Cardiovascular Risk Prediction

Saurabh Yadav: saurabhyd423@gmail.com

Abstract: Identifying people at risk of cardiovascular diseases (CVD) is a cornerstone of preventive cardiology. Risk prediction models currently recommended by clinical guidelines are typically based on a limited number of predictors with sub-optimal performance across all patient groups. Data-driven techniques based on machine learning (ML) might improve the performance of risk predictions by agnostically discovering novel risk predictors and learning the complex interactions between them. We tested whether ML techniques based on a state-of-the-art automated ML framework (AutoPrognosis) predicts the patient has a 10-year risk of future coronary heart disease (CHD).

1) INTRODUCTION:

Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. According to the WHO, an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three quarters of these deaths took place in low- and middle-income countries.

Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. In the United States, for example, it is estimated that someone has a heart attack every 40 seconds and about **805,000 Americans** have a heart attack every year (CDC 2019).

Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

In this project, I've built a model for predicting whether a patient has a 10-year risk of developing coronary heart disease(CHD) using different Machine Learning techniques on the Framingham, Massachusetts town dataset.

2) OBJECTIVE:

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.

3) DATA SUMMARY

The data set is publicly available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The data set provides the patients' information. It includes over 4,000 records and 15 attributes. 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")
- **Age**: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) Behavioral
- **is_smoking**: whether or not the patient is a current smoker ("YES" or "NO")
- **Cigs Per Day**: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Medical(history)

- **BP Meds**: whether or not the patient was on blood pressure medication (Nominal)
- **Prevalent Stroke**: whether or not the patient had previously had a stroke (Nominal)
- **Prevalent Hyp**: whether or not the patient was hypertensive (Nominal)
- **Diabetes**: whether or not the patient had diabetes (Nominal)

Medical(current)

- **Tot Chol**: total cholesterol level (Continuous)
- **Sys BP**: systolic blood pressure (Continuous)
- **Dia BP**: diastolic blood pressure (Continuous)

- **BMI**: Body Mass Index (Continuous)
- **Heart Rate**: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
- **Glucose**: glucose level (Continuous)

Predict variable (desired target)

- 10 year risk of developing **coronary heart disease (CHD)** — (binary: “1”, means “There is a risk”, “0” means “There is no risk”).

4) TOOL DEVELOPMENT

The full code for this article can be found here. It is implemented in Python and different classification algorithms are used. Below is a brief description of the general approach that I employed:

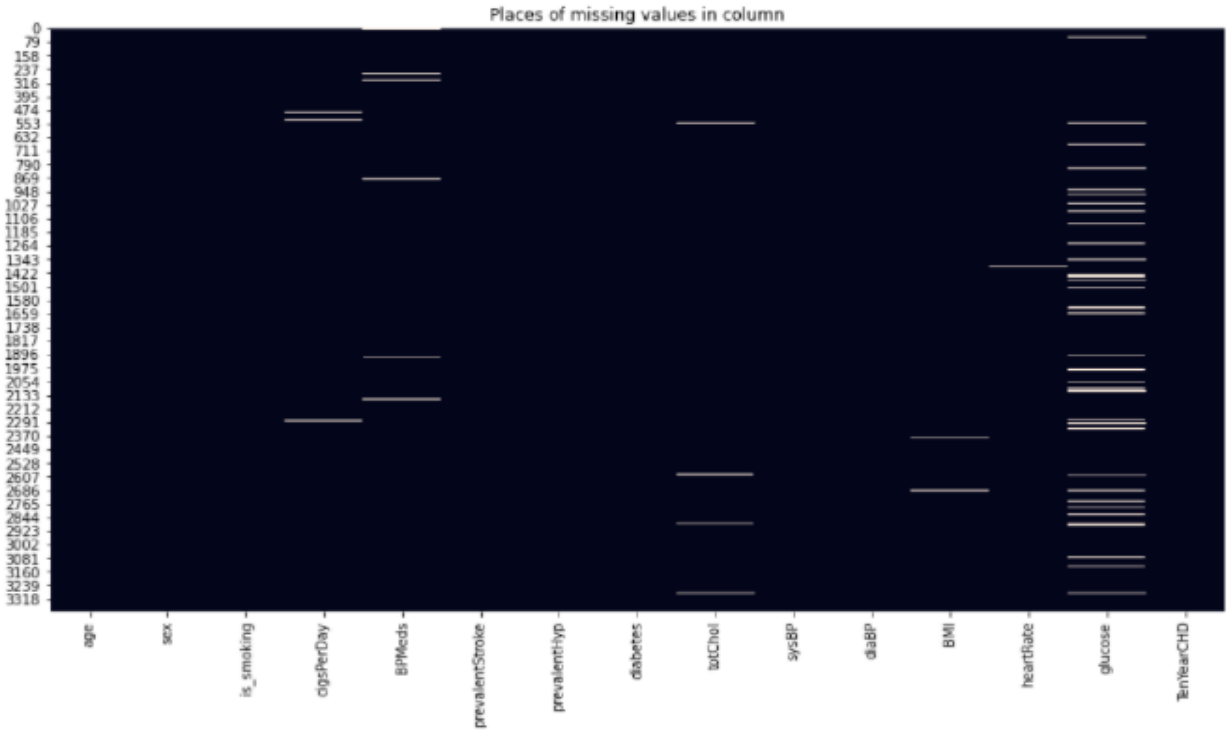
1) **Data cleaning and preprocessing**: Here I checked and dealt with missing and duplicate variables from the data set as these can grossly affect the performance of different machine learning algorithms (many algorithms do not tolerate missing data).
2) **Exploratory Data Analysis**: Here I wanted to gain important statistical insights from the data and the things that I checked for were the distributions of the different attributes, correlations of the attributes with each other and the target variable and I calculated important odds and proportions for the categorical attributes.
3) **Feature Selection**: Since having irrelevant features in a data set can decrease the accuracy of the models applied, I used Tree-based: SelectFromModel which is an embedded method that uses algorithms that have built-in feature selection methods which were later used to build different models.
4) **Model development and comparison**: I used four classification models, i.e., Logistic Regression, K-Nearest Neighbors, Decision Trees and Support Vector Machine, After which I compared the performance of the models using their accuracy and F1 scores. I then settled with the best performing model.

