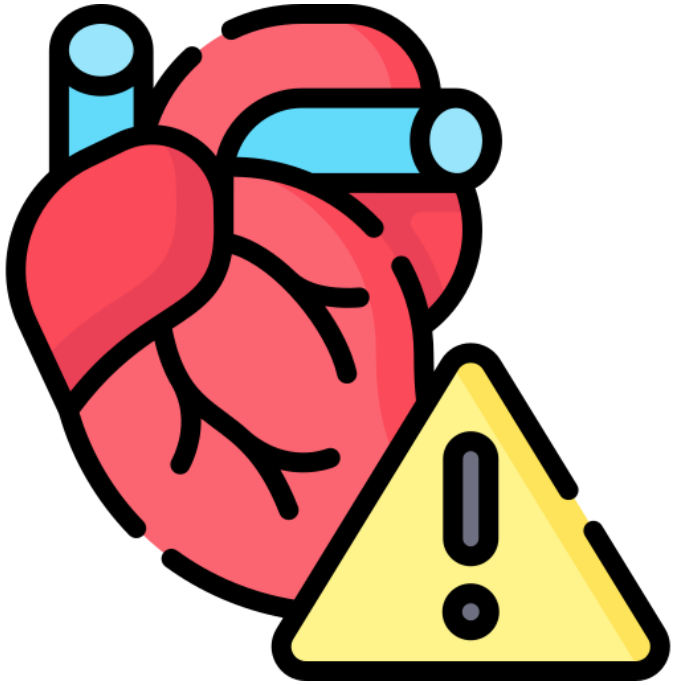
PREDICTING HEART DISEASE RISK

BY GETRUDE SHABIHA

Introduction



Project Objective

To develop a machine learning model that predicts the risk of heart disease.

Importance of Predicting Heart Disease Risk

- Early intervention for better healthcare outcomes.
- Preventive measures to improve patient well-being.

Dataset Overview

```
In [44]: data.shape
Out[44]: (4240, 16)
```

Brief Description of Dataset:

- It contains various health-related features such as age, sex, cholesterol levels, and blood pressure
- The dataset has a total of 4240 rows and 16 columns of data

Significance of Dataset:

- Foundation for predicting heart disease risk.
- The dataset has multiple data points of people who experience heart problems.

Dataset Overview

```
Missing Values:
sex                 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
```

The following steps were taken to clean the dataset:

- The dataset has multiple null values in different axes.
- Missing data is removed from the dataset by replacing categorical variables with the mode and continuous variables with the mean.
- The cleaned dataset is maintained with 4240 rows and 16 columns.

