# Capstone Project

## Cardiovascular Risk Prediction

**Dharmeshbhai Patel**

# Content

# Problem Statement:

- The objective of the project is to come up with the machine learning model to predict whether a patient has 10-year risk of developing coronary heart disease (CHD) using the residents of the town of Framingham, Massachusetts dataset.

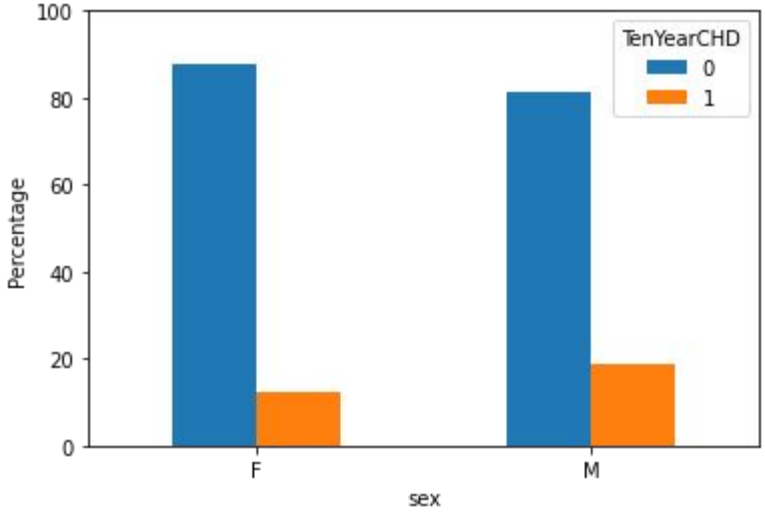
# Data Summary:

- The dataset provides the patients' information.
- It includes over 4,000 records and 15 attributes.
- Variables Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.
- Attributes:
  - Demographic:
    - Sex: male or female("M" or "F")(Nominal)
    - Age: Age of the patient(Continuous)
  - Behavioral:
    - is_smoking: Whether or not the patient is a current smoker ("YES" or "NO") (Nominal)
    - cigsPerDay: the number of cigarettes that the person smoked on average in one day. (Continuous)

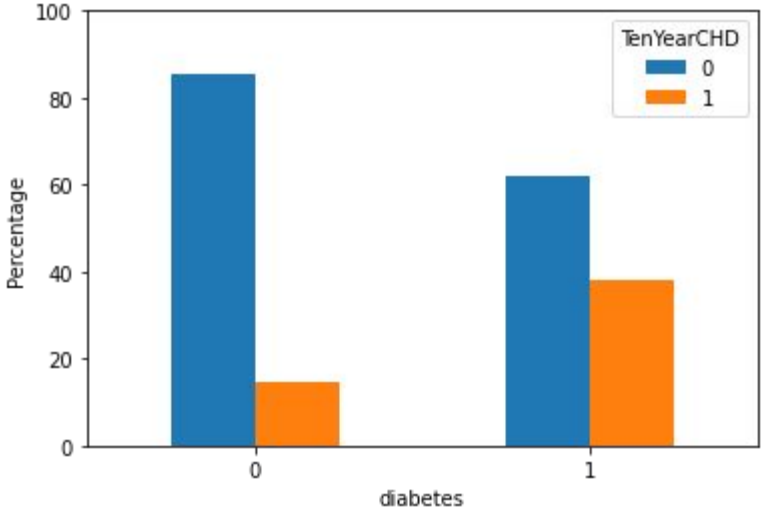
# Data Summary(contd..):