4.1) Data Cleaning and preprocessing:

We drop the **education** and **id** columns because it has no correlation with heart disease.

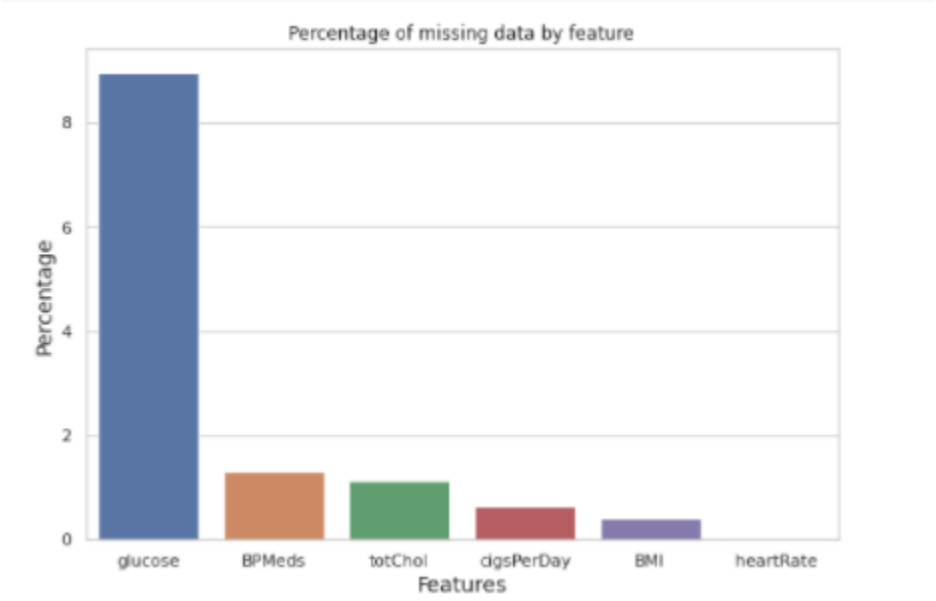
	age	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	64	F	YES	3.0	0.0	0	0	0	221.0	148.0	85.0	NaN	90.0	80.0	1
1	36	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0	0
2	46	F	YES	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.0	0
3	50	M	YES	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.0	1
4	64	F	YES	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.0	0

There were no duplicate entries in the data set but some had missing values and the table below gives a summary of these:
Let's visualize the location of missing values.



Percentage of missing values per feature is shown below:

	Total	Percentage
glucose	304	8.967552
BPMeds	44	1.297935
totChol	38	1.120944
cigsPerDay	22	0.648968
BMI	14	0.412979
heartRate	1	0.029499

