CSCE 5214 : Software Development for Artificial Intelligence

Heart Disease Risk Prediction

Group 2

Laasya Priya Jyesta

Mohan Kalyan Guntupalli

Nagalakshmi Reddy

Vinay Kumar Parvathini

Motivation

- Heart disease often progresses silently, making **early detection critical** to prevent severe outcomes.
- Many patients remain **undiagnosed or unaware** of their risk level until a major cardiac event occurs.
- Traditional risk assessment tools may **oversimplify** health factors and fail to capture complex interactions between variables.
- Machine learning can **analyze patterns and relationships** across multiple health indicators more effectively than manual evaluation.
- A predictive model can help **healthcare professionals make timely, data-driven decisions**, enabling early intervention and improved patient outcomes.

Significance

- Helps in early identification of individuals at high risk, enabling preventive treatment before serious heart complications occur.
- Supports data-driven clinical decision-making, reducing reliance on subjective judgment or limited experience.
- Provides a more accurate and personalized risk assessment compared to traditional scoring methods.
- Can be integrated into healthcare systems or digital health applications for real-time risk monitoring.
- Contributes to reducing healthcare costs and improving patient outcomes by preventing advanced-stage heart disease.

Objective and Goal

Objective

To develop a machine learning model that predicts the likelihood of an individual developing Coronary Heart Disease (CHD) within the next 10 years.

The model will analyze clinical, behavioral, and demographic features to provide accurate and early risk assessment.

Goal

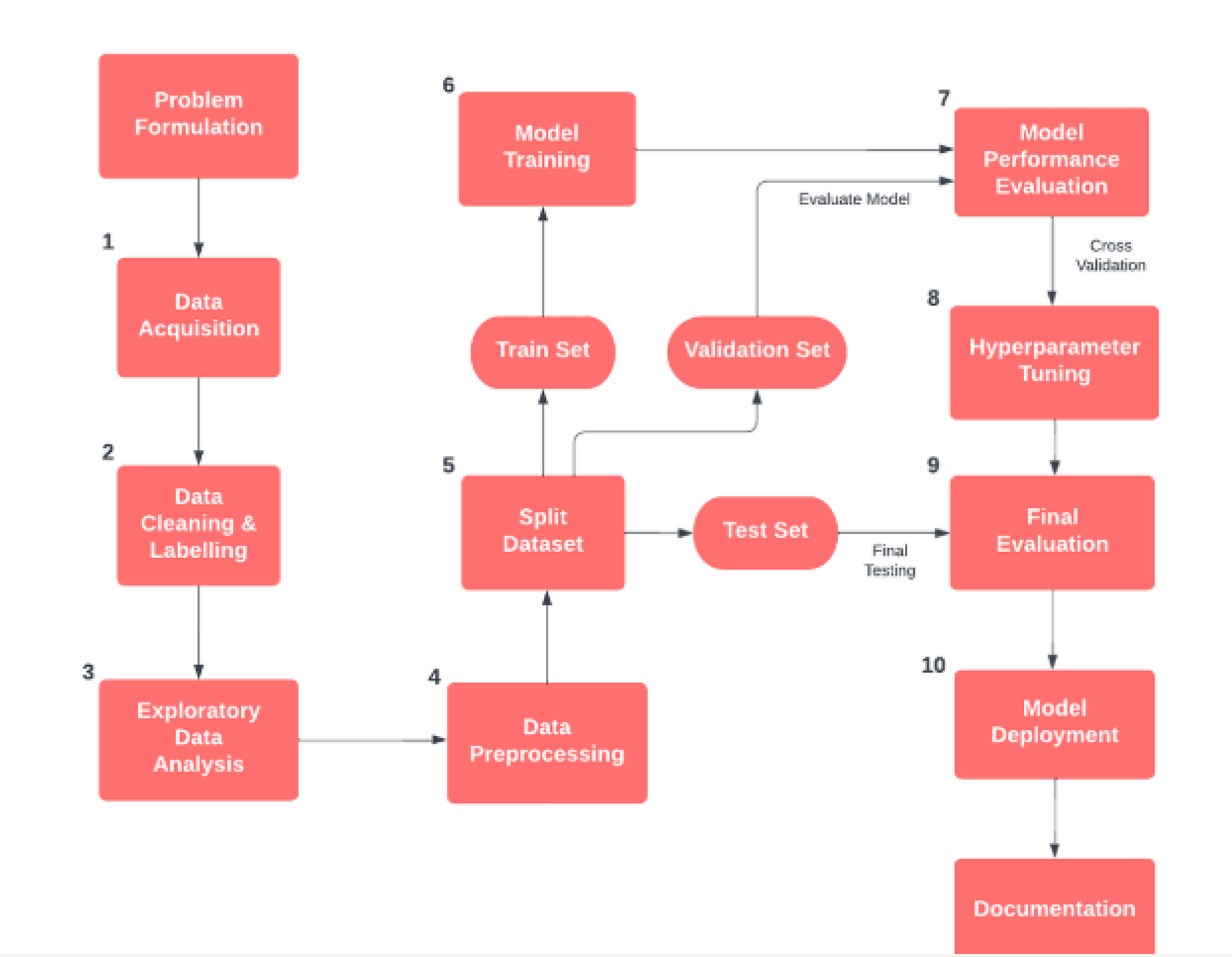
Collect and preprocess the Framingham Heart Study dataset for reliable model input.

- Identify key features that contribute most significantly to CHD risk.
- Train and evaluate multiple machine learning models to find the best-performing one.
- Compare models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- Recommend the most accurate and interpretable model for real-world healthcare use.

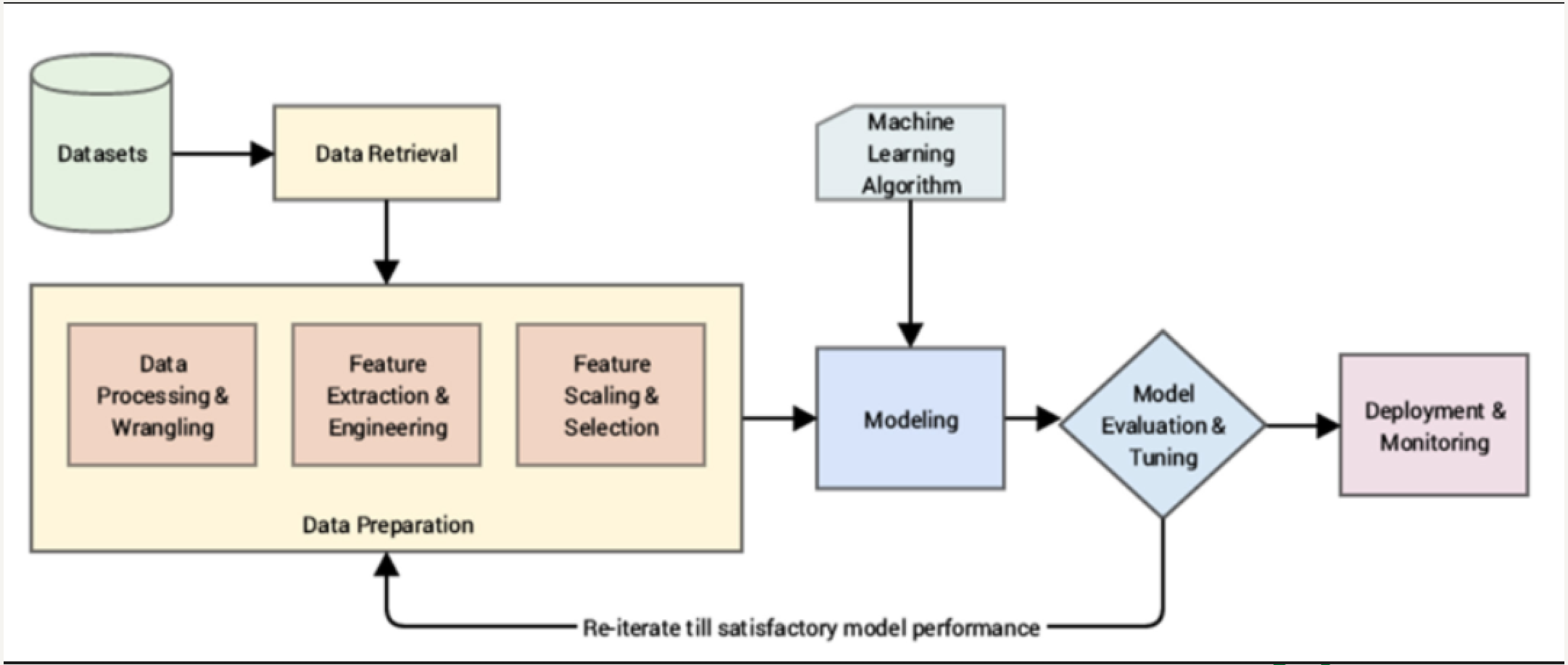
Framework and Tools

Purpose	**Tools / Frameworks**	**Description**
Data Handling & Cleaning	**Pandas, NumPy**	Used to load, clean, manipulate, and structure the Framingham dataset for analysis.
Class Imbalance Correction	**SMOTE (Imbalanced-learn)**	Balances the dataset by generating synthetic samples for the minority class to improve model fairness and performance.
Feature Selection	**Boruta Algorithm**	Identifies the most significant health and lifestyle variables contributing to CHD risk.
Model Development	**Scikit-learn, XGBoost**	Used to build and train various machine learning models for predicting heart disease risk.
Model Performance Evaluation	**Accuracy, Precision, Recall, F1-Score, ROC-AUC**	Ensures the model is accurate, reliable, and capable of distinguishing between high-risk and low-risk individuals.
Data Visualization	**Matplotlib, Seaborn**	Creates graphs and plots to interpret feature relationships, performance metrics, and model outcomes.
Development Environment	**Jupyter Notebook**	Interactive environment used to write code, run experiments, and display results smoothly.