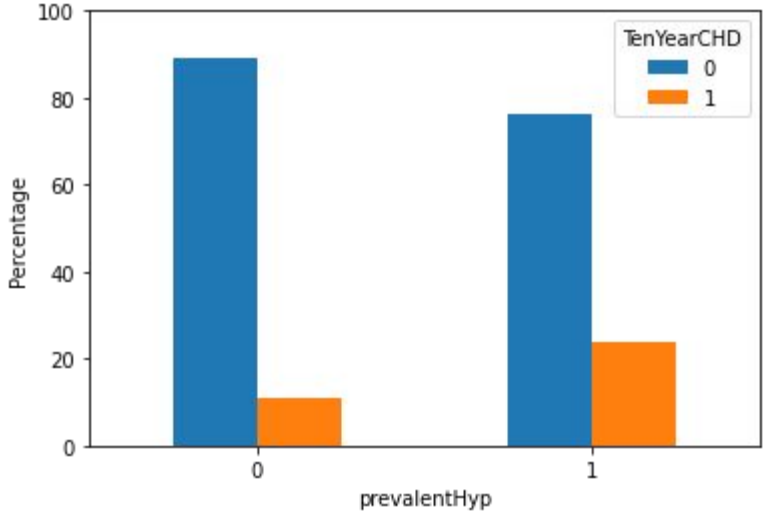
- Medical( history):
  - BPmeds: whether or not the patient was on blood pressure medication (Nominal)
  - prevalentStroke: whether or not the patient had previously had a stroke (Nominal)
  - prevalentHyp: whether or not the patient was hypertensive (Nominal)
  - diabetes: whether or not the patient had diabetes (Nominal)
- Medical(current)
  - totChol: total cholesterol level (Continuous)
  - sysBP: systolic blood pressure (Continuous)
  - diaBP: diastolic blood pressure (Continuous)
  - BMI: Body Mass Index (Continuous)
  - heartRate: heart rate (Continuous)
  - glucose: glucose level (Continuous)
- Predict variable (desired target):
  - TenYearCHD:10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) –Discrete variable

