# Cardiovascular Risk Prediction

Kirtesh Verma, Pravin Bejjo, and

Sahil Pardeshi

Data Science Trainees,

Almabetter, Nashik

## ➢ Introduction:

Heart disease is one the major cause of morbidity and mortality globally. A heart attack happens when the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked and the heart can't get oxygen. If blood flow isn't restored quickly, the section of heart muscle begins to die. Doctors and Scientists across the globe have started to look into Machine Learning Techniques to develop screening tools. In this project, we shall be giving you a walk through on the development of a screening tool for predicting whether a patient has a 10-year risk of developing coronary heart disease (CHD) based on their present health conditions using different Machine Learning Techniques.

## ➢ Problem Statement:

Heart disease is the leading cause of morbidity and mortality worldwide, killing more people each year than any other cause. In this project, we shall be giving you a walk through on the development of a screening tool for predicting whether a patient has a 10-year risk of developing coronary heart disease (CHD) using different Machine Learning techniques. The given dataset provides the patients' information. It includes over 3,390 records and 17 attributes. Each attribute is a potential risk factor. There are both demographic, behavioural, and medical risk factors given for the analysis.

## ➢ Data-set description:

**Breakdown of Our Features:**

- **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) Behavioural**
- **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)

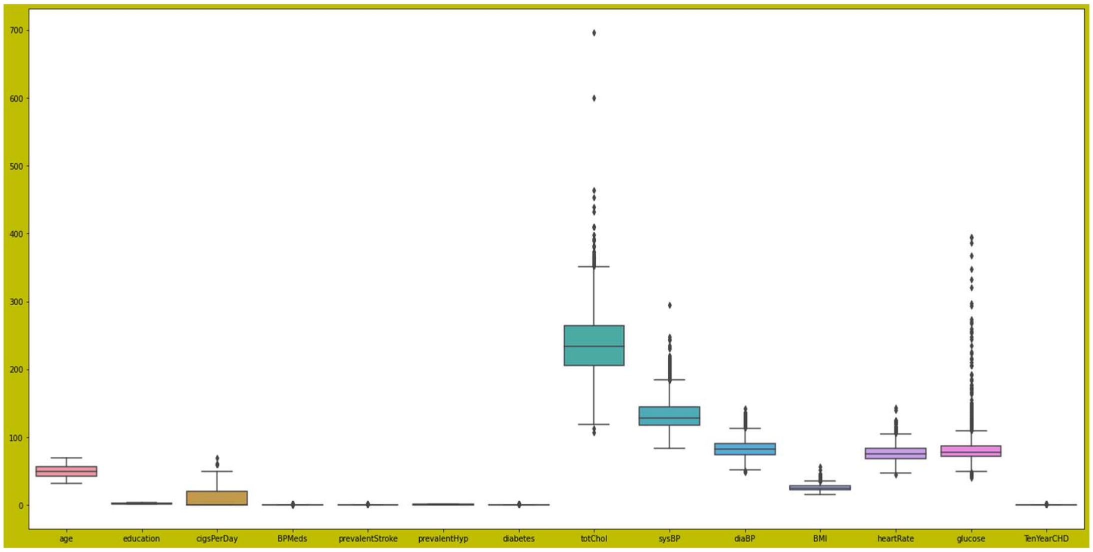
- **10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”) -DV**

## ➢ Steps Involved

### I. Performing EDA (Exploratory data analysis) and data preprocessing

1) Exploring head and tail of the data to get insights on the given data.
2) Checking the Null values or missing values are present in the dataset or not.
3) Check the duplicate values.
4) Check the majority and minority set of a target variable.
5) Check if any of the NaN values belongs to the minority class.
6) **KNNImputer** shall be used to impute the NaN values for continuous data.