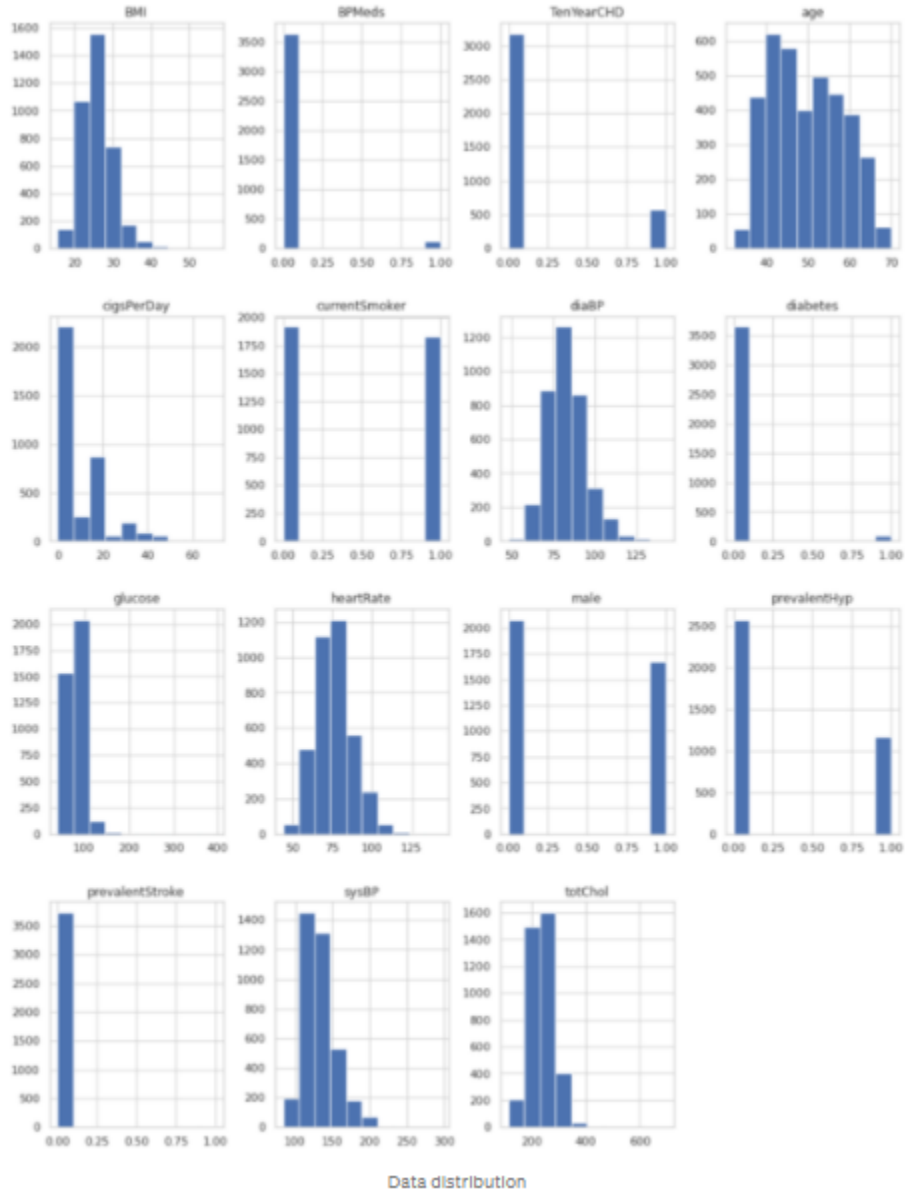


At **8.97%**, the blood glucose entry has the highest percentage of missing data. The other features have very few missing entries.

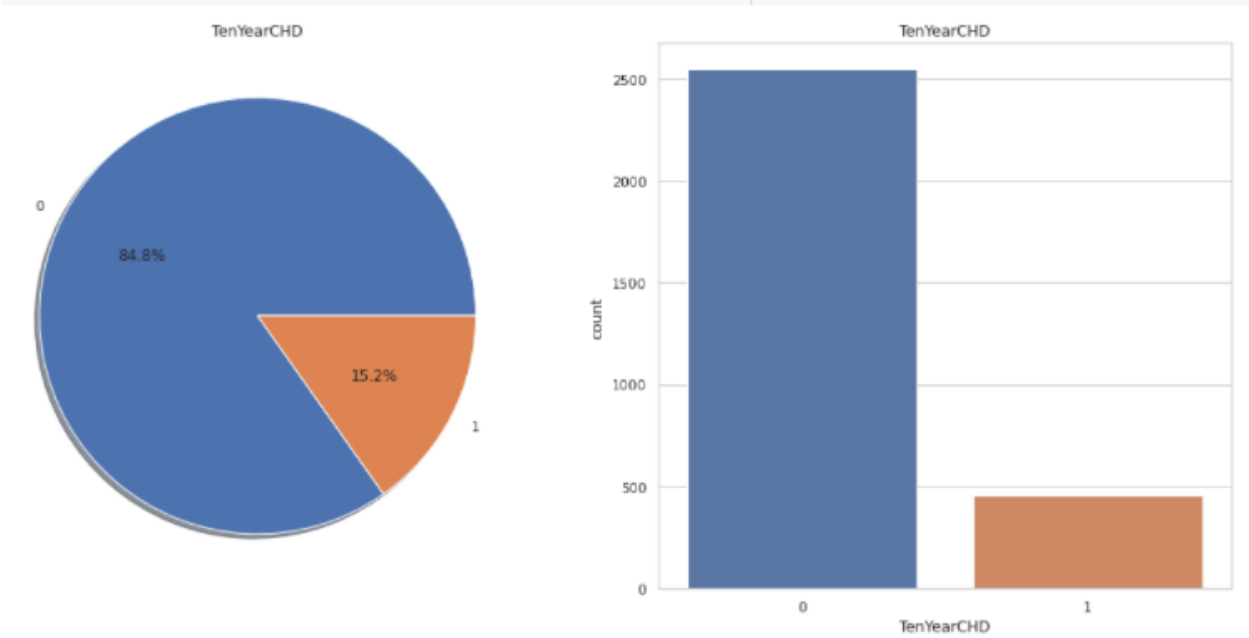
Since the missing entries account for only **11%** of the total data, we can drop these entries without losing alot of data.