Dataset Variable Description

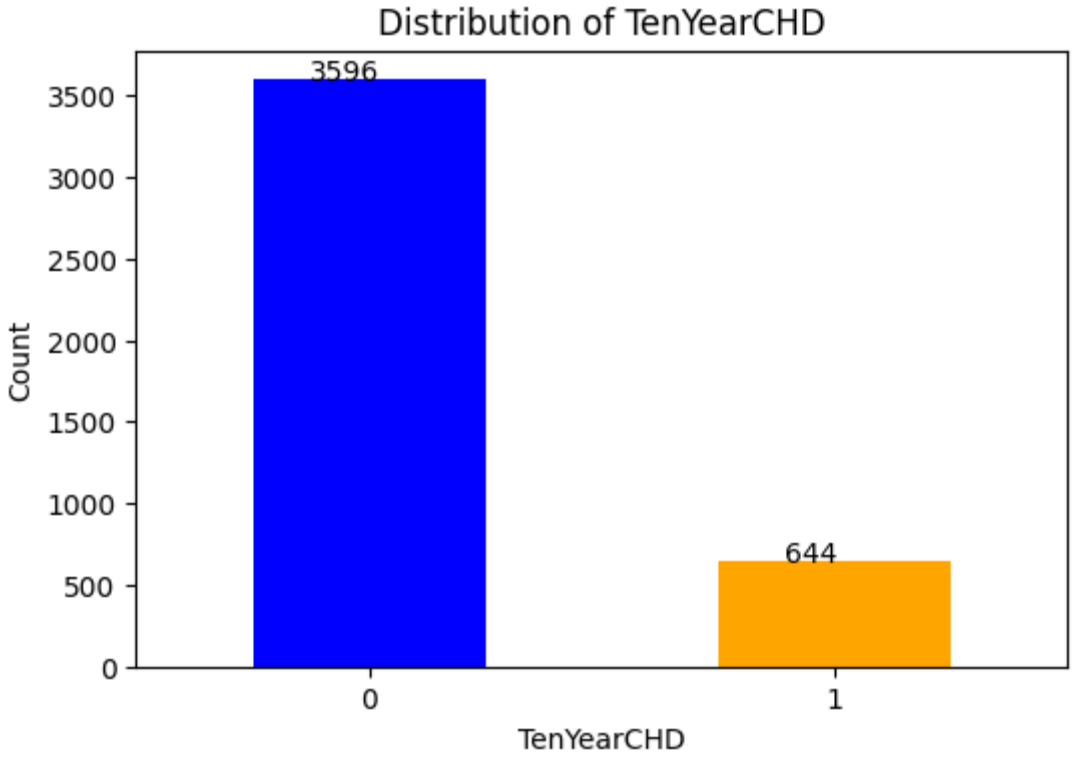
Variable Name	Description
Sex	Indicates the gender of the individual (e.g 1 for male, 0 for female)
Age	Represents the age of the participant at the time of data collection.
Education	Represents the participant's educational level, which can be categorized into different levels.
CurrentSmoker	A binary variable indicating whether the individual is currently a smoker (eg 1 For yes, 0 for no)
cigsPerDay	Describe the number of cigarettes a person smokes per day, providing insights into their smoking habits.
BpMeds	A binary value indicating whether the participant is currently using any blood pressure related medications
PrevalentStroke	A binary value indicating whether the participant has experienced any stroke related complications
PrevalentHyp	A binary value indicating whether the participant has experienced any hypertension related complications
diabetes	Indicates whether the participant has diabetes (eg 1 for yes, 0 for no)

Dataset Variable Description

Variable Name	Description
Totchol	Indicates the total cholesterol level in the blood, a key metric for assessing cardiovascular health.
sysBP	Represents the higher number in blood pressure readings, indicating the pressure in arteries when the heart beats.
diaBp	Represents the lower number in blood pressure readings, indicating the pressure in the arteries when the heart is at rest between beats.
BMI	BMI is a measure of body fat based on an individual's height and weight. It is often used to assess whether a person is underweight, normal weight, overweight, or obese.
heartRate	Describes the number of hearts per minute, providing information about the individual's heart function.
glucose	Represent the glucose levels of the participant per mm of blood
TenYearCHD	A binary value representing the participants' risk of coronary heart disease in a 10 year span