# Visualize the target and Sex variable

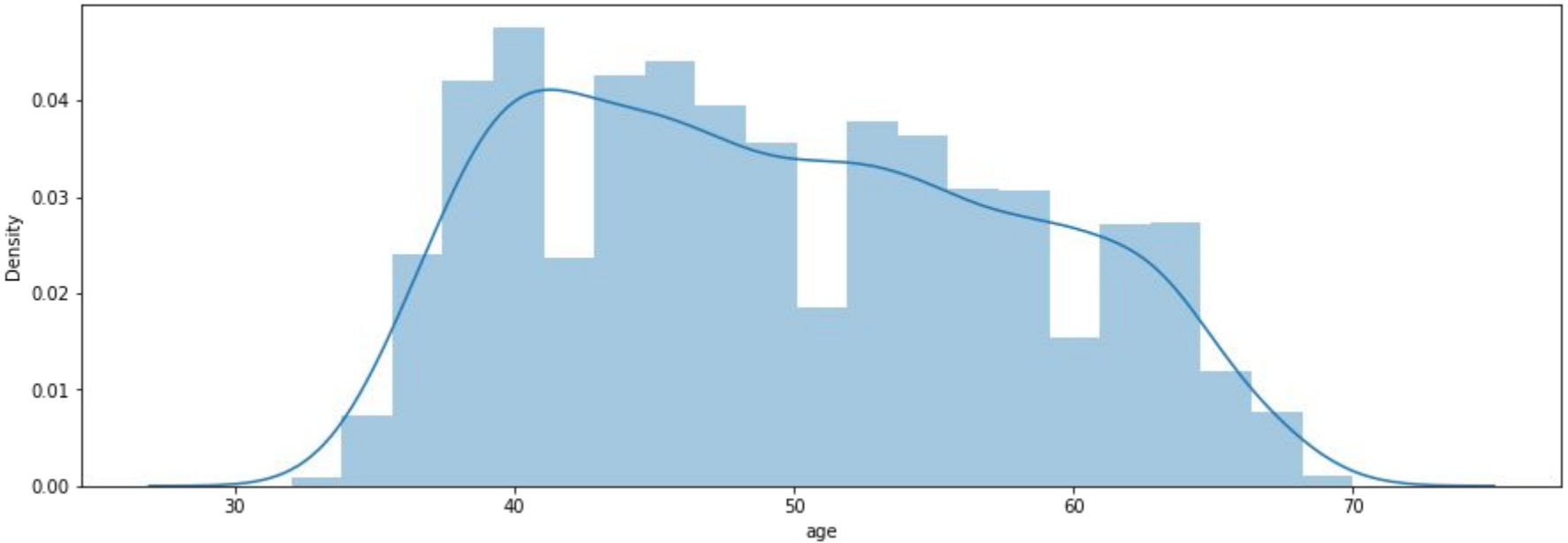


# Visualization of Diabetes and hypertensive:

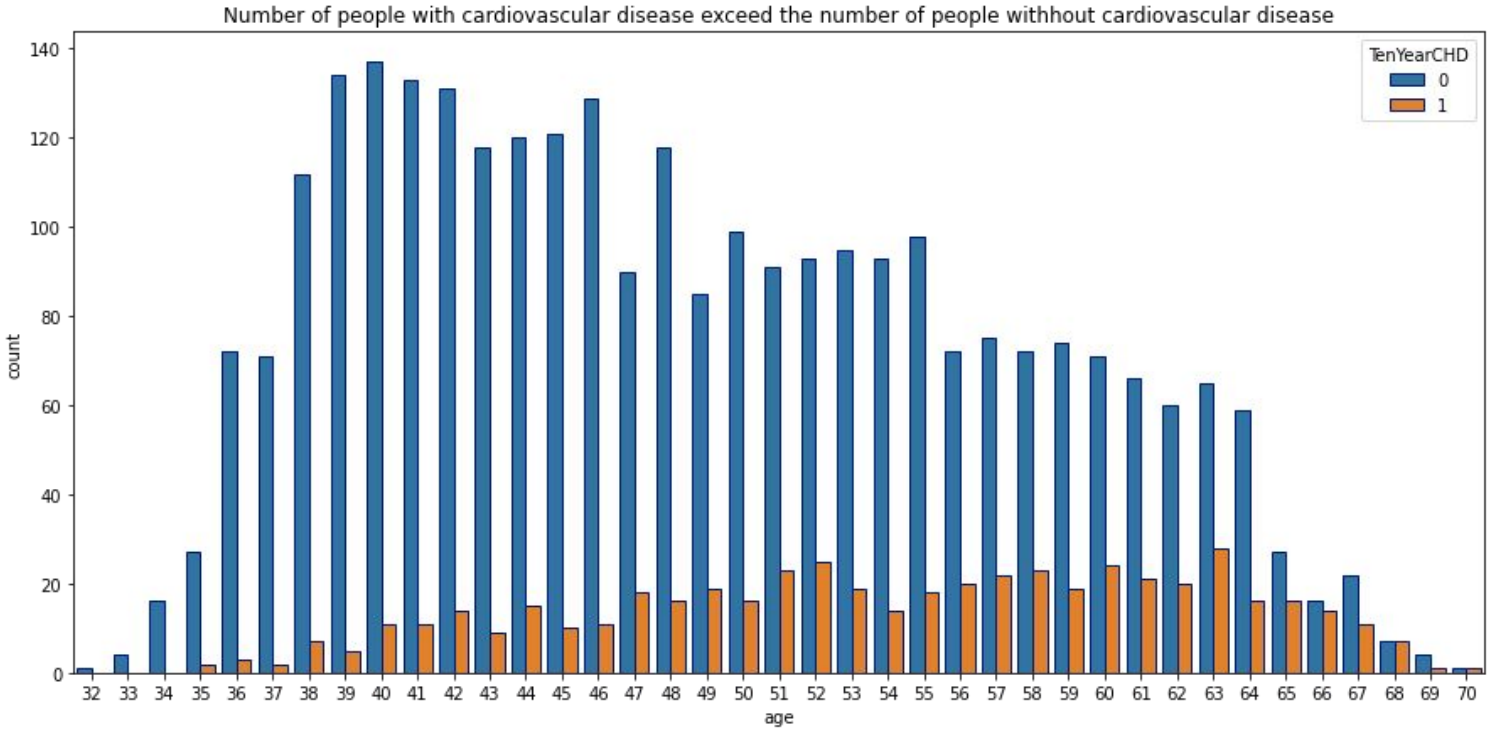


# Distribution of Age variable



# Visualize the target and age variable



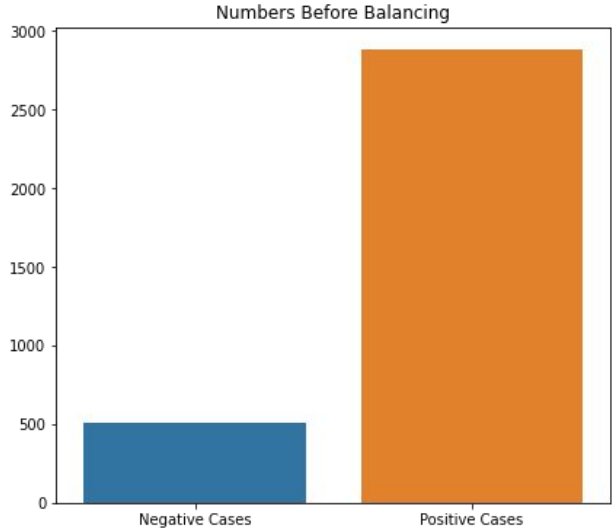
# Feature Engineering:

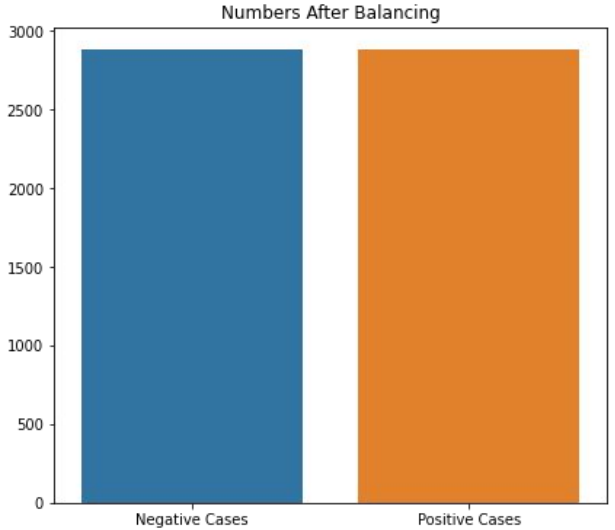
- We have the features like ‘id’ and ‘education’ which does not provide much more information so we remove that columns.
- We’ve the columns ‘sex’ and ‘is_smoking’ which are of string type so we convert them into integer by applying the function which converts the following:
  - In sex feature M(Male) will be converted to 1 and F(Female) will be converted to 0.
  - In is_smoking feature YES will be converted to 1 and NO will be converted to 0.

# Visualize the target and age variable

**First we balance our dataset because for every positive case there are about 5-6 negative cases.**

**•To handle this problem we will balance the dataset using the Synthetic Minority Oversampling Technique (SMOTE)**



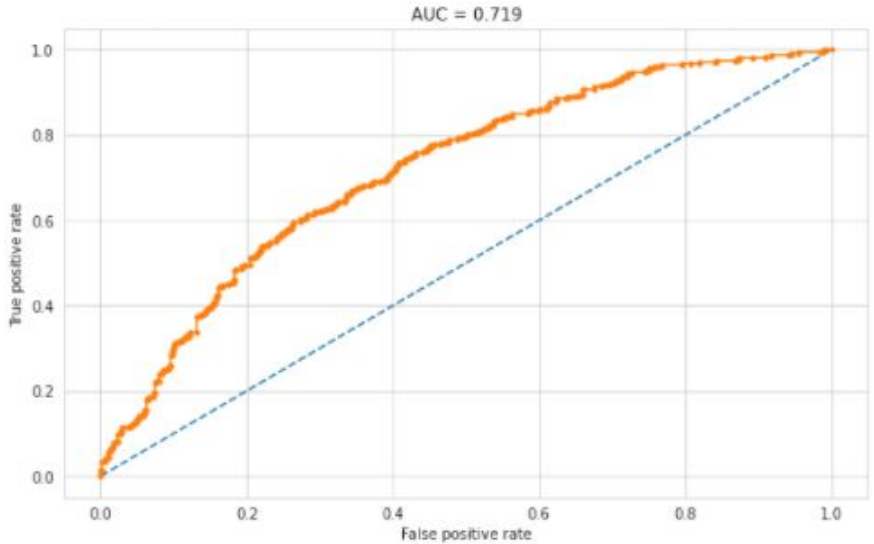
# Algorithm Used

Here, we'll be using this 4 algorithms along with GridsSearchCV for finding optimum parameters:

1. Logistic Regression
2. Random Forrest
3. XG-Boost
4. Support Vector Machine

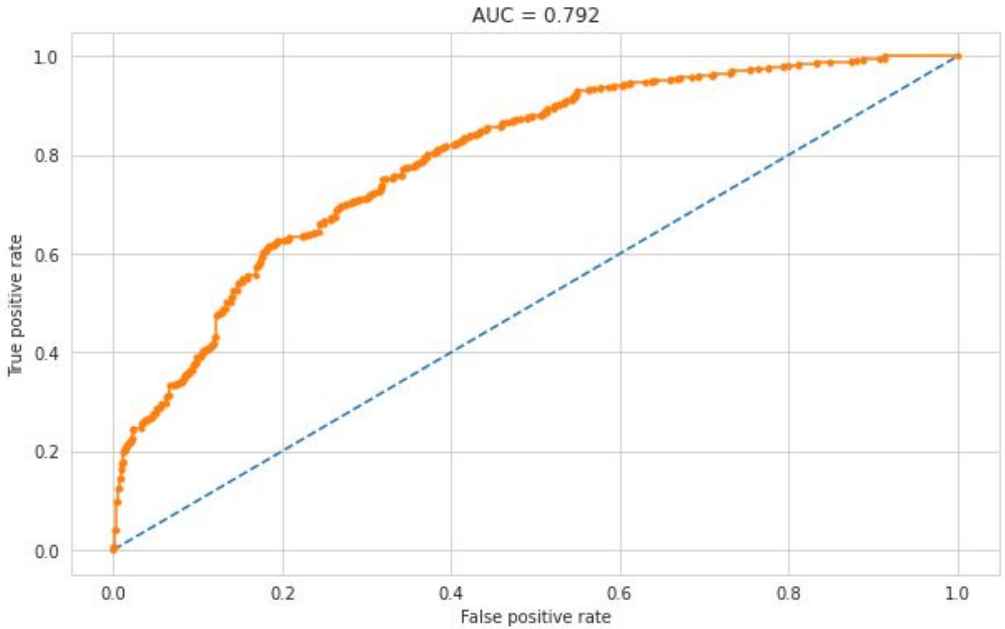
# Logistic Regression

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.65 | 0.66 | 589 |
| 1 | 0.65 | 0.67 | 0.66 | 563 |
| accuracy | | | 0.66 | 1152 |
| macro avg | 0.66 | 0.66 | 0.66 | 1152 |
| weighted avg | 0.66 | 0.66 | 0.66 | 1152 |



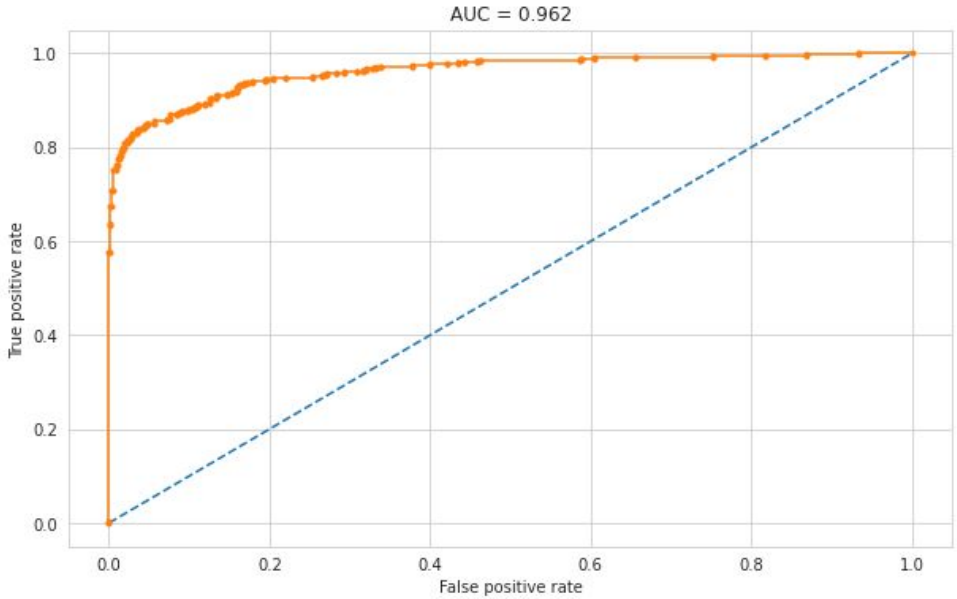
# Random Forest

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.75 | 0.66 | 0.70 | 589 |
| 1 | 0.68 | 0.77 | 0.72 | 563 |
| accuracy | | | 0.71 | 1152 |
| macro avg | 0.72 | 0.71 | 0.71 | 1152 |
| weighted avg | 0.72 | 0.71 | 0.71 | 1152 |