4.2) Exploratory Data Analysis

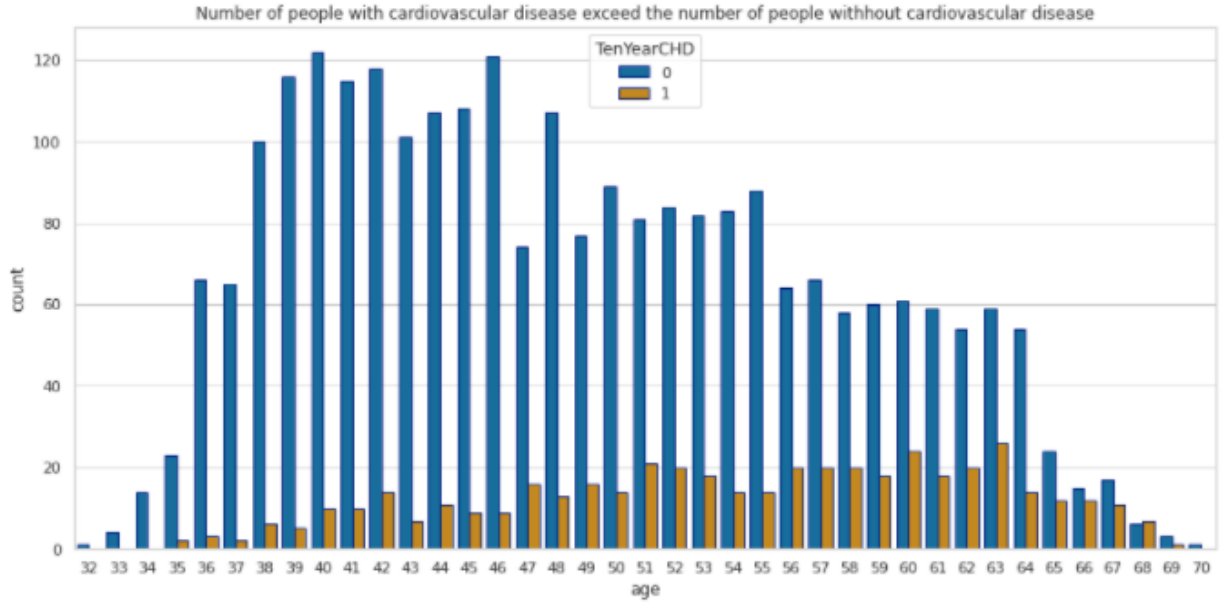
The first step was to check the distribution of different attributes and this was best visualized by histograms.



It is easy to pick out the categorical and continuous variables from the distribution plots. Also, it can be seen that none of the respondents had prevalent stroke and very few were diabetic, on blood pressure medication or hypertensive. These distributions also raised the suspicion that the data set might not be properly balanced and to confirm this I compared the number of positive and negative cases and true to my suspicions there were 2547 respondents without CHD and 572 patients with CHD.