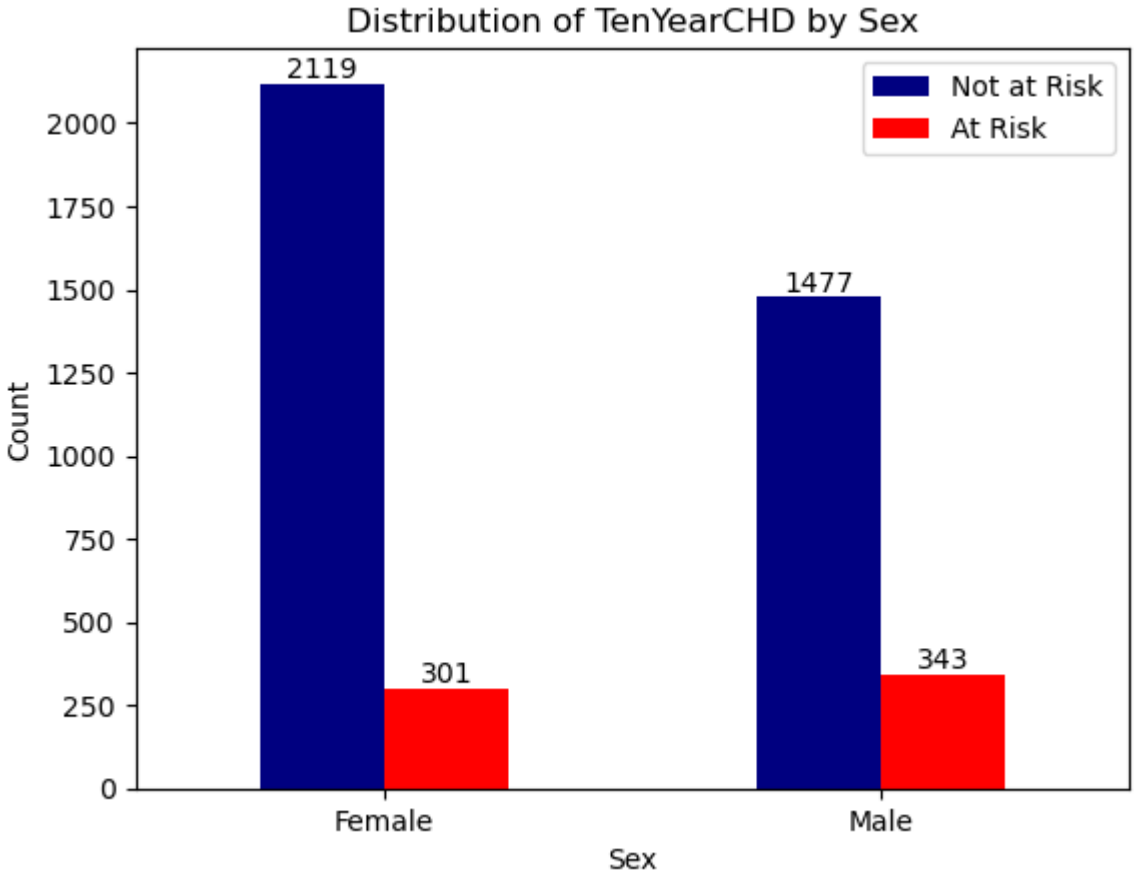
Exploratory Data Analysis (EDA)

In this section, various variables are explored by visualizing them to see their distribution and how they compare in the dataset.



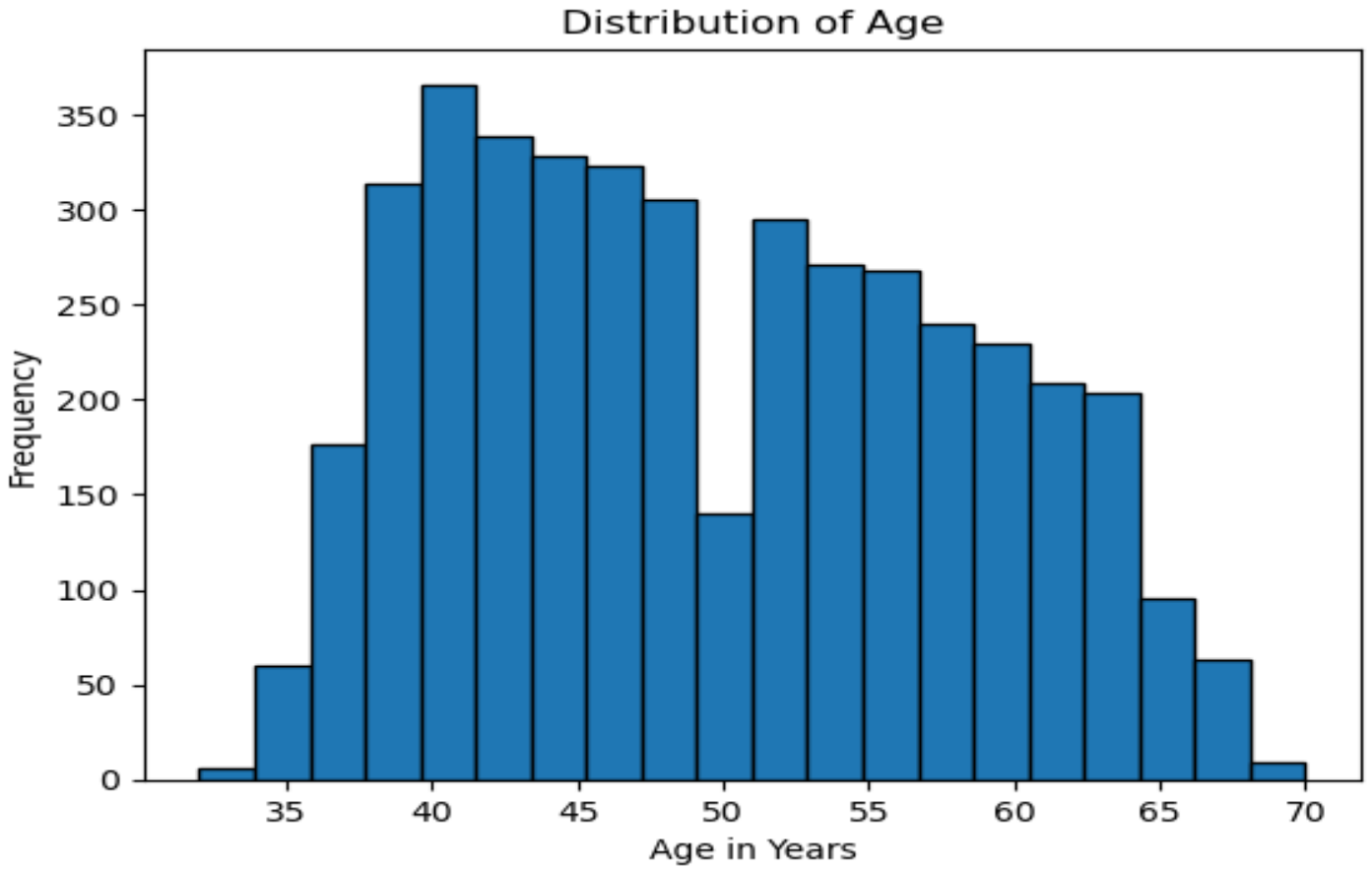
- In this exploration we are finding out the number of people diagnosed with coronary heart disease
- According to the dataset :
  - A total of 644 people were diagnosed with Coronary heart disease
  - A total of 3596 people were not diagnosed with Coronary heart disease