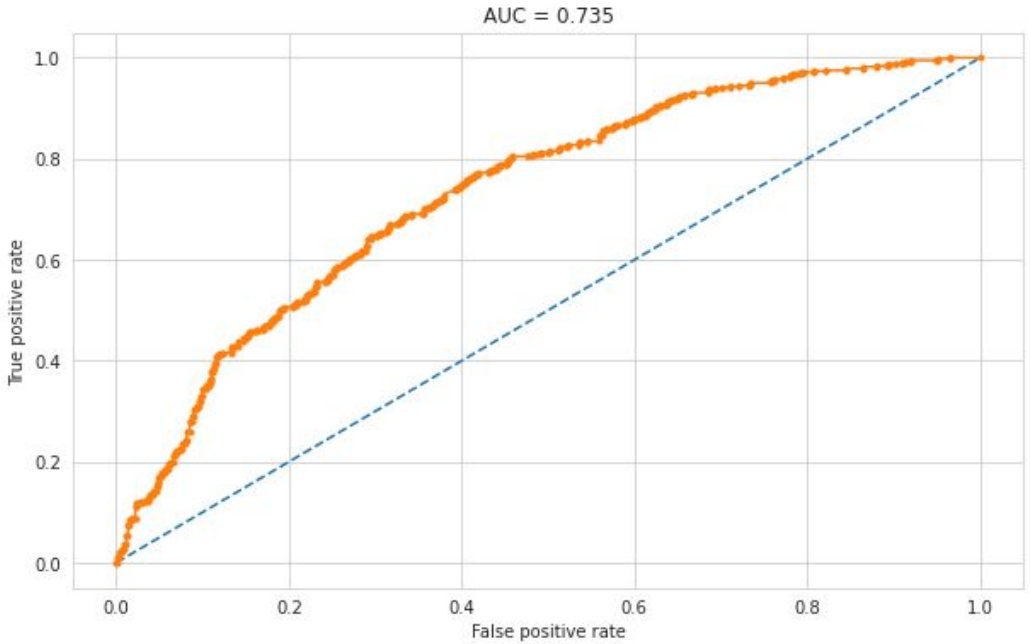
# XG-Boost

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.91 | 0.90 | 589 |
| 1 | 0.91 | 0.87 | 0.89 | 563 |
| accuracy | | | 0.89 | 1152 |
| macro avg | 0.89 | 0.89 | 0.89 | 1152 |
| weighted avg | 0.89 | 0.89 | 0.89 | 1152 |



# Support Vector Machine

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.72 | 0.59 | 0.65 | 589 |
| 1 | 0.64 | 0.75 | 0.69 | 563 |
| accuracy | | | 0.67 | 1152 |
| macro avg | 0.68 | 0.67 | 0.67 | 1152 |
| weighted avg | 0.68 | 0.67 | 0.67 | 1152 |



# Final Comparison Table

| | Test Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| Logistic regression | 0.66 | 0.65 | 0.67 | 0.66 | 0.72 |
| Random Forest | 0.71 | 0.68 | 0.77 | 0.72 | 0.79 |
| XG Boost | 0.89 | 0.91 | 0.87 | 0.89 | 0.96 |
| Support vector machine | 0.67 | 0.64 | 0.75 | 0.69 | 0.73 |

# Challenges:

- Although we have done feature selection based on their relevance to the target variable, it was challenging to come up with new engineered features that could explain hidden patterns in the data and classify our target variable better.
- We might need to work more on feature engineering and improve our precision. We might as well expect data samples with positive risk of CHD to be available in future.

# Conclusion:

- Balancing the dataset by using the SMOTE technique helped in improving the models' sensitivity.
- Slightly more males are suffering from Cardiovascular heart disease than females.
- The people who have Cardiovascular heart disease is almost equal between smokers and non smokers.
- The percentage of people who have Cardiovascular heart disease is higher among the diabetic patients and also those patients with prevalent hypertension have more risk of Cardiovascular heart disease compare to those who don't have hypertensive problem.
- The percentage of people who are on medication of blood pressure have more risk of Cardiovascular heart disease compare to those who are not on medication.
- The optimum model for predicting Cardiovascular heart disease is XG-Boost.

# Thank You