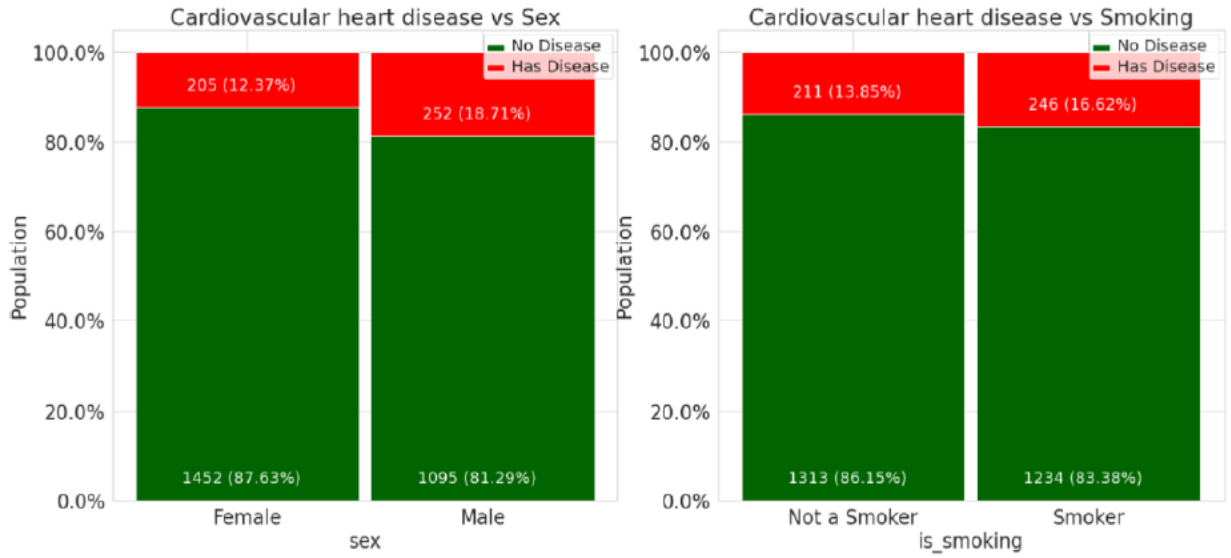
Let's look at the number of people with cardiovascular disease exceeding the number of people without cardiovascular disease.

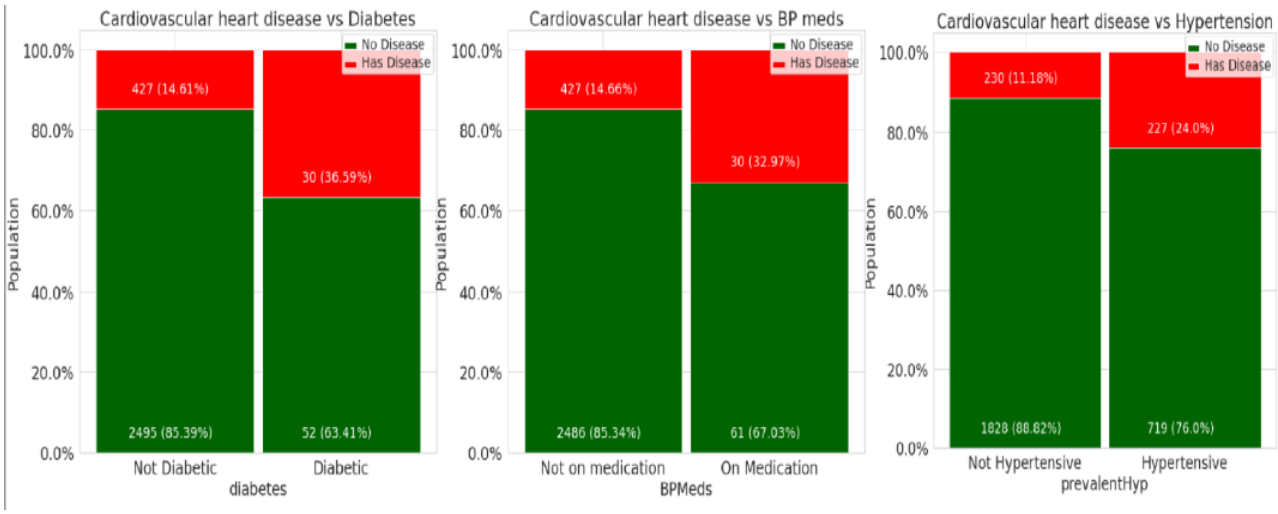


As we can see in above plot The people with the highest risk of developing heart disease are betwwen the ages of **51 and 63.**

Because the number of sick people generally increases with age.

To gain more insight into the data I checked the proportions of positive and negative cases in each category.