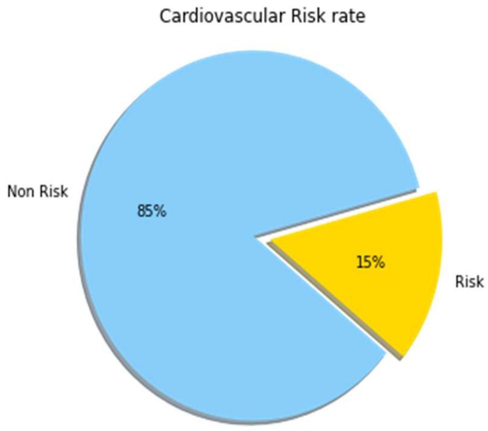
7) **SimpleImputer** shall be used to impute the NaN values for categorical data.
8) **missing_value_continuous** function to handle missing values of continuous data.
9) **variables missing_value_categorical** function to handle missing values of categorical data.
10) Treating with outliers, some of the outliers are important for our dataset so, we cannot remove them. This would further affect our machine learning model.



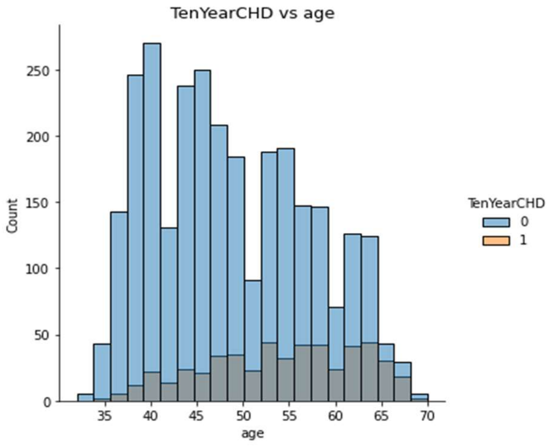
## II. Drawing conclusion from the data

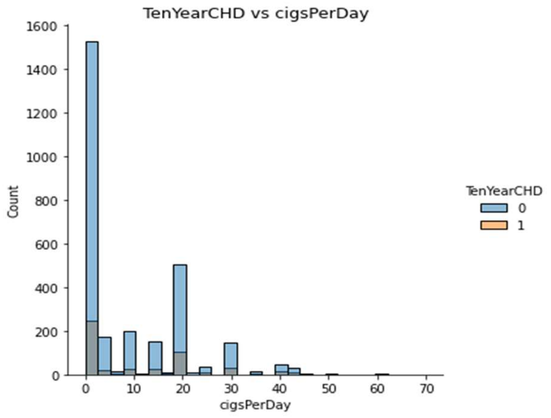
Plotting necessary graphs which provides relevant information on our data like:

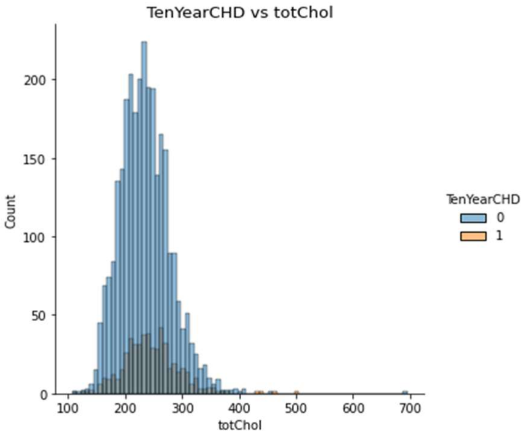
1) Let us plot the majority and minority set of the target variable