Exploratory Data Analysis (EDA)



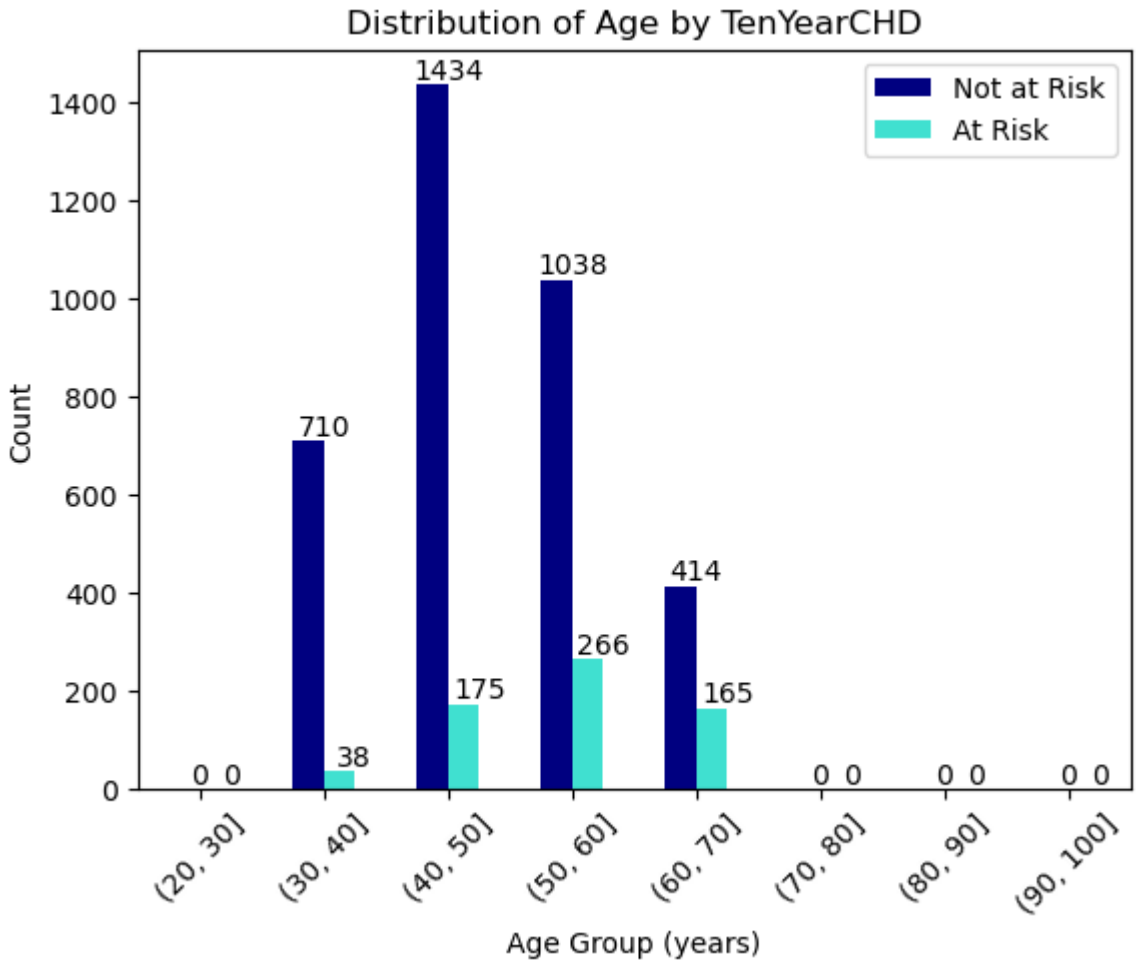
- In this exploration we look at the gender distribution versus the risk of coronary heart disease
- The male gender are at a higher risk of getting coronary heart disease as compared to their female counterparts