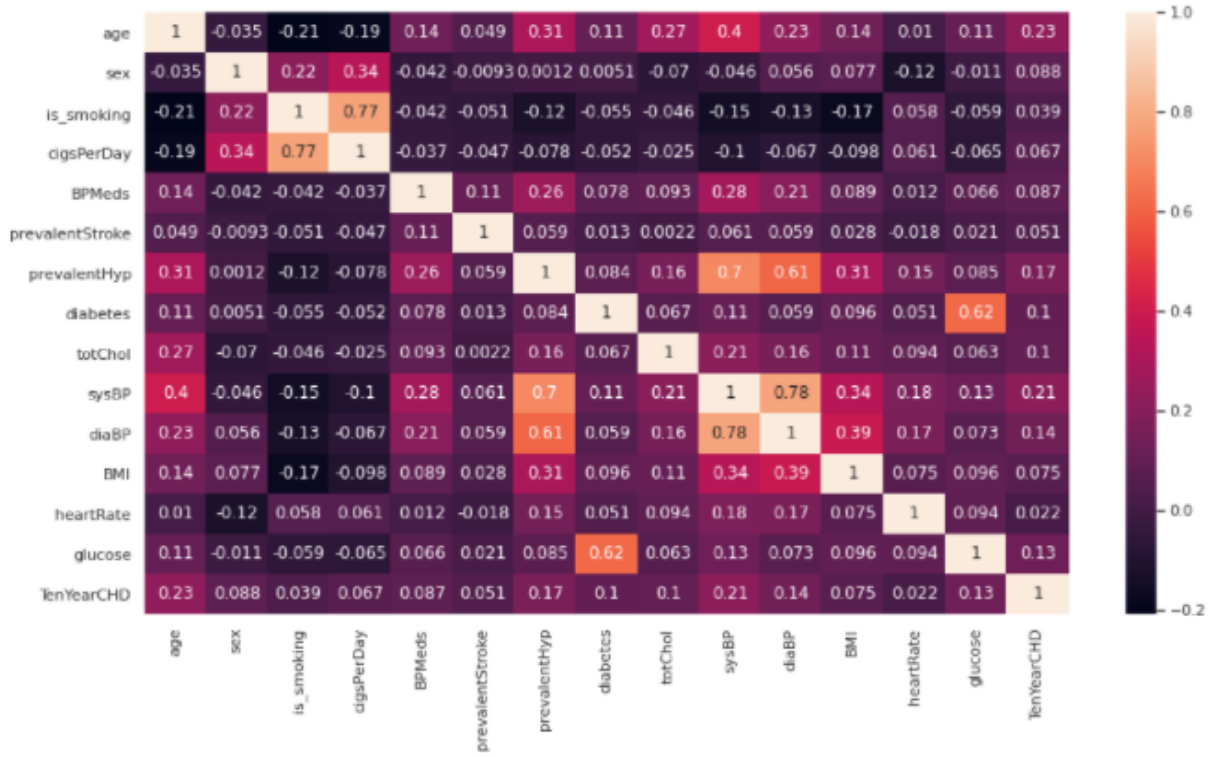


Due to the imbalanced nature of the data set it was difficult to make conclusions but based on what is observed but these are the conclusions that could be drawn:

- Slightly more males are suffering from Cardiovascular heart disease than females.
- The number of 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 compared to those who don't have hypertensive problems.

- The percentage of people who are on medication of blood pressure have more risk of Cardiovascular heart disease compared to those who are not on medication.

The final step was to check the correlation of the different features with the target variable and with each other as this would not only give a good estimate of the strength of the features as predictors of coronary heart disease but also reveal any co-linearity among the features.



From the above correlation plot we can conclude that,

- There are no features with more than 0.5 correlation with the Ten year risk of developing CHD and this shows that the features are poor predictors. However the features with the highest correlations are age, prevalent hypertension(prevalentHyp) and systolic blood pressure(sysBP).
- Also there are a couple of features that are highly correlated with each other and it makes no sense to use both of them in building a machine learning model.

These includes:

- Blood glucose and diabetes;
- systolic and diastolic blood pressures;
- cigarette smoking and the number of cigarettes smoked per day.

Therefore we need to carry out feature selection to pick the best features.

4.3) Feature Selection

The results from the correlation matrix prompt the need for feature selection. To do this I employed the Select From Model algorithm which is a wrapper method built around the random forest classification algorithm. It tries to capture all the important, interesting features in a data set with respect to an outcome variable.

It works in the following way:

- First, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features).
- Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where a higher score means a more important feature.
- At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z-score than the maximum Z-score of its shadow features) and constantly removes features which are deemed highly unimportant.
- Finally, the algorithm stops either when all features get confirmed or rejected or it reaches a specified limit of random forest runs.

After running the algorithm for 100 iterations the top selected features were: Age, total cholesterol, systolic blood pressure, diastolic blood pressure, BMI, heart rate and blood glucose.

I then calculated the odds ratio of the top features and the ten year risk of developing CHD and these were the results:

	5%	95%	Odds Ratio
age	1.010284	1.035552	1.022840
totChol	0.995854	1.000481	0.998165
sysBP	1.017473	1.032083	1.024752
diaBP	0.958523	0.983091	0.970729
BMI	0.922652	0.972119	0.947062
heartRate	0.962550	0.978167	0.970327
glucose	1.001974	1.009041	1.005501

.