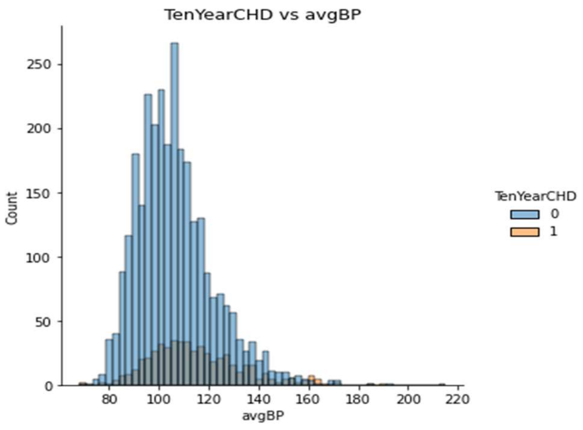


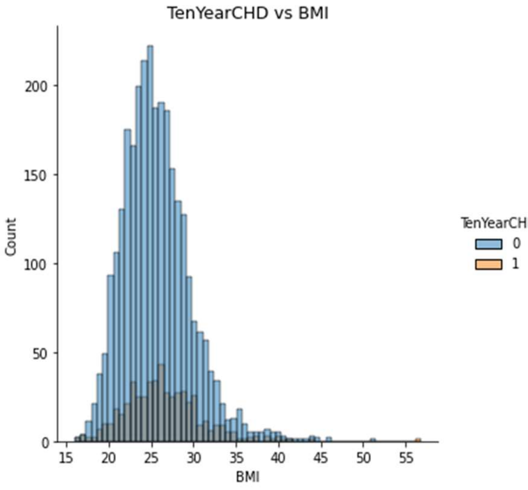
## **2)** EDA on continuous features.