Exploratory Data Analysis (EDA)



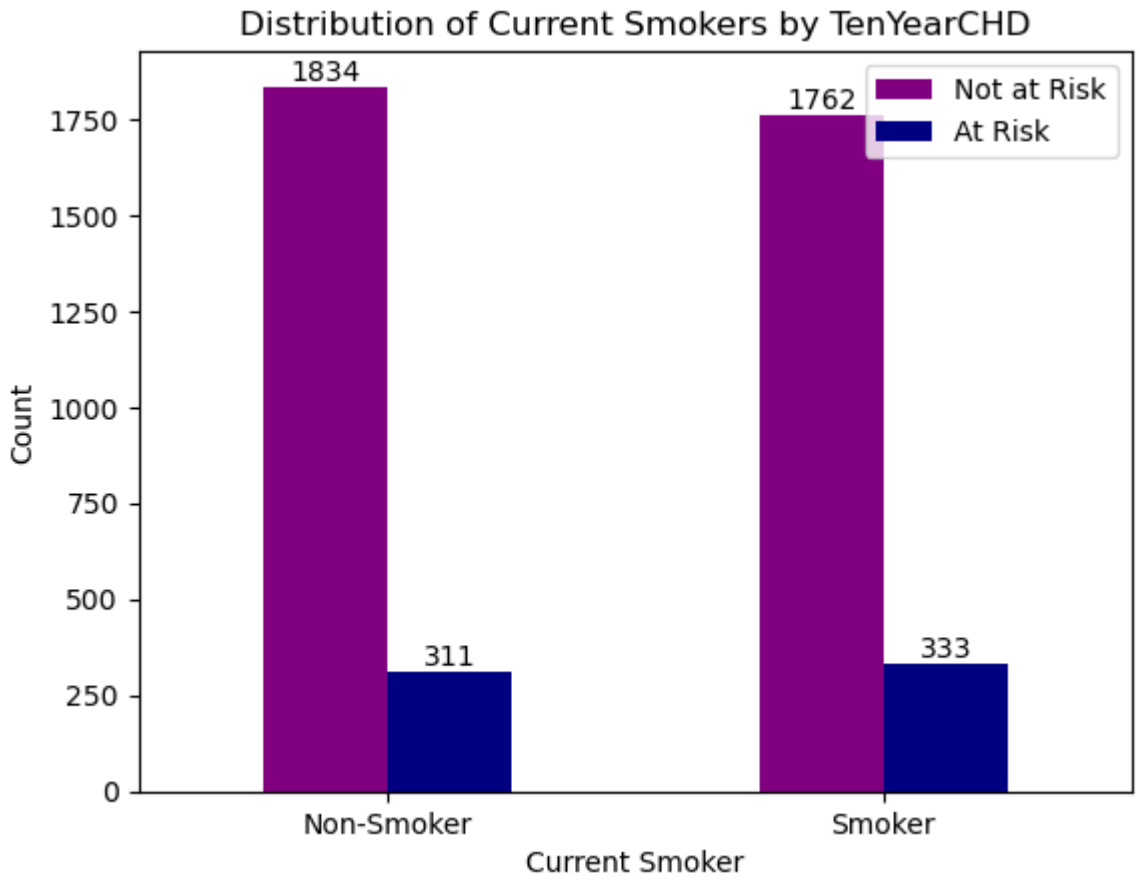
- This exploration analysis shows the age distribution in the dataset