Holding all other features constant, the odds of getting diagnosed with heart disease increases with about 2% for every increase in age and systolic blood pressure.

The other factors show no significant positive odds.

In the output, **Iterations** refer to the number of times the model iterates over the data, trying to optimize the model.

Explanation of some of the terms in the summary table:

- **coef** : the coefficients of the independent variables in the regression equation.

- **Log-Likelihood** : the natural logarithm of the Maximum Likelihood Estimation(MLE) function. MLE is the optimisation process of finding the set of parameters which result in best fit.
- **LL-Null** : the value of log-likelihood of the model when no independent variable is included(only an intercept is included).
- **Pseudo R-squ.** : a substitute for the R-squared value in Least Squares linear regression. It is the ratio of the log-likelihood of the null model to that of the full model.

4.4) Model Development And Comparison

Since our dataset is imbalanced i.e for every positive case there are about 5-6 negative cases. We may end up with a classifier that is biased to the negative cases. The classifier may have a high accuracy but poor precision and recall.

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

SMOTE :

SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are then added between the chosen point and its neighbors.

SMOTE algorithm works in 4 simple steps:

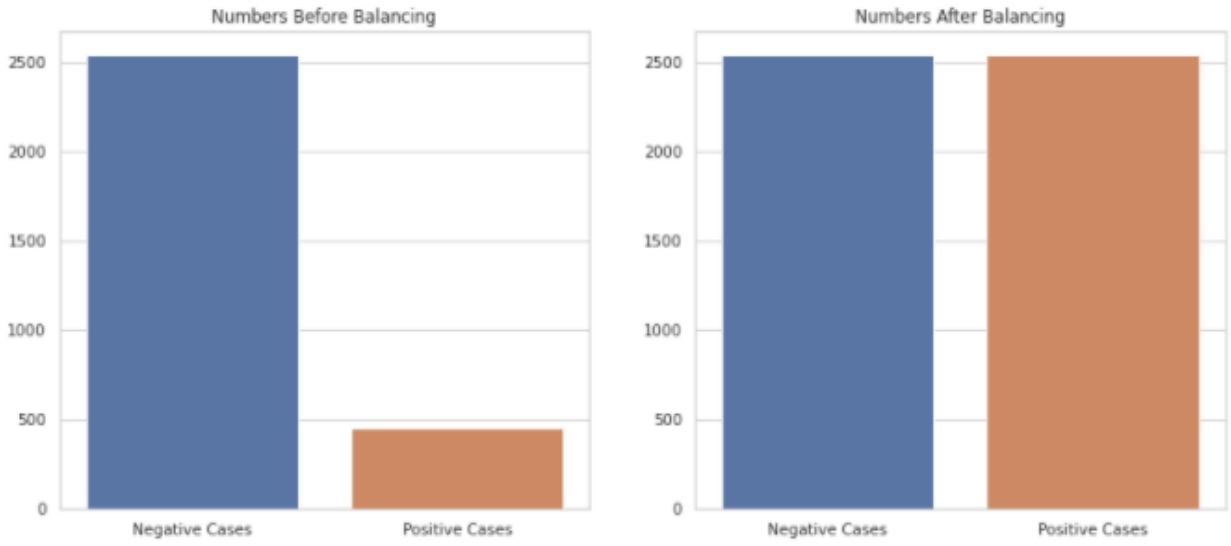
- Choose a minority class as the input vector.
- Find its k nearest neighbors (k_neighbors is specified as an argument in the SMOTE() function).
- Choose one of these neighbors and place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbor.
- Repeat the steps until the data is balanced.

This procedure can be used to create as many synthetic examples for the minority class as are required. It suggests first using random undersampling to trim the number of examples in the majority class, then using SMOTE to oversample the minority class to balance the class distribution.

As seen after applying SMOTE, the new dataset is much more balanced.

```
Original dataset shape 3004

Resampled dataset shape 5094
```



After balancing the data set, I scaled the features to speed up the training of the classifiers and then split the data into a training and test set at a ratio of 0.8 to 0.2 respectively.