Model Diagram



workflow Diagram


Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

Description of Tools & Components

- **Pandas:** Used to load, clean, and manage the dataset in a structured dataframe format.
- **NumPy:** Supports efficient numerical and mathematical operations on data arrays.
- **SMOTE:** Balances the dataset by generating synthetic samples for the minority class.
- **Boruta Algorithm:** Identifies and selects the most important features for prediction.
- **Scikit-learn:** Provides machine learning algorithms and evaluation metrics for model building.
- **XGBoost:** Implements high-performance gradient boosting for accurate classification.
- **Matplotlib:** Creates visual plots and graphs to represent data and model results.
- **Seaborn:** Produces enhanced, visually appealing statistical visualizations.
- **Jupyter Notebook:** Serves as an interactive environment to write code, analyze data, and display results.

Dataset

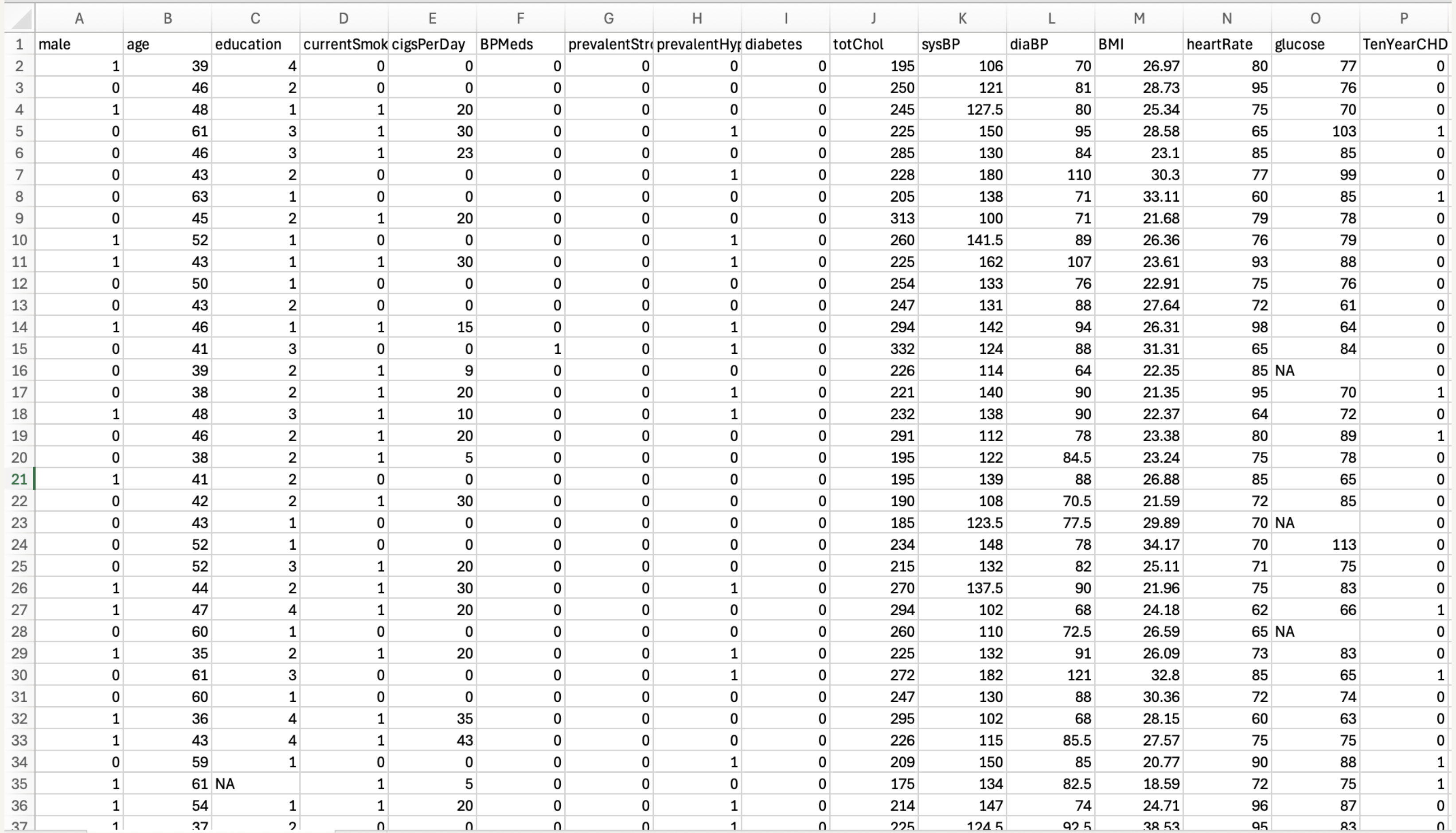
- Used **Framingham Heart Study dataset** (public cardiovascular health dataset).
- **Goal:** Predict whether an individual will develop **Coronary Heart Disease (CHD)** within 10 years.
- Contains **4,000+ patient records** with demographic, lifestyle, and clinical health measurements.
- **Target variable:** *TenYearCHD* (1 = CHD occurred, 0 = No CHD).
- Converted to a **binary classification problem** based on presence or absence of CHD risk.
- Removed records with missing values and applied **SMOTE** to balance the dataset since CHD-positive cases were significantly fewer.

Dataset

Key Features

- **Age** – Patient’s age in years.
- **Gender** – 1 = Male, 0 = Female.
- **Current Smoker** – 1 if currently smoking, else 0.
- **CigsPerDay** – Number of cigarettes smoked per day (if smoker).
- **BPMeds** – 1 if taking blood pressure medication, else 0.
- **Prevalent Hypertension** – 1 if diagnosed with high blood pressure, else 0.
- **Prevalent Stroke** – 1 if patient has had a stroke previously, else 0.
- **Diabetes** – 1 if diabetic, else 0.
- **TotChol** – Total cholesterol level (mg/dL).
- **SysBP** – Systolic blood pressure (mm Hg).
- **DiaBP** – Diastolic blood pressure (mm Hg).
- **BMI** – Body Mass Index (calculated from height and weight).
- **HeartRate** – Resting heart rate (beats per minute).
- **Glucose** – Blood glucose level (mg/dL).
- **TenYearCHD** – **Target variable:** 1 if CHD occurred within 10 years, 0 otherwise.

Snapshot of Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	male	age	education	currentSmok	cigsPerDay	BPMeds	prevalentStr	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
2	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
3	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
4	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
5	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
6	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
7	0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
8	0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
9	0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
10	1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0
11	1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0
12	0	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0
13	0	43	2	0	0	0	0	0	0	247	131	88	27.64	72	61	0
14	1	46	1	1	15	0	0	1	0	294	142	94	26.31	98	64	0
15	0	41	3	0	0	1	0	1	0	332	124	88	31.31	65	84	0
16	0	39	2	1	9	0	0	0	0	226	114	64	22.35	85	NA	0
17	0	38	2	1	20	0	0	1	0	221	140	90	21.35	95	70	1
18	1	48	3	1	10	0	0	1	0	232	138	90	22.37	64	72	0
19	0	46	2	1	20	0	0	0	0	291	112	78	23.38	80	89	1
20	0	38	2	1	5	0	0	0	0	195	122	84.5	23.24	75	78	0
21	1	41	2	0	0	0	0	0	0	195	139	88	26.88	85	65	0
22	0	42	2	1	30	0	0	0	0	190	108	70.5	21.59	72	85	0
23	0	43	1	0	0	0	0	0	0	185	123.5	77.5	29.89	70	NA	0
24	0	52	1	0	0	0	0	0	0	234	148	78	34.17	70	113	0
25	0	52	3	1	20	0	0	0	0	215	132	82	25.11	71	75	0
26	1	44	2	1	30	0	0	1	0	270	137.5	90	21.96	75	83	0
27	1	47	4	1	20	0	0	0	0	294	102	68	24.18	62	66	1
28	0	60	1	0	0	0	0	0	0	260	110	72.5	26.59	65	NA	0
29	1	35	2	1	20	0	0	1	0	225	132	91	26.09	73	83	0
30	0	61	3	0	0	0	0	1	0	272	182	121	32.8	85	65	1
31	0	60	1	0	0	0	0	0	0	247	130	88	30.36	72	74	0
32	1	36	4	1	35	0	0	0	0	295	102	68	28.15	60	63	0
33	1	43	4	1	43	0	0	0	0	226	115	85.5	27.57	75	75	0
34	0	59	1	0	0	0	0	1	0	209	150	85	20.77	90	88	1
35	1	61	NA	1	5	0	0	0	0	175	134	82.5	18.59	72	75	1
36	1	54	1	1	20	0	0	1	0	214	147	74	24.71	96	87	0
37	1	37	2	0	0	0	0	1	0	225	124.5	92.5	38.53	95	83	0

Data Preprocessing Technique

Missing Value Handling

•df.isnull().sum() was used to identify missing values in features like cholesterol and glucose.