Exploratory Data Analysis (EDA)



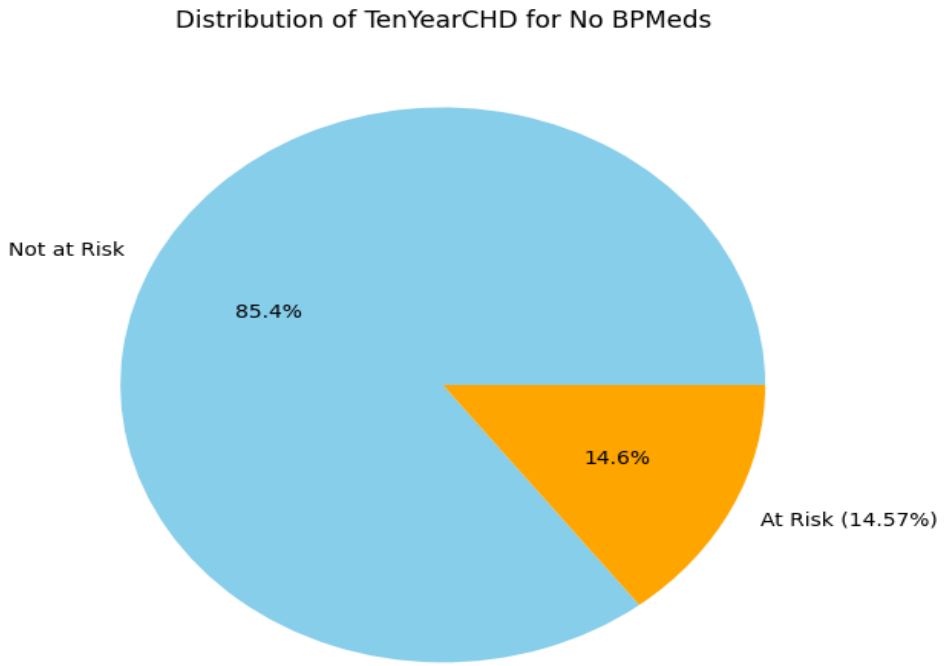
- In this exploration we look at the age distribution and the possible risk of coronary heart disease among these age groups.
- Participants in the age bracket of 50-60 years experience the highest risk while the age bracket of 30-40 years have the lowest risk.

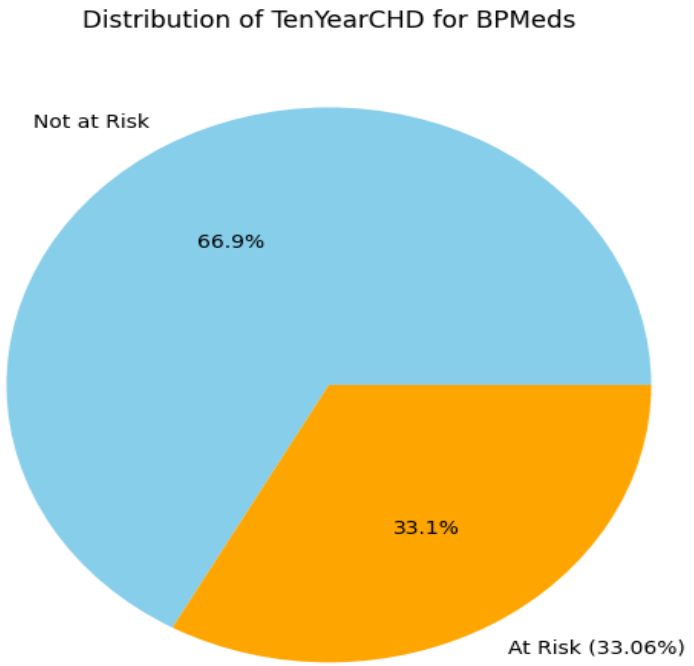
Exploratory Data Analysis (EDA)