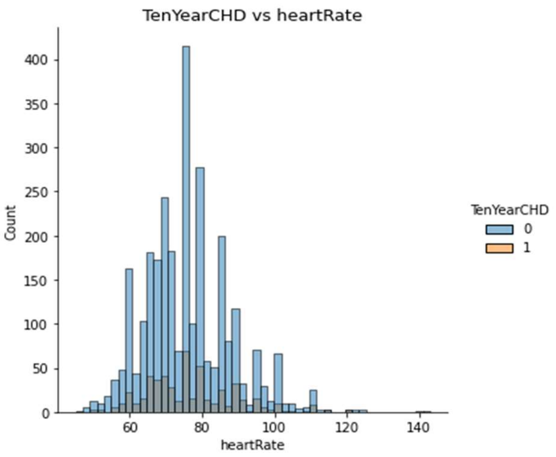


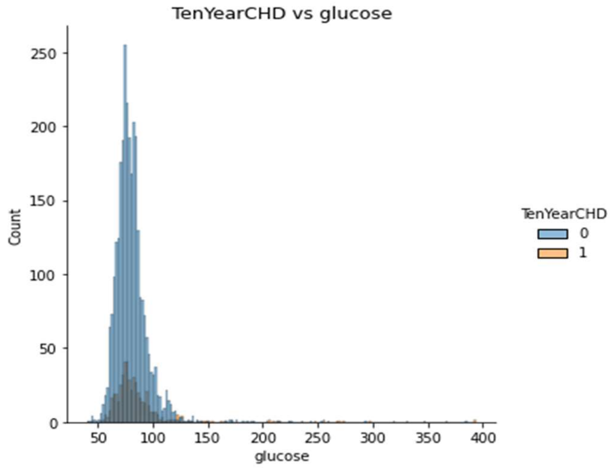






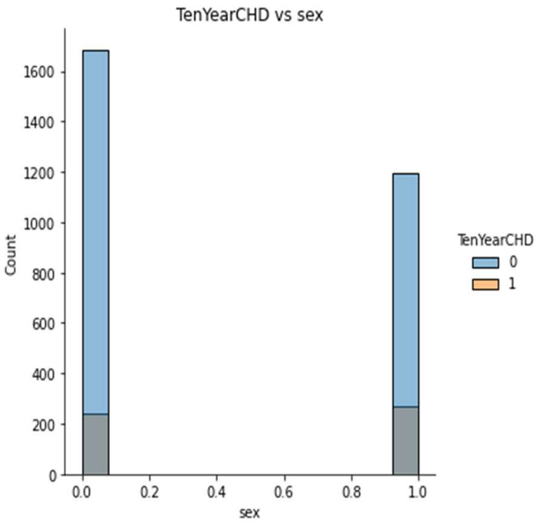


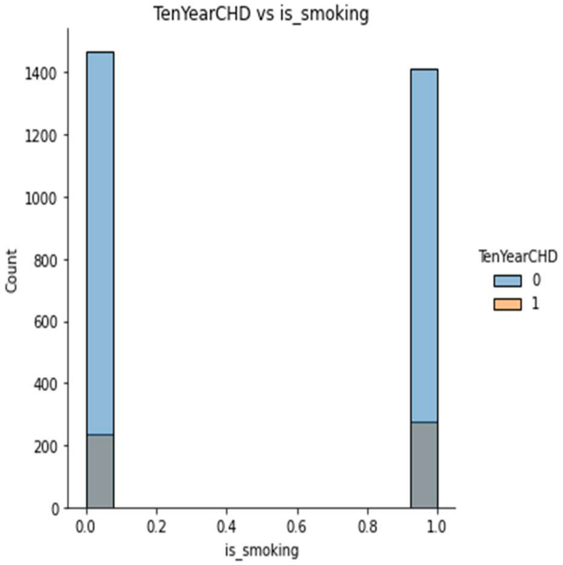


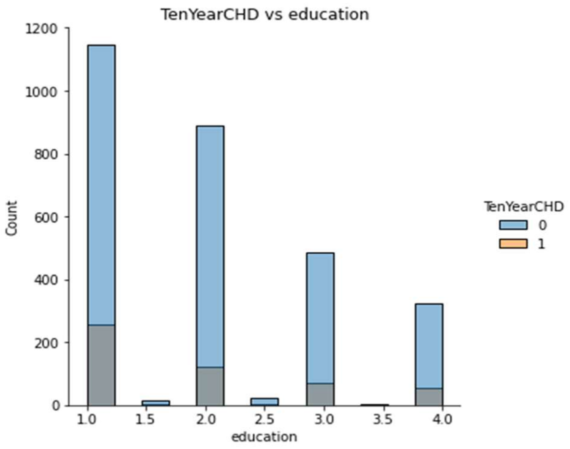


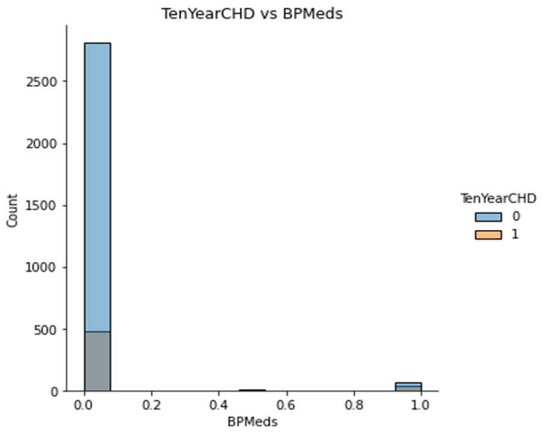
- **Glucose slightly shows a bit of right skewness. But since we are concerned about people with cardiovascular issues we consider the right skewed values as an important information so we left it as it is.**
- **cigsPerDay is not following Gaussian/Normal distribution and from the distplot as well as the kde plot we did not get any inference w.r.t target variable.**