•Records with missing values were **removed** to maintain dataset quality and avoid biased model predictions.

Duplicate Removal Check

•df.duplicated().sum() was used to check for duplicate patient records.

•No significant duplicates were found, ensuring data integrity.

Class Imbalance Handling

•The dataset had **fewer CHD-positive cases** compared to CHD-negative cases.

•**SMOTE** (Synthetic Minority Oversampling Technique) was applied to balance the classes for fair and accurate model training.

Feature Selection

•The **Boruta algorithm** was used to identify key predictors such as Age, SysBP, DiaBP, Glucose, Cholesterol, BMI, and Cigarettes per Day.

Visualization for Outlier & Distribution Analysis

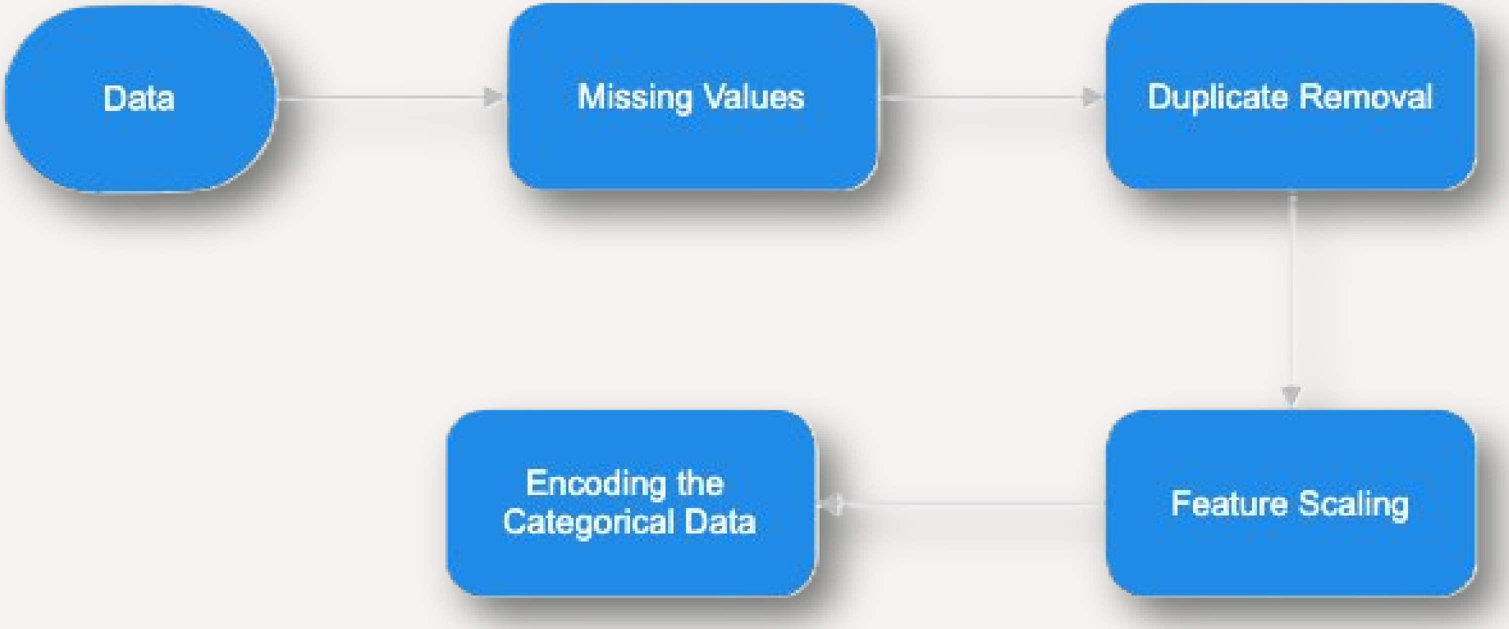
•Histograms and box plots were used to inspect the spread of values and detect possible outliers.

•Correlation heatmaps were used to visualize relationships among clinical and behavioral features.

Normalization / Scaling

•**Z-score scaling** was applied to standardize continuous variables, ensuring equal weight across features during model training.

Data Preprocessing Technique



Implementation

Data Loading

•Loaded the **Framingham Heart Study dataset** using Pandas into a DataFrame for processing and analysis.

Data Inspection

•Checked for missing values, incorrect entries, and data types using df.info() and df.isnull().sum().

•Removed rows with missing values to ensure high-quality training data.

Visual Analysis

•Used **histograms, box plots, and heatmaps** to understand feature distribution and identify patterns.

•Examined correlations among variables such as Age, Blood Pressure, Cholesterol, and Glucose.

Data Preprocessing

•Applied **Z-score normalization** to scale continuous features.

•Addressed **class imbalance** by applying **SMOTE** to ensure balanced representation of CHD and non-CHD cases.

Feature Selection

•Used the **Boruta algorithm** to determine the most important predictors, including Age, SysBP, DiaBP, Glucose, BMI, and CigsPerDay.

Implementation

Hyperparameter Tuning

- **GridSearchCV** was applied to find the best hyperparameters for models such as **SVM, Random Forest, and XGBoost**.
- Optimization focused on improving **accuracy, recall, and ROC-AUC** scores.

Model Implementation

- Trained and evaluated multiple machine learning models for CHD prediction:
  - **Logistic Regression** – baseline, interpretable model
  - **Decision Tree** – rule-based classification
  - **Random Forest** – ensemble model improving stability
  - **K-Nearest Neighbors (KNN)** – distance-based classifier
  - **Support Vector Machine (SVM)** – best performance for non-linear relationships
  - **Gradient Boosting** – sequential boosting approach
  - **XGBoost** – optimized boosting with regularization

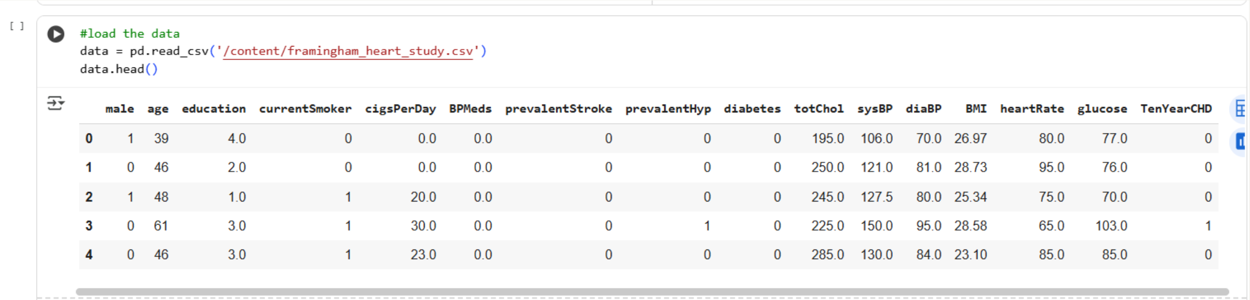
Explainability & Interpretation

- **Feature importance** analysis was performed using **Random Forest and XGBoost outputs**.
- Key predictors such as **Age, Systolic BP, Glucose, Cholesterol, BMI, and Smoking intensity** were consistently highlighted.

Bias & Reliability Consideration

- Class imbalance was addressed using **SMOTE**, ensuring fair prediction for both CHD and non-CHD groups.
- Model evaluation metrics (Accuracy, Precision, Recall, F1, ROC-AUC) ensured **reliable and clinically meaningful performance**.