- In this exploration we look at the current cigarette smoking and the possible risk of coronary heart disease.
- Participants who smoked were at a greater risk of developing coronary heart diseases compared to their nonsmoking counterparts.

Exploratory Data Analysis (EDA)

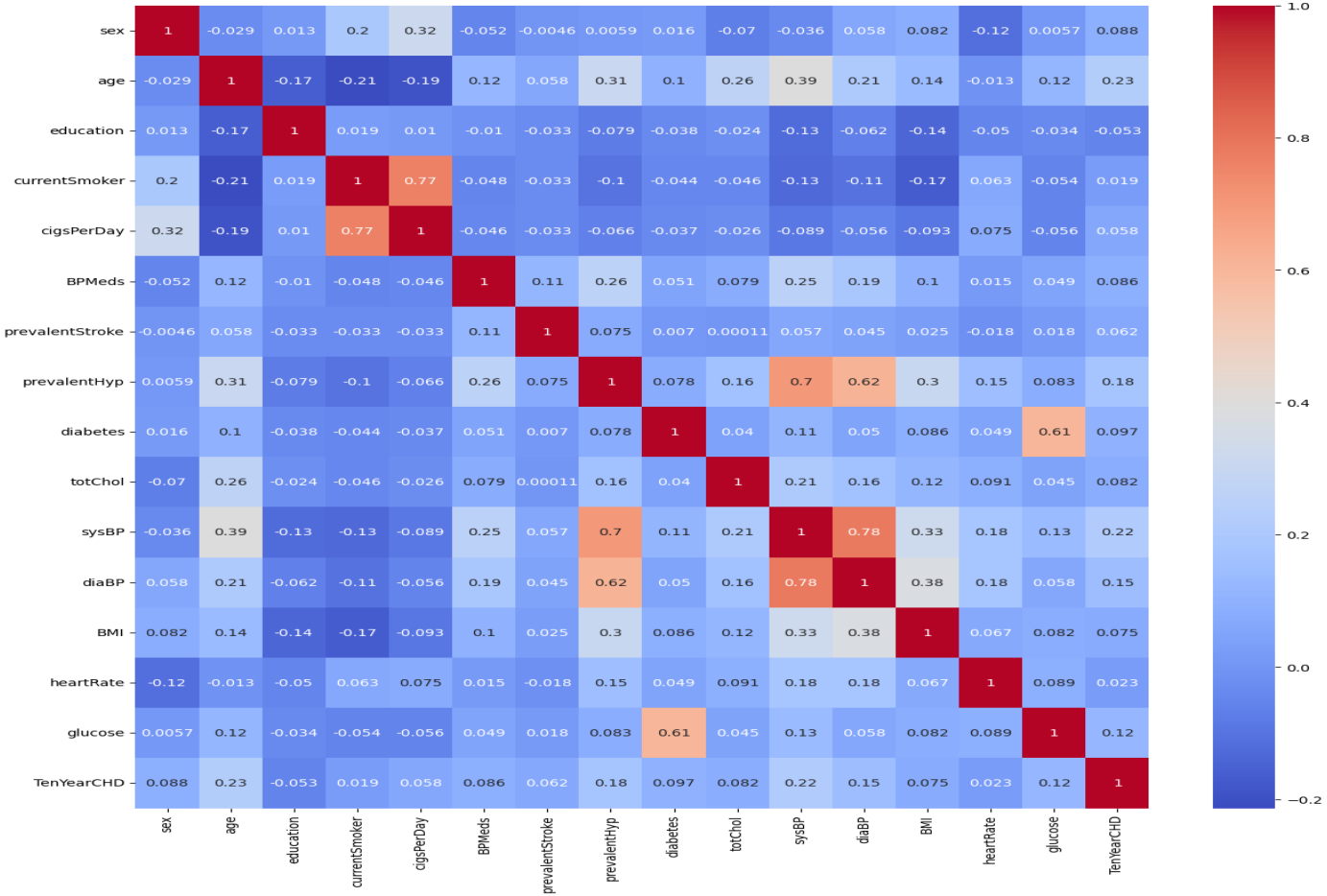




In this exploration we look at the current usage of blood pressure medications and the possible risk of coronary heart disease.

Participants who did not use any medications had a 14.6% risk while those who used medications had a higher risk rate at 33.1%

Feature Selection



- The best feature that can be used to predict the risk for heart disease are chosen to remove nonessential variables.
- The following Techniques Are Used for Feature Selection:
  - Univariate feature selection.
  - Feature importance from models.
  - Correlation analysis.
- The Most Relevant Features That Were Identified are:
  - currentSmoker, cigsPerDay, prevalentHyp, diabetes, sysBP, diaBP and glucose since they registered a high correlation greater than 0.5 as shown on the heatmap

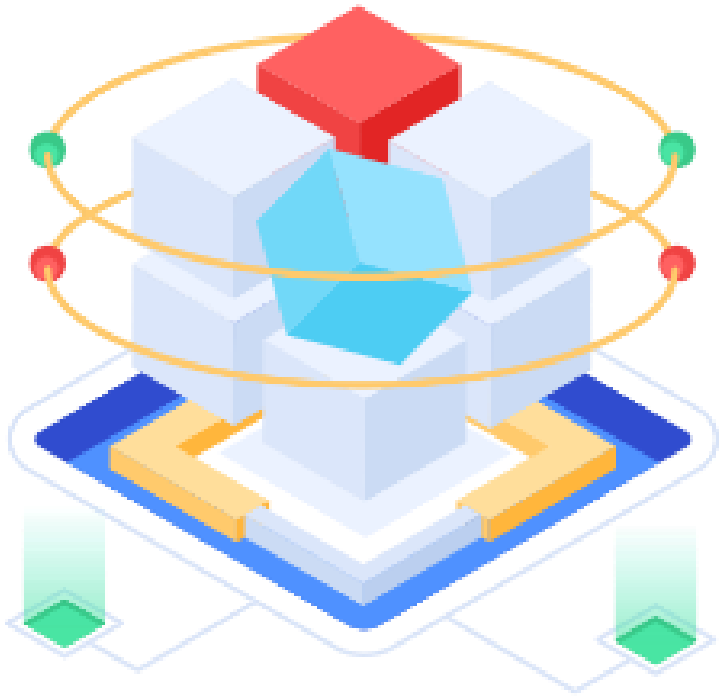
Model Selection

The following machine learning algorithms were chosen for the data analysis.

- Logistic regression.
- Decision tree.
- Random forest.
- K-Nearest Neighbours
- Support Vector Machine

Reason behind the model selections

Considerations for interpretability, performance, and suitability.



Model Training and Validation

- The training and validation process for each model includes splitting the dataset into training and test samples.

- Each model is trained and validated individually.

- The performance and accuracy of each model is evaluated based on its accuracy score, confusion matrix and ROC graph.

Model results

- Below are the performance results of the models.

Model	Accuracy	Recall	F1-Score	ROC_AUC
Decision Tree	0.75	0.85	0.85	0.51
Logistic Regression	0.86	1.00	0.92	0.63
K-Nearest Neighbors (KNN)	0.84	0.97	0.91	0.58
Support Vector Machine(SVM)	0.85	1.00	0.92	0.69
Random Forest	0.85	0.99	0.92	0.68

Conclusion

- Machine learning algorithms can be used to predict the possible risk of heart diseases based on the data it is trained with.
- Some variables are closely linked with the target variable than others.
- The best model for predicting this kind of data is logistic regression.

THANK YOU