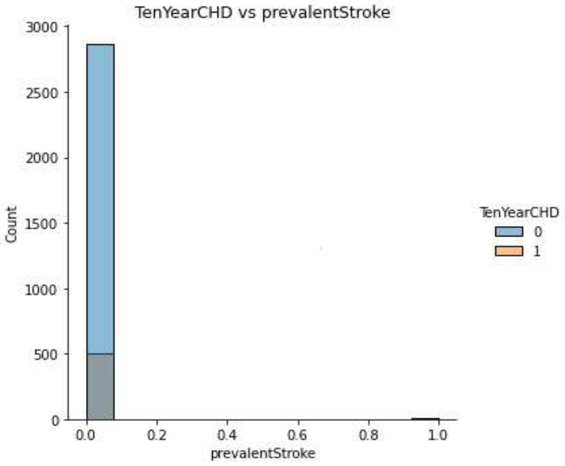
### **3)** EDA on discrete features

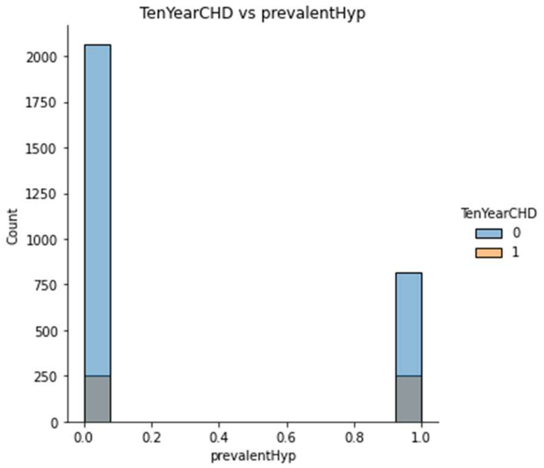


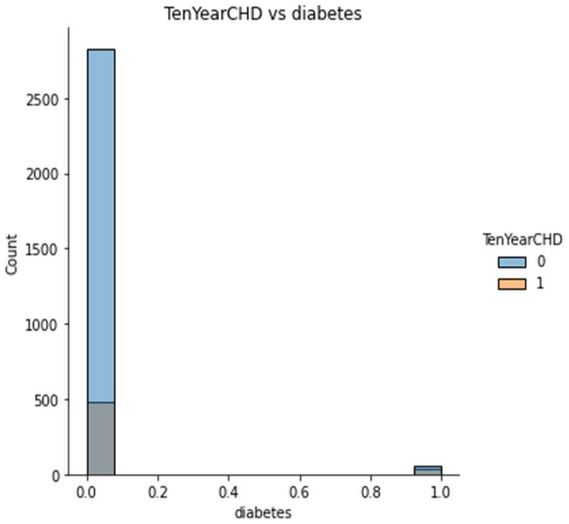








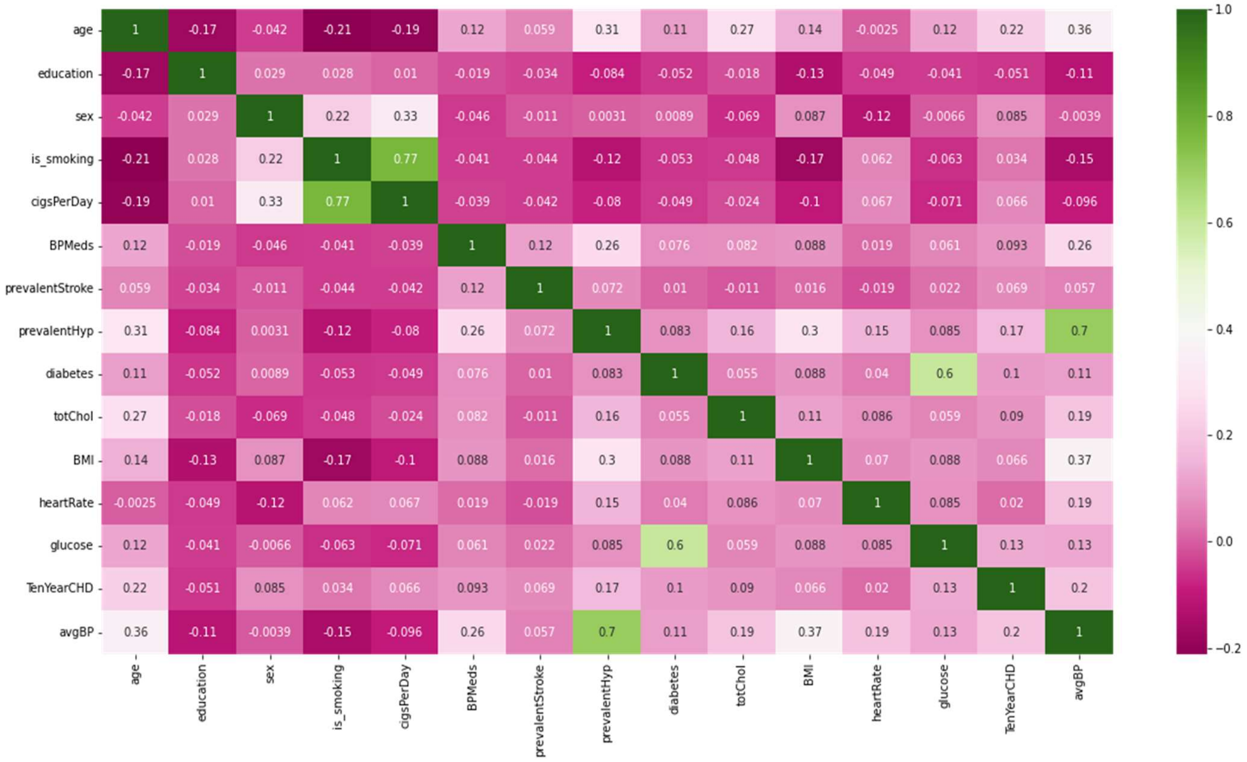




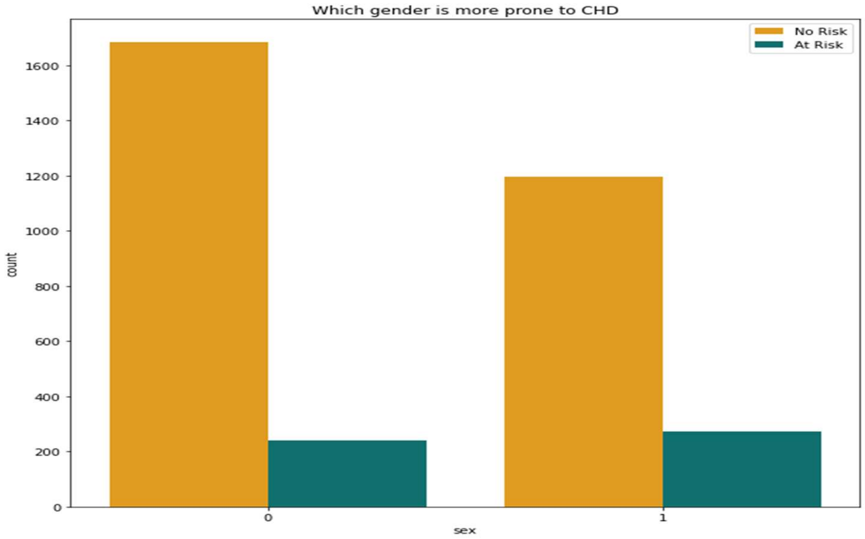
- **Education is important, as it is evident that if people are aware, they take care and precautions in order to avoid the risk of CHD.**