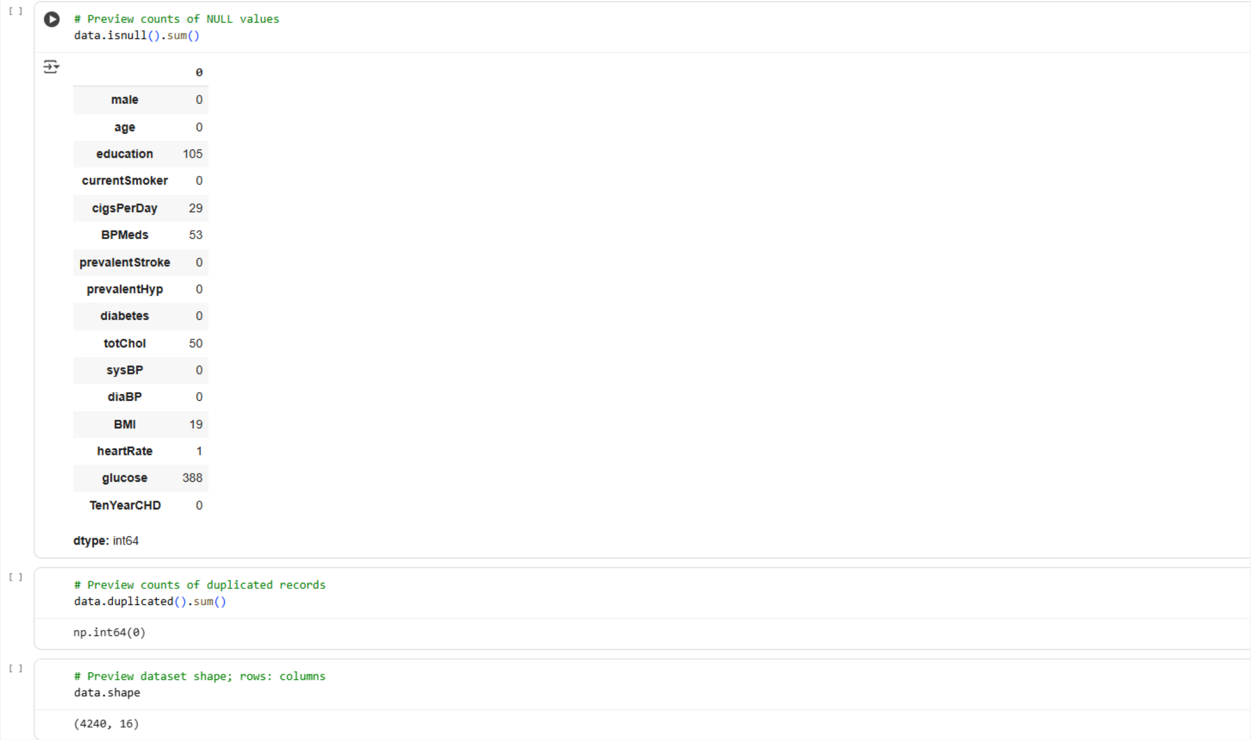
Loaded the Dataset



```
#load the data
data = pd.read_csv('/content/framingham_heart_study.csv')
data.head()
```

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Checked for missing and Duplicate values



```
# Preview counts of NULL values
data.isnull().sum()
```

	0
male	0
age	0
education	105
currentSmoker	0
cigsPerDay	29
BPMeds	53
prevalentStroke	0
prevalentHyp	0
diabetes	0
totChol	50
sysBP	0
diaBP	0
BMI	19
heartRate	1
glucose	388
TenYearCHD	0

dtype: int64

```
# Preview counts of duplicated records
data.duplicated().sum()
```

```
np.int64(0)
```

```
# Preview dataset shape; rows: columns
data.shape
```

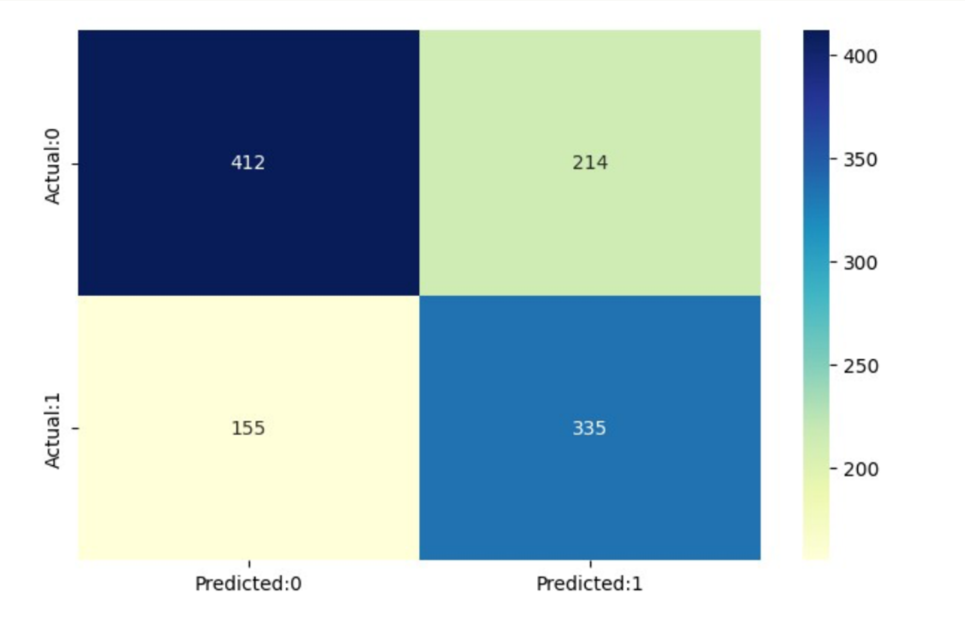
```
(4240, 16)
```

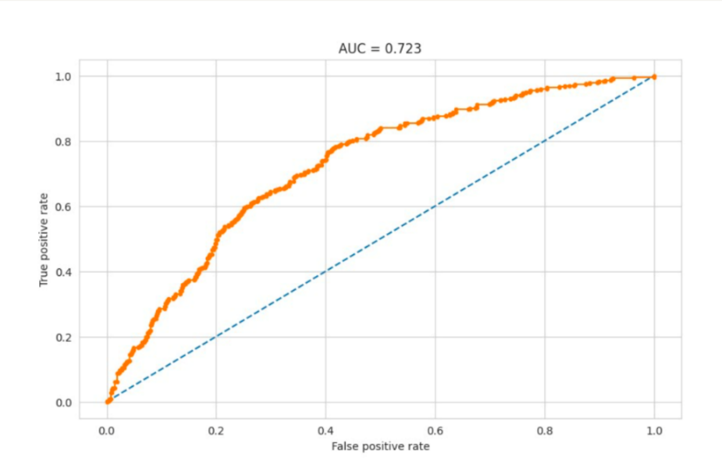
Algorithms Used

1. **Logistic Regression:**

A linear classification model that estimates the probability of a person developing CHD (1) or not (0) based on health-related features.

- Hyperparameter Tuning:
- Used Grid Search with 5-fold cross-validation to find the best model settings.
- Parameters tested:
  C values: 0.01, 0.1, 1, 10, 100
  Penalty types: l1 and l2
  Solver: liblinear
- Best Parameters found: C = 1, penalty = l2, solver = liblinear.



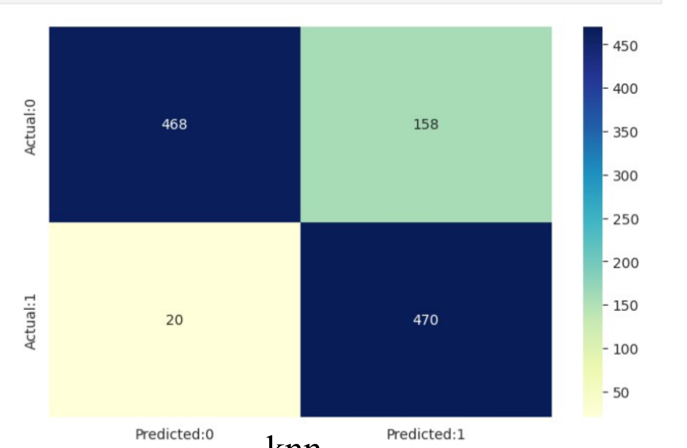


Algorithms Used

2. K-Nearest Neighbors (KNN)
Classifies individuals as CHD Positive or Negative based on the labels of the closest patients in the dataset.
Performance depends on the choice of **K** and proper **feature scaling** due to varying health measurement ranges.

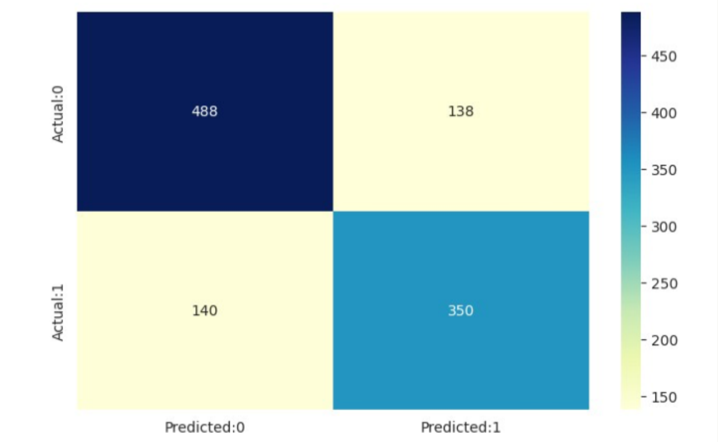
3. Support Vector Machine (SVM)
Finds the best separating boundary between CHD and non-CHD cases using patient health indicators.
Effective for high-dimensional medical data but can be computationally heavy for large datasets.

4. Decision Tree
Splits patients into CHD risk categories based on thresholds in features like age, blood pressure, and glucose.
Very easy to interpret for medical explanation, but can **overfit** the training data.

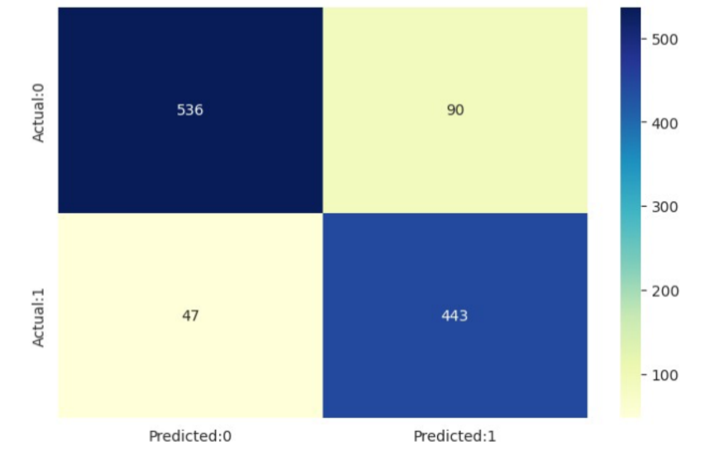


knn

Decision tree



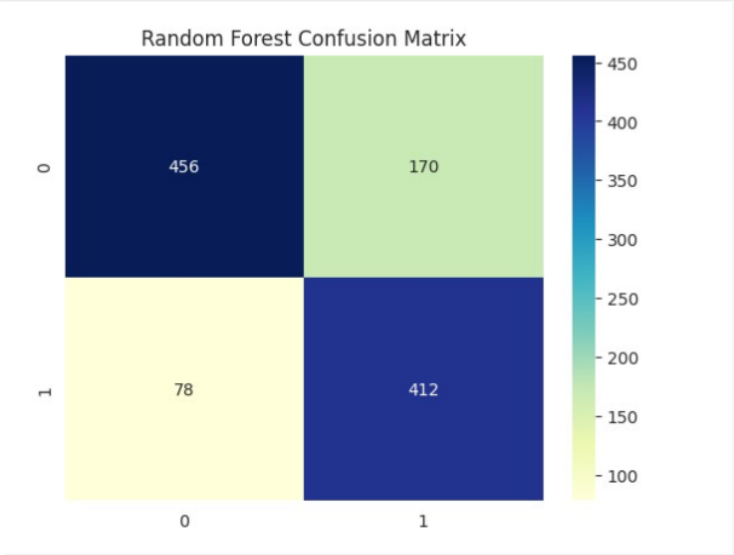
SVM

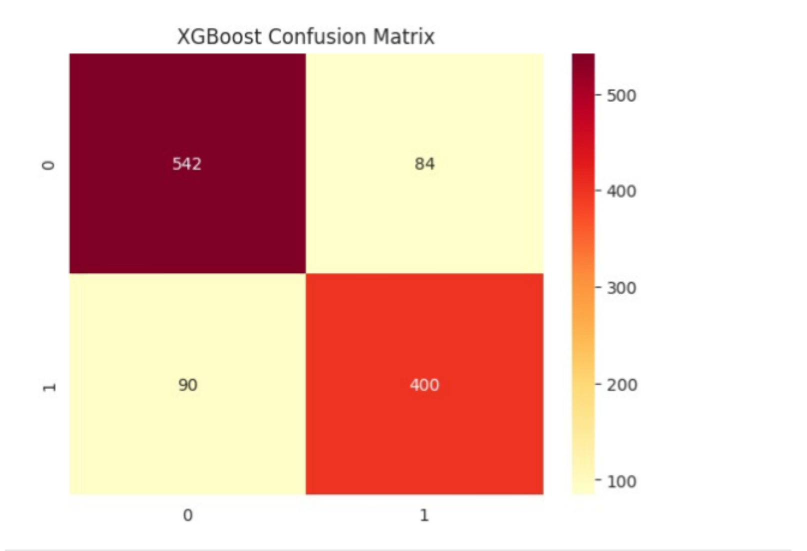


Algorithms Used

4. Random Forest
Ensemble model that builds **multiple decision trees** and combines their outputs to classify CHD risk.
More **robust and generalizable** than a single decision tree, reducing overfitting and improving prediction accuracy on medical data.

5. XGBoost Classifier
Boosting-based model that builds trees **sequentially**, where each new tree corrects errors made by the previous ones.
Highly **accurate and efficient**, especially effective in handling **imbalanced CHD data**, but requires careful tuning to avoid overfitting.





Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

•EDA was performed to understand data distribution, patterns, and relationships before modeling.

•Helped identify key health indicators influencing CHD risk.

Correlation Heatmap

•Pearson correlation matrix generated to measure relationships between features.

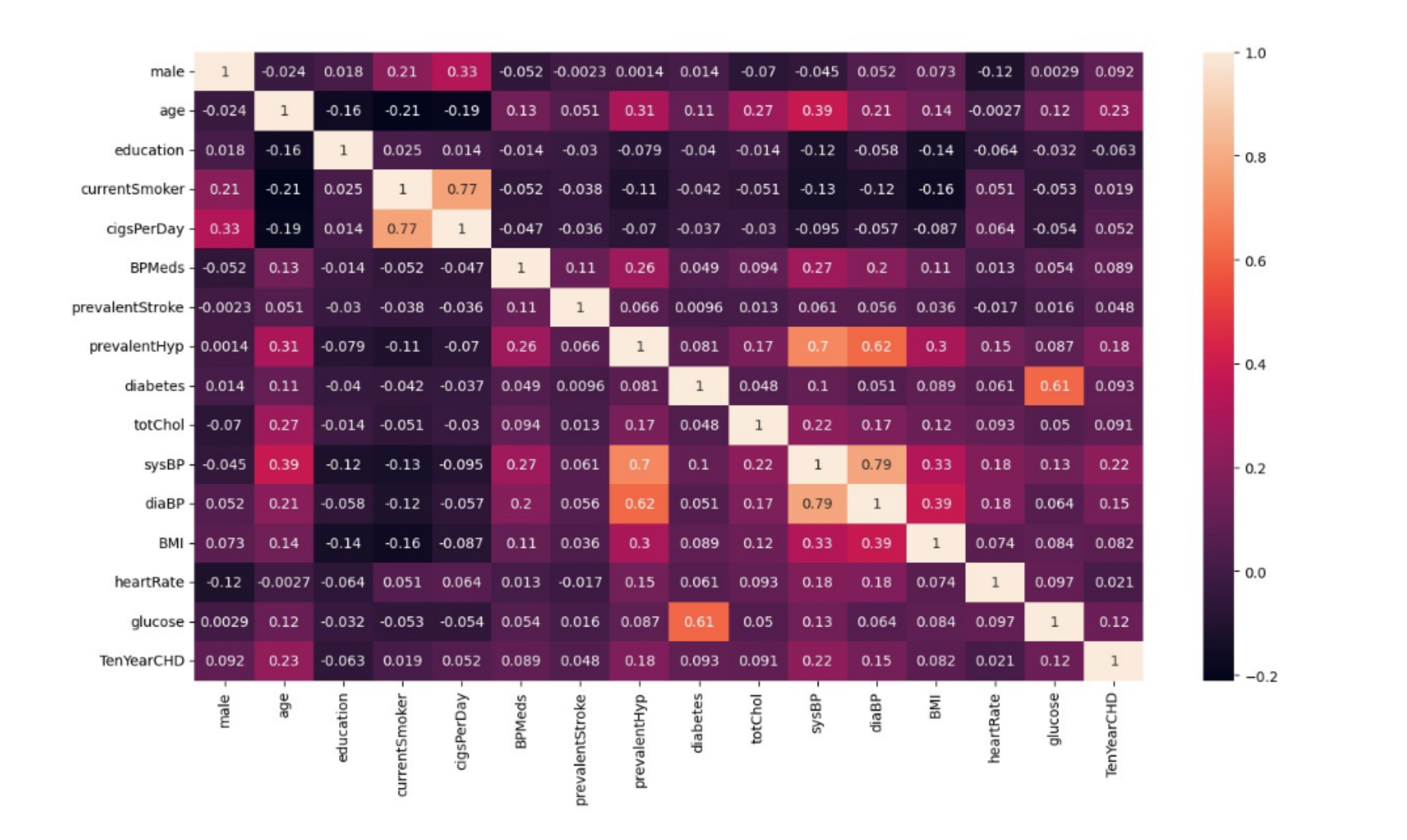
•Visualized using **Seaborn heatmap**.

Key Observations

•Age, SysBP, DiaBP, Glucose, and Cholesterol show strong positive correlation with CHD risk.

•BMI and Smoking-related features show moderate correlation.