Using the training set, I trained four classifiers, i.e.,:

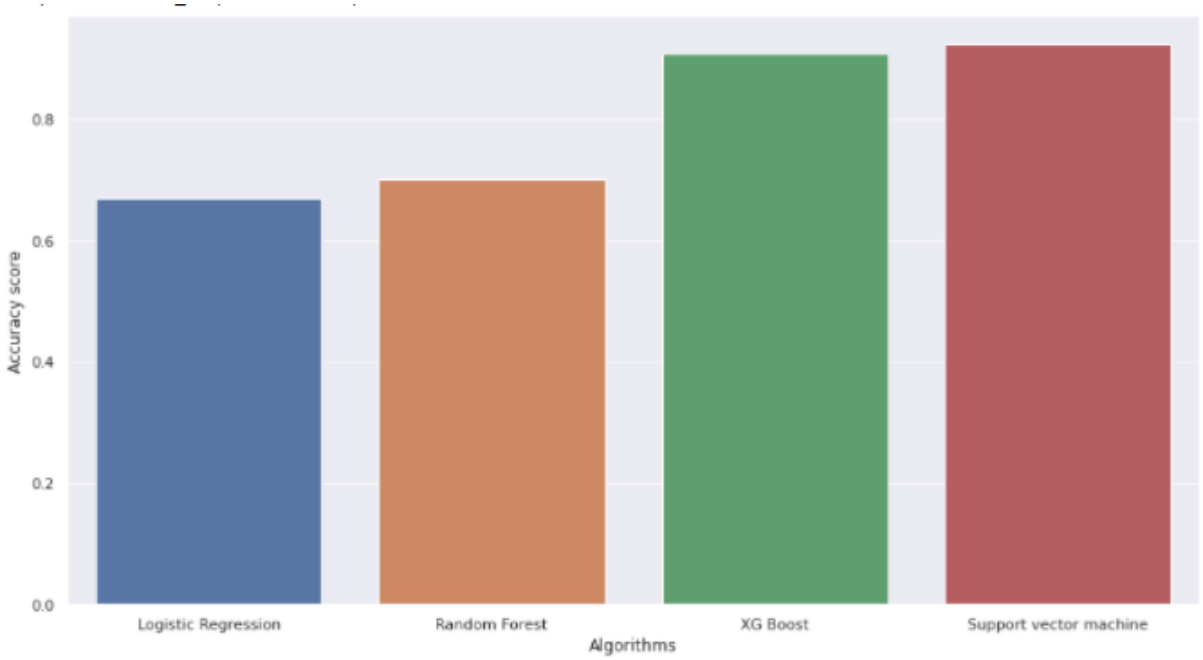
1. **Logistic Regression:** Logistic regression aims to measure the relationship between a categorical dependent variable and one or more independent variables (usually continuous) by plotting the dependent variables' probability scores.
2. **Random Forest:** Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.
3. **XGboost:** XGBoost stands for eXtreme Gradient Boosting. The name xgboost, though, actually refers to the engineering goal to push the limit of computational resources for boosted tree algorithms.
4. **Support Vector Machine (SVM):** Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

After training each model and tuning their hyper-parameters using grid search, I evaluated and compared their performance using the following metrics:

1. **The accuracy score:** which is the ratio of the number of correct predictions to the total number of input samples. It measures the tendency of an algorithm to classify data correctly.
2. **The F1 Score**: Which is defined as the weighted harmonic mean of the test's precision and recall. By using both precision and recall it gives a more realistic measure of a test's performance. (Precision, also called the positive predictive value, is the proportion of positive results that truly are positive. Recall, also called sensitivity, is the ability of a test to correctly identify positive results to get the true positive rate).
3. **The Area under the ROC Curve (AUC):** Which provides an aggregate measure of performance across all possible classification thresholds. It gives the probability that the model ranks a random positive example more highly than a random negative example.

Here are the results:

	Test Accuracy	Precision	Recall	F1 Score	AUC
Logistic regression	0.67	0.65	0.68	0.67	0.72
Random Forest	0.70	0.67	0.74	0.71	0.78
XG Boost	0.91	0.91	0.90	0.90	0.96
Support vector machine	0.92	0.91	0.93	0.92	0.98



Observation from above table:

- **XGBoost**, **Support vector machine** gives highest Accuracy, Recall, Precision and AUC score.
- Highest recall is given by **Support vector machine**
- Highest AUC is given by **Support vector machine**

Overall we can say that the support vector machine was the best performing model across all metrics. It's best parameters were a radial kernel, a C value of 10 and a gamma value of 1. Its high AUC and F1 score also show that the model has a high true positive rate and is thus sensitive to predict if one has a high risk of developing CHD , i.e., getting a heart attack within 10 years.

5) CHALLENGES

- Handling the missing values.
- Making data more accurate.
- Selection of important features.

6) CONCLUSION

- The number of people who have Cardiovascular heart disease is almost equal between smokers and non-smokers.
- The top features in predicting the ten year risk of developing Cardiovascular Heart Disease are 'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'.
- The Support vector machine with the radial kernel is the best performing model in terms of accuracy and the F1 score and Its high AUC-score shows that it has a high true positive rate.
- Balancing the dataset by using the SMOTE technique helped in improving the models' sensitivity.
- With more data(especially that of the minority class) better models can be built.