- **BPMeds, Prevalent stroke, diabetes has very low variance, thus we are unable to come up with any generalized conclusion about the co-relation between their history and prevalence of cardiovascular risk.**

**4)** Correlation heat map

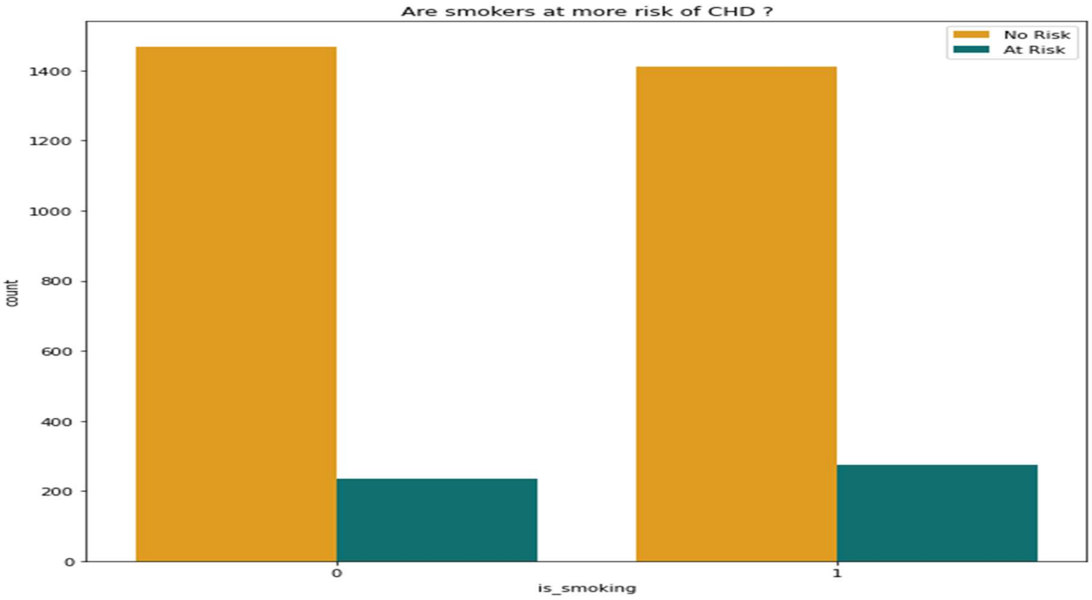


**5)** Which gender is prone to coronary heart disease?