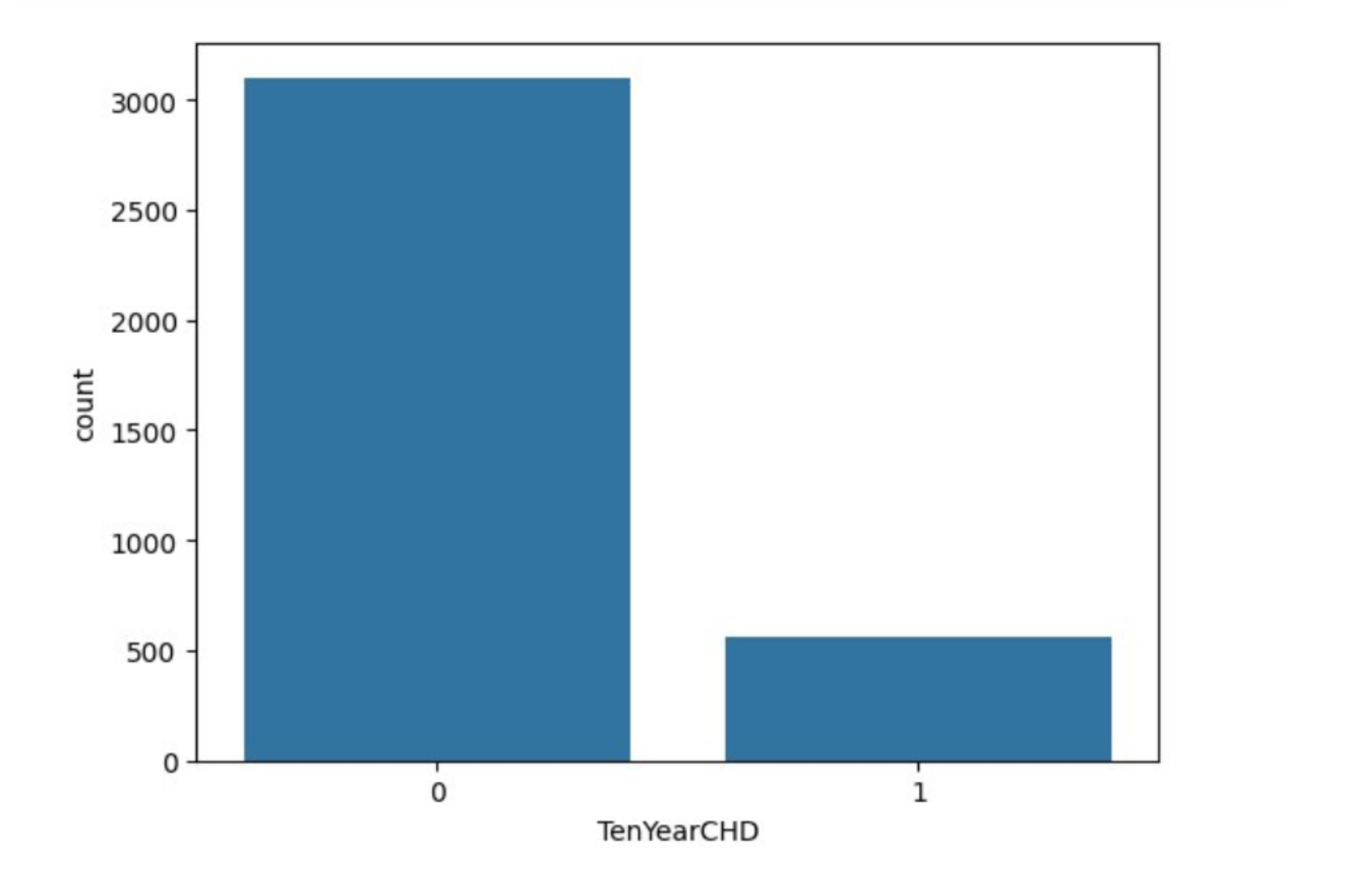
•Categorical variables like Gender were not included directly in the heatmap.



Exploratory Data Analysis (EDA)

Distribution of Ten-Year CHD Outcome

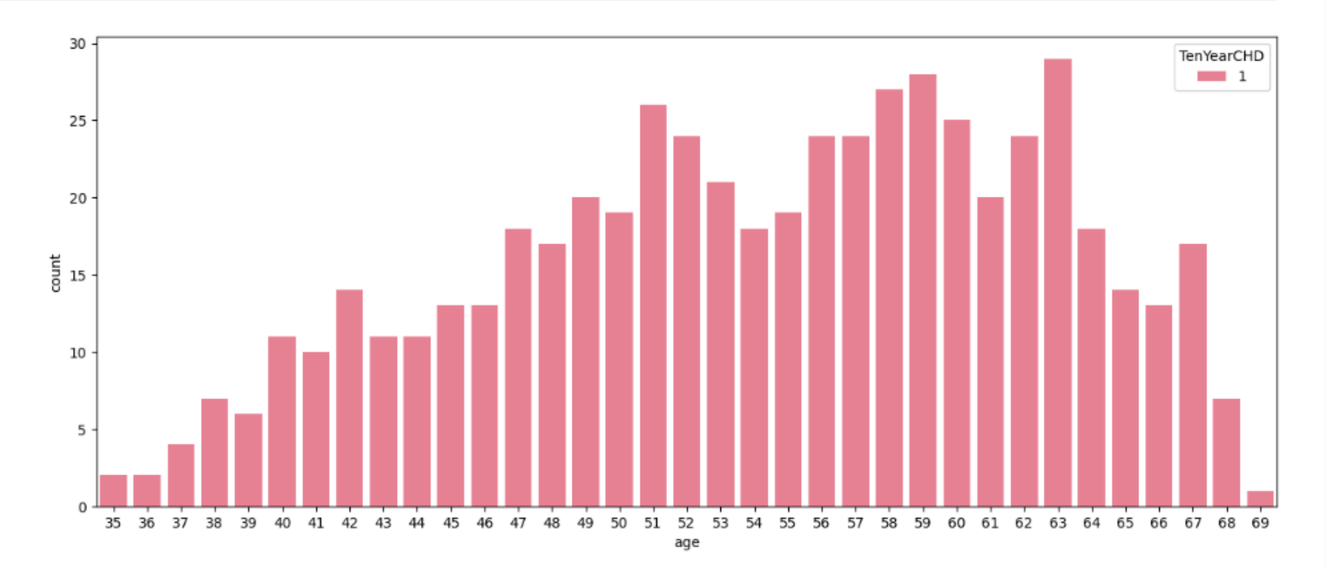
The chart clearly shows that the number of individuals **without CHD** is much higher than those with CHD.
This indicates that the dataset is **imbalanced**, with fewer positive CHD cases.
Such imbalance can lead models to **favor predicting "No CHD"** by default.
Therefore, **balancing techniques like SMOTE** are needed before model training.



Exploratory Data Analysis (EDA)

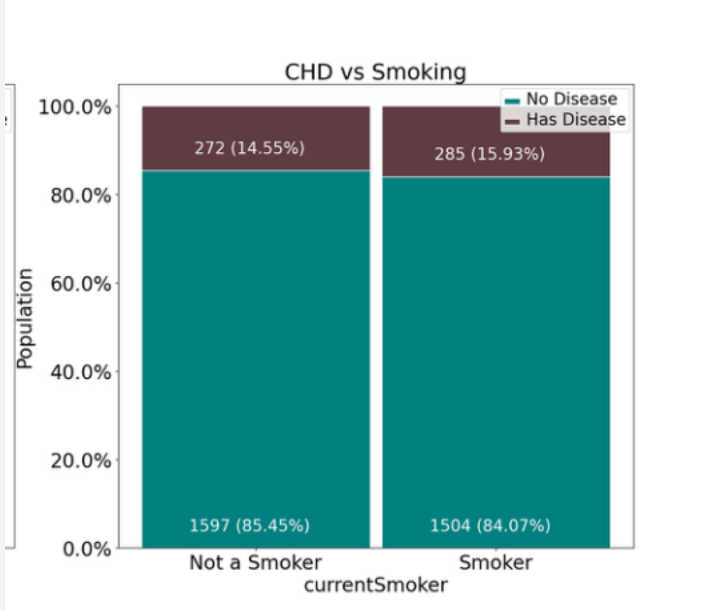
CHD Cases Across Age Groups

The chart shows CHD-positive cases increase steadily with age.
Individuals above 50 years appear significantly more likely to develop CHD.
This visually emphasizes that age is a major risk factor.
It supports the idea that preventive screening should start earlier.



Smoking Status vs CHD Occurrence

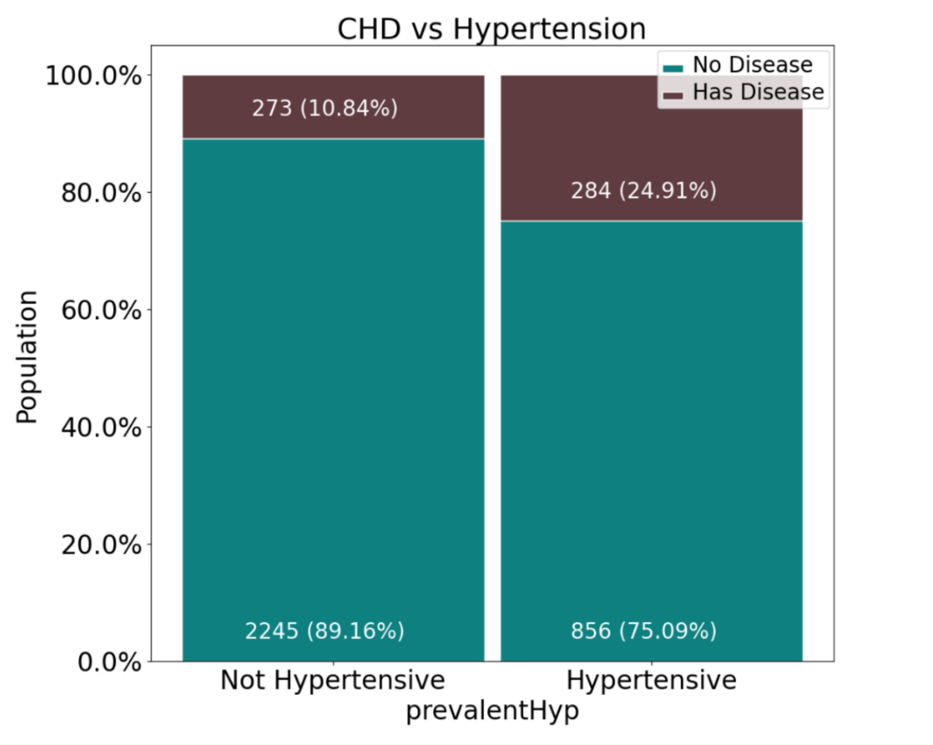
This chart compares CHD rates between smokers and non-smokers.
Smokers show a higher proportion of CHD-positive cases, highlighting smoking as a key lifestyle risk factor.
This aligns with medical findings
linking smoking intensity to cardiovascular disease.



Exploratory Data Analysis (EDA)

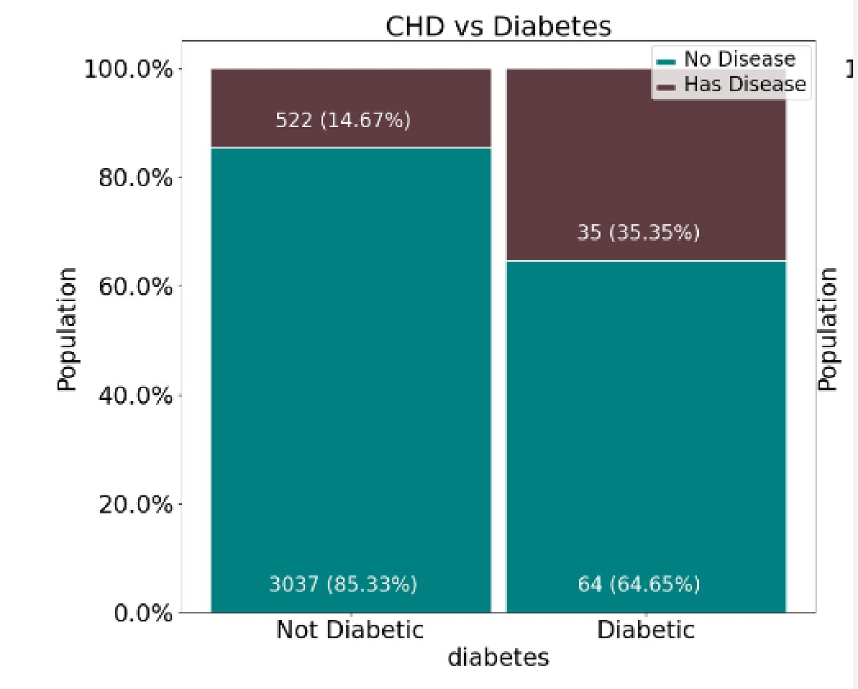
- **Blood Pressure Levels**

The distribution shows individuals with higher systolic and diastolic blood pressure have greater CHD risk.
This confirms hypertension as a critical clinical indicator.
The variation also shows BP values differ significantly across the population.
Thus, blood pressure must be strongly considered during risk assessment.



- **Glucose Level Distribution**

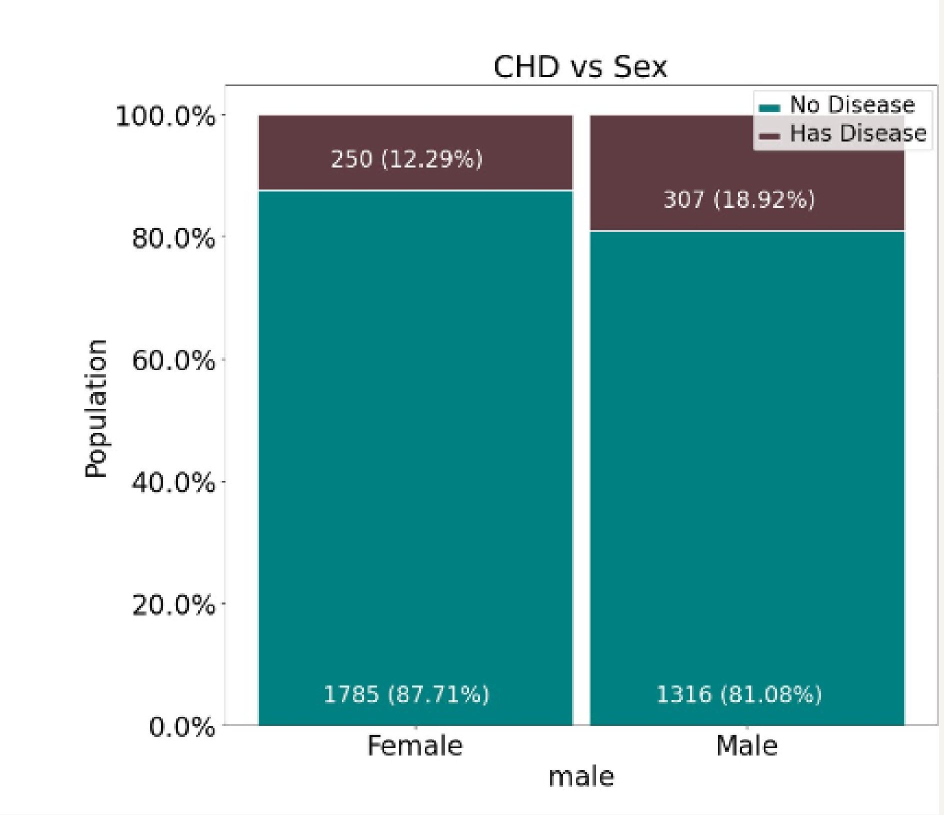
This chart compares CHD occurrence between individuals with and without diabetes.
People with diabetes show a higher rate of CHD compared to non-diabetic individuals.
This indicates that diabetes is a major contributing risk factor for heart disease.
Therefore, managing blood sugar levels is essential to reduce CHD risk.



Exploratory Data Analysis (EDA)

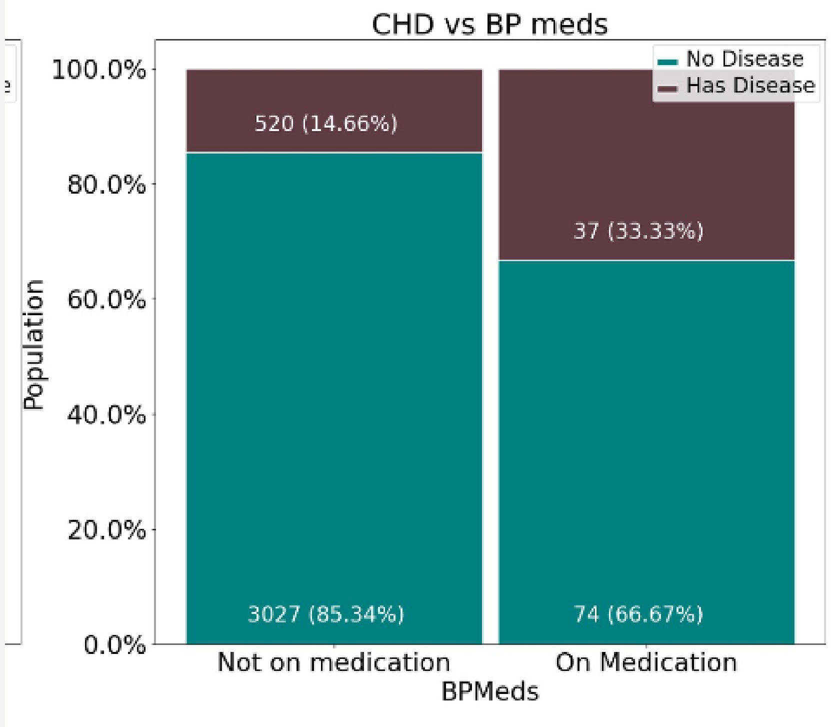
- **CHD vs Sex**

This chart compares CHD occurrence between males and females.
We observe that males have a higher rate of CHD compared to females.
This suggests that gender plays a significant role in heart disease risk.
Men may require closer cardiac monitoring and preventive care.



- **CHD vs BP Meds**

This chart shows CHD cases among individuals taking vs not taking blood pressure medication.
We see that CHD cases still occur even among those on BP medications, though usage often reflects existing hypertension.
This indicates that medication alone may not eliminate CHD risk.
Lifestyle management and continuous medical monitoring remain essential for prevention.



Model Training

Train multiple machine learning models to classify individuals as **"CHD Positive (1)"** or **"CHD Negative (0)"** based on health and lifestyle features,and evaluate their performance using standard classification metrics.

Training Dataset:
Input Features:
Age, Sex, Smoking Status, Cigarettes per Day, Systolic Blood Pressure, Diastolic Blood Pressure, Cholesterol, BMI, Heart Rate, Glucose, and Hypertension indicators.

Output Label:
TenYearCHD (1 if the individual developed CHD within 10 years, otherwise 0).

Dataset Split:
Training Set: 80%
Test Set: 20%

Scaler Used:
StandardScaler was applied to normalize numerical feature values to ensure equal model weighting.

Evaluation Metrics

Accuracy measures the overall correctness of the model by calculating the proportion of all correct predictions (both CHD Positive and CHD Negative) out of the total predictions made.

It shows how often the model predicts heart disease risk correctly.

Precision focuses on the positive predictions (CHD Positive predicted).

It tells us, out of all individuals the model predicted as having CHD, how many actually did develop CHD.

High precision means **fewer false positives** (wrongly predicting CHD when it is not present).

Recall (Sensitivity) measures how well the model identifies actual CHD cases.

It answers: out of all individuals who truly developed CHD, how many did the model correctly detect?

High recall means **fewer false negatives** (failing to detect real CHD cases).

F1-Score is the harmonic mean of precision and recall, balancing both metrics.

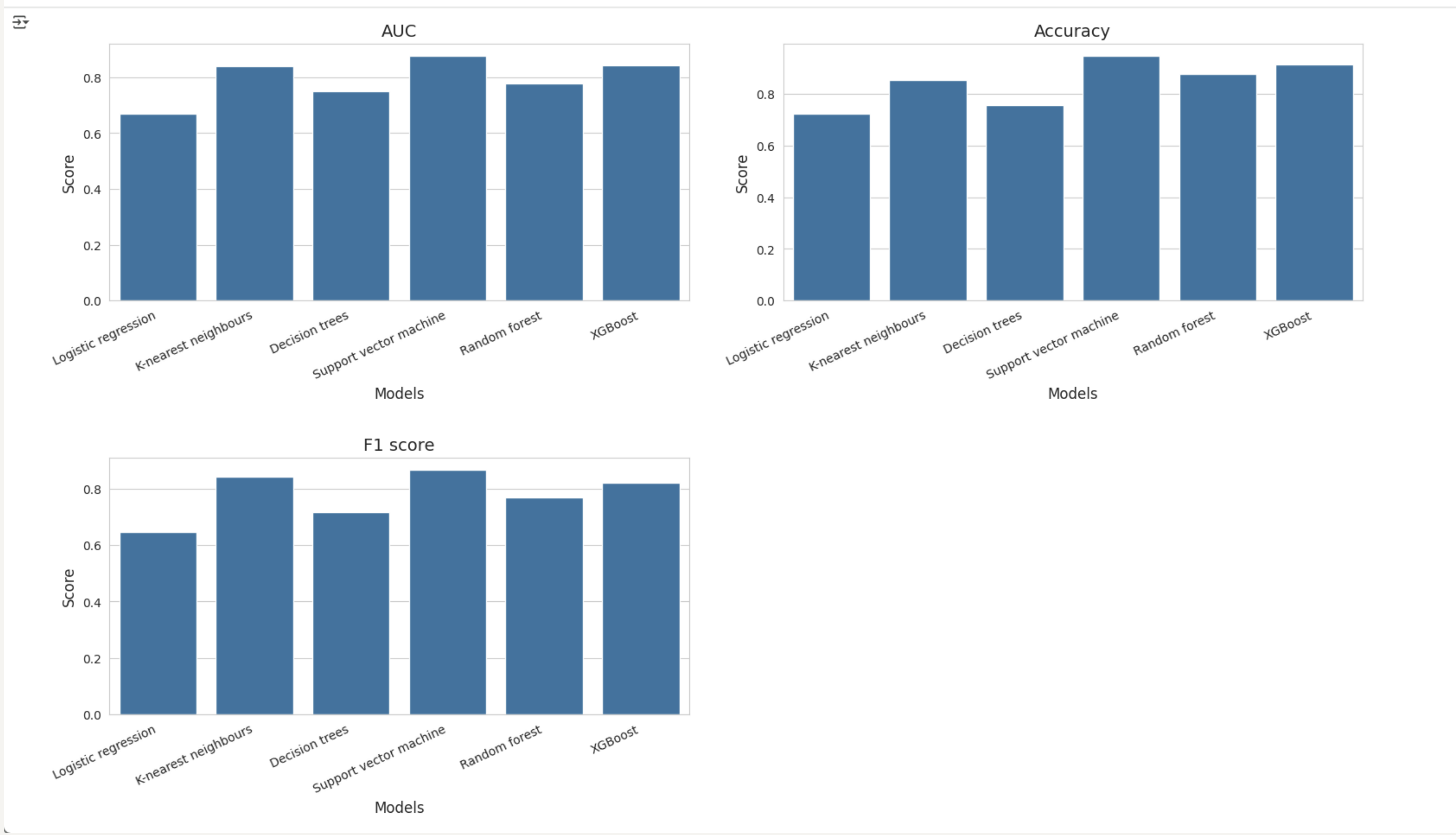
It is especially useful because CHD cases are **less frequent** than non-CHD cases (class imbalance).

A higher F1-score indicates the model is performing well in identifying CHD without too many false alarms.

Confusion Matrix is a table that breaks predictions into four outcomes:

1. **True Positives (TP):** Model correctly predicted CHD.
2. **True Negatives (TN):** Model correctly predicted No CHD.
3. **False Positives (FP):** Model predicted CHD, but the person did not have it.
4. **False Negatives (FN):** Model predicted No CHD, but the person actually had it.

This matrix helps clearly **visualize where the model is performing well and where it is making mistakes**, which is crucial in medical prediction scenarios.


AUC
Accuracy
F1 score
Score
0.0
0.2
0.4
0.6
0.8
Logistic regression
K-nearest neighbours
Decision trees
Support vector machine
Random forest
XGBoost
Models

ML Model Metrics

Model	Accuracy	Precision (CHD=1)	Recall (CHD=1)	F1-Score (CHD=1)
Logistic Regression	**0.67**	**0.61**	**0.68**	**0.64**
K-Nearest Neighbors	**0.84**	**0.83**	**0.82**	**0.82**
Support Vector Machine	**0.88**	**0.83**	**0.90**	**0.86**
Decision Tree	**0.75**	**0.72**	**0.71**	**0.71**
Random Forest	**0.78**	**0.71**	**0.84**	**0.76**
XGBoost Classifier	**0.84**	**0.82**	**0.79**	**0.81**

Summary of Evaluation Metrics

Among all the models evaluated, the Support Vector Machine showed the strongest overall performance with high accuracy, recall, and F1-score, indicating that it effectively distinguishes between classes and maintains balanced precision. K-Nearest Neighbors and XGBoost also performed consistently well, achieving high and well-balanced scores across metrics, making them reliable alternatives. Random Forest demonstrated good recall but lower precision, suggesting it identifies positive cases well but may produce more false positives. The Decision Tree model showed moderate performance, likely affected by overfitting and less generalization capability. Logistic Regression had the lowest performance across metrics, indicating that it may not capture the complexity of the data as effectively as the other models.

Conclusion

Developed a machine learning model to **predict Coronary Heart Disease (CHD)** using clinical and lifestyle-related features from the Framingham Heart Study dataset.

Used a **binary classification approach**: CHD = 1 (Positive) and CHD = 0 (Negative).

Trained and evaluated six models: **Logistic Regression, KNN, SVM, Decision Tree, Random Forest, and XGBoost**.

SVM delivered the best overall performance:

Accuracy: ~87%

High Recall, meaning it detected CHD cases more effectively

Chosen for final interpretation due to **high sensitivity, balanced precision**, and strong ROC curve.

XGBoost and KNN also performed well, showing **strong F1-scores** and stable predictions.

Random Forest improved recall but produced slightly higher false positives.

Decision Tree and **Logistic Regression** showed moderate performance, indicating limited generalization on this medical dataset.

Key influential features (based on correlation and model insights):

•Systolic Blood Pressure

•Cholesterol and Glucose Levels

•Age

•Smoking Status

•Diabetes Indicator

These clinical variables significantly contribute to CHD risk patterns.

CHD vs Diabetes and CHD vs Smoking charts confirmed **strong associations** between lifestyle/medical factors and heart disease.

ROC curve analysis demonstrated that **SVM and XGBoost** provided the highest discriminatory power.

This project highlights the importance of **data-driven decision support in healthcare**, enabling early detection of CHD and supporting preventive care strategies.

THANK YOU