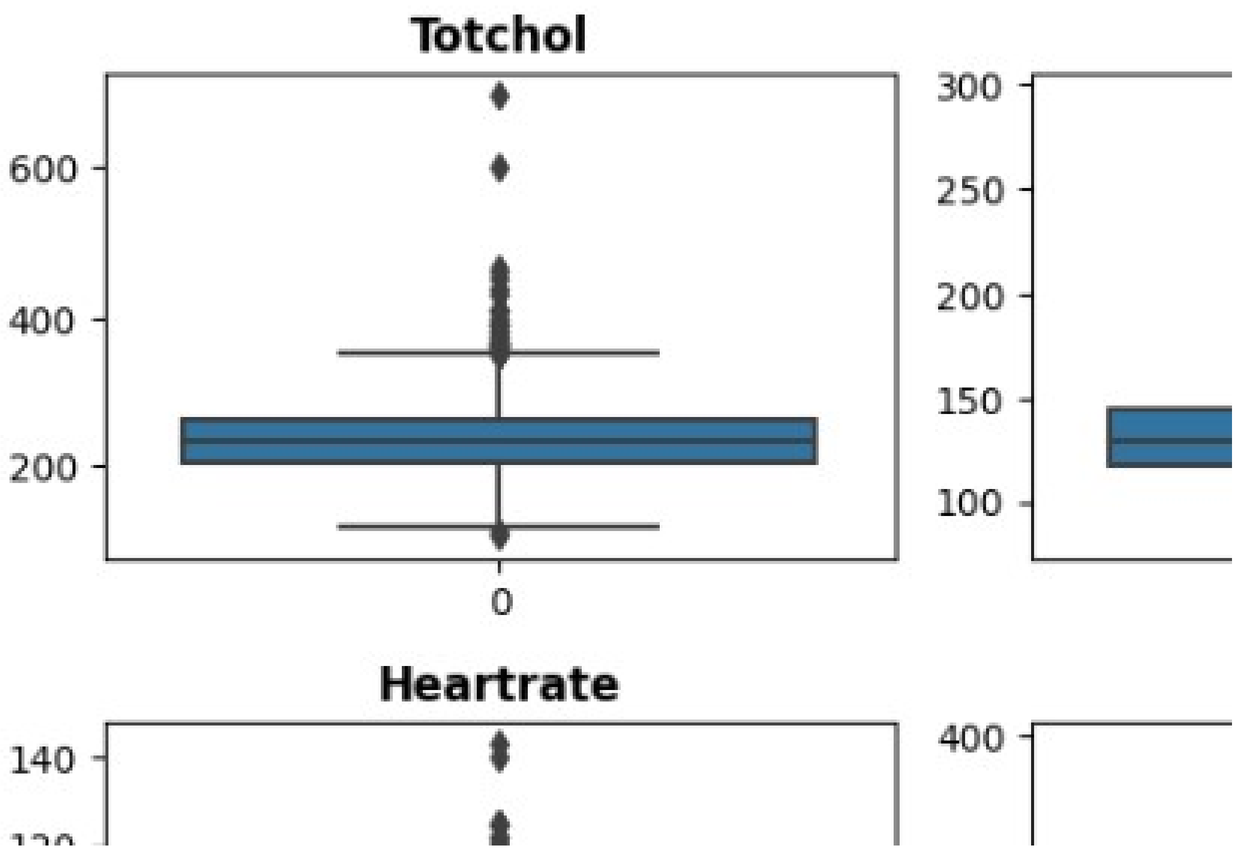
- Numeric Feature Distributions Through Box Plots



- **Cigarettes Per Day:** Right-skewed, majority smokes less, potential high-smoking outliers.
- **Diastolic Blood Pressure:** Symmetrical, no major outliers.
- **Body Mass Index:** Right-skewed, majority lower BMI, potential high-BMI outliers.

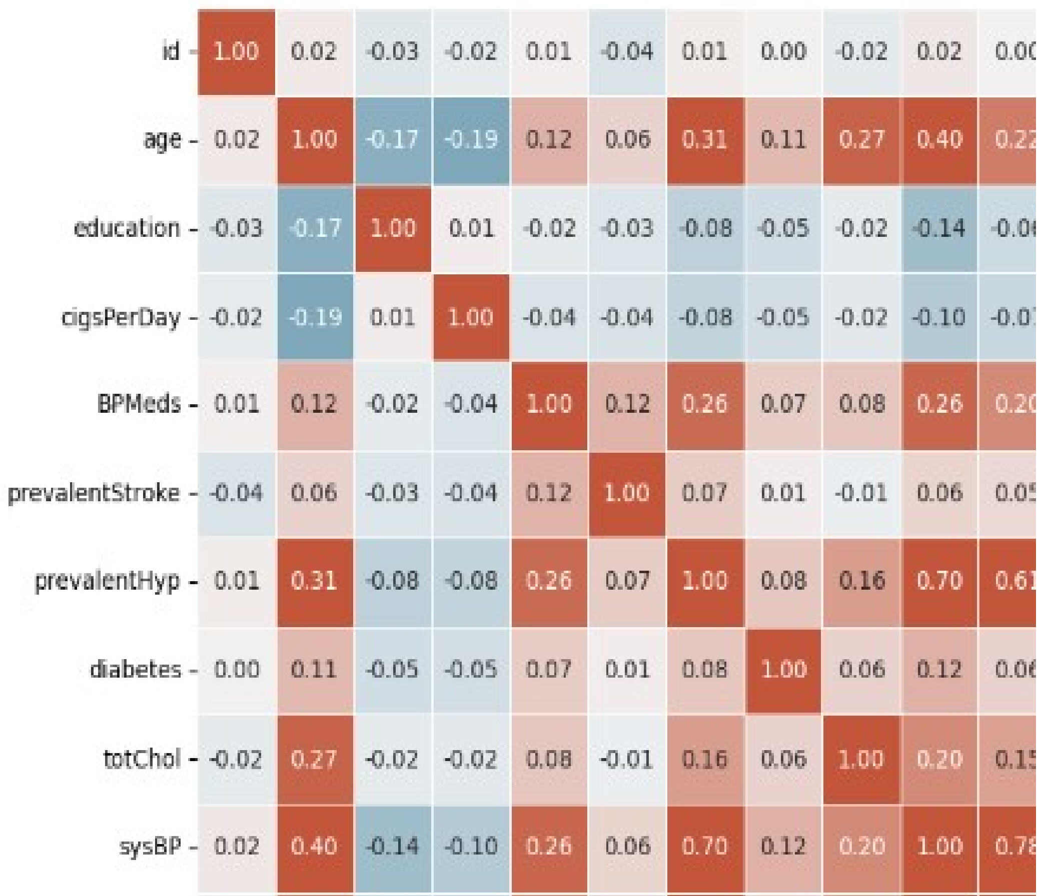
Exploratory Data Analysis (EDA)

- Numeric Feature Distributions Through Box Plots



- **Total Cholesterol:** Skewed, potential high-cholesterol outliers
- **Systolic Blood Pressure:** Slightly right-skewed, potential high-pressure outliers.
- **Heart Rate:** Skewed, potential high-heart-rate outliers.
- **Glucose:** Highly right-skewed, majority with lower levels, potential high-glucose outliers.

Exploratory Data Analysis (EDA)



- **Strong positive correlation:** Systolic and diastolic blood pressure have a strong positive correlation, meaning they tend to increase or decrease together.
- **Moderate positive correlation:** Diabetes and glucose have a moderate positive correlation (0.62), suggesting a tendency for higher glucose levels with diabetes.
- **Negligible influence:** Education level doesn't appear to be significantly correlated with CHD, implying it likely has little influence on CHD risk and can be dropped from the analysis.

Hypothesis Testing

Hypothetical Statement – 1

Null Hypothesis (H0): No CHD risk: Average age?

Alternative Hypothesis (H1):

CHD risk: Different average age?

Two-sample independent t-test

T-statistic: 13.771446174098248
P-value: 1.8455541117753288e-38

Hypothetical Statement – 2

Null Hypothesis (H0): SysBP vs DiaBP: No Correlation

Alternative Hypothesis (H1): SysBP vs DiaBP: Correlation

Pearson correlation coefficient test

Pearson Correlation Coefficient: 0.7819075063169498

Hypothetical Statement – 3

Null Hypothesis (H0): No age difference vs stroke.

Alternative Hypothesis (H1): Age difference vs stroke

Two-sample independent t-test

T-statistic: 3.4424169317511253
P-value: 0.0005835249029062626

REJECT NULL HYPOTHESIS

Feature Engineering

- **Outlier Handling Using Row Removal**
  - Extreme values in specific features were removed to ensure data representativeness.
  - Thresholds were set based on domain knowledge and medical guidelines:
    - Cigarettes/day > 50
    - Diastolic BP > 140
    - Systolic BP > 250
    - BMI > 50
    - Heart rate > 130
    - Glucose > 300
    - Total cholesterol > 500

- **Label Encoding for Categorical Variables:**
  - Categorical variable has ordinal relationship (meaningful order).
  - Assigns unique integers to categories based on their order.

Feature Manipulation

VIF Technique

	variables	
0	age	40
1	cigsPerDay	1
2	totChol	31
3	sysBP	11
4	diaBP	124

- Combining sysBP and diaBP into meanBloodPressure didn't resolve the issue.

	variables
0	age
1	cigsPerDay
2	totChol
3	BMI

High VIF Values Suggest Multicollinearity:

Variables with high VIF:
age, cigsPerDay, totChol, sysBP, diaBP, BMI, heartRate, glucose

- VIF values for meanBloodPressure remain high.
- Explore additional feature engineering or selection techniques to address multicollinearity.

Skewness Correction

```
Original Skewness:
age
cigsPerDay
totChol
BMI
heartRate
```

```python
#Applying Transformations

# Skew for sqrt transformation
new_df["cigsPerDay"] = np.sqrt(new_df['cigsPer

new_df["age"] = np.log10(new_df['age']+1)
new_df["totChol"] = np.log10(new_df['totChol']
new_df["meanBloodPressure"] = np.sqrt(new_df['
```

```
age
cigsPerDay
totChol
BMI
heartRate
```

Scaling Data

- Focused on critical health parameters: age, cigarettes per day, total cholesterol, mean blood pressure, BMI, heart rate, and glucose.

- Applied Z-score normalization using StandardScaler to standardize feature values.

- Ensures data transformation with a mean of 0 and a standard deviation of 1.

Data Splitting & Handling Imbalanced Dataset

- **Strategic Splitting:**
  - 80% training set for model learning.
  - 20% test set for unbiased performance evaluation.
  - Balances model training with generalization assessment.
- **Addressing Imbalance:**
  - SMOTE oversamples minority class to create synthetic samples.
  - Tomek links remove overlapping instances to improve class separation.
- **SMOTE's Purpose:**
  - Counteracts class imbalance by generating additional minority class data.
- **Tomek Links' Purpose:**
  - Enhances class distinction by eliminating potentially confusing instances.

Model Implementation

- **Model Selection:**
  - **Logistic Regression** - A linear model used for binary classification, estimating the probability of an instance belonging to a particular class.
  - **Decision Tree -** A tree-like model that recursively splits data based on feature conditions to make decisions or classifications.
  - **Random Forest-** An ensemble of decision trees that aggregates their predictions to improve accuracy and robustness.
  - **SVM -** A model that finds a hyperplane to separate data into classes, maximizing the margin between them in a high-dimensional space.
  - **XGBoost -** An efficient gradient boosting algorithm that sequentially builds a series of weak learners to enhance predictive performance.
  - **Naive Bayes -** A probabilistic classification algorithm based on Bayes' theorem, assuming independence between features for simplicity.

Model Comparison

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Logistic Regression	0.642	0.27	0.75	0.4
Decision Tree	0.813	0.31	0.15	0.2
KNN	0.663	0.21	0.42	0.28
Random Forest	0.786	0.22	0.13	0.16
SVM	0.623	0.27	0.78	0.4
XGBoost	0.717	0.27	0.48	0.35
Naive Bayes	0.69	0.28	0.61	0.38

Decision Tree: Highest accuracy but risks overfitting due to lower precision and recall.

Random Forest: Balanced precision and recall, sacrificing some accuracy compared to Decision Tree.

XGBoost: Good balance across all metrics (accuracy, precision, recall).

Naive Bayes: Decent accuracy with balanced metrics, simpler to implement.

KNN: Moderate accuracy, falls behind top performers in precision and recall.

Logistic Regression: Lowest accuracy, similar imbalance as SVM.

Hyperparameter Tuning

- ***GridSearchCV for Hyperparameter Tuning***
  - Systematically evaluates all possible hyperparameter combinations within a defined grid.
  - Ensures reliable performance estimates by training and testing on different data folds.

Logistic Regression - Hyperparameter Optimization

Best Hyperparameters:
C: 0.01
Penalty: 'l2'
Solver: 'liblinear'

Performance Improvement:
Accuracy: 0.642 (Before) → 0.642 (After)
Precision (Class 1): 0.27 → 0.2724
Recall (Class 1): 0.75 → 0.7570
F1-Score (Class 1): 0.4 → 0.4138

Decision Tree - Hyperparameter Tuning

Best Hyperparameters:
Criterion: 'gini'
Max Depth: None
Min Samples Split: 2

Performance Improvement:
Accuracy: 0.813 (Before) → 0.8131 (After)
Precision (Class 1): 0.31 → 0.3137
Recall (Class 1): 0.15 → 0.1495
F1-Score (Class 1): 0.2 → 0.2031

KNN - Hyperparameter Tuning

Best Hyperparameters:
N Neighbors: 5
Weight Function: 'uniform'
Algorithm: 'auto'

Performance Improvement:
Accuracy: 0.663 (Before) → 0.6632 (After)
Precision (Class 1): 0.21 → 0.2143
Recall (Class 1): 0.42 → 0.4206
F1-Score (Class 1): 0.28 → 0.2824

Model Selection

Naive Bayes

- **Strengths:**
  - Interpretability with probability-based predictions.
  - Fast training, even with large datasets.
  - Requires minimal hyperparameter tuning.
- **Weaknesses:**
  - Sensitive to assumptions about feature independence.
  - Lower accuracy compared to sophisticated algorithms.
- **Recommendation:**
  - Choose Naive Bayes for interpretability, fast training, and simplicity when accuracy is acceptable.

XGBoost

- **Strengths:**
  - High accuracy and predictive power.
  - Handles complex data and non-linear relationships.
  - Regularization features for preventing overfitting.
  - Scalability for efficient processing of large datasets.
- **Weaknesses:**
  - Challenges in interpretability.
  - Computational complexity, especially for large datasets.
- **Recommendation:**
  - Choose XGBoost for the highest accuracy, complex data, and scalability.

Conclusion

- **Data Overview:**
  - Information on health parameters like age, sex, cholesterol, blood pressure, BMI, and lifestyle.
  - Handled missing values, treated outliers, and encoded categorical variables.
- **EDA Insights:**
  - Revealed distribution patterns, correlations, and potential risk factors for cardiovascular diseases.
- **Machine Learning Models:**
  - Implemented Logistic Regression, Decision Trees, Random Forest, SVM, XGBoost, and Naive Bayes.
- **Model Evaluation:**
  - XGBoost showed a Good performance with an accuracy of 71.7%
- **Imbalanced Dataset:**
  - Addressed imbalance using SMOTE to enhance model performance.
- **Key Factors:**
  - Age, blood pressure, and cholesterol identified as critical contributors.
- **Recommendations:**
  - Focus on individuals with advanced age, high blood pressure, and abnormal cholesterol for early intervention and preventive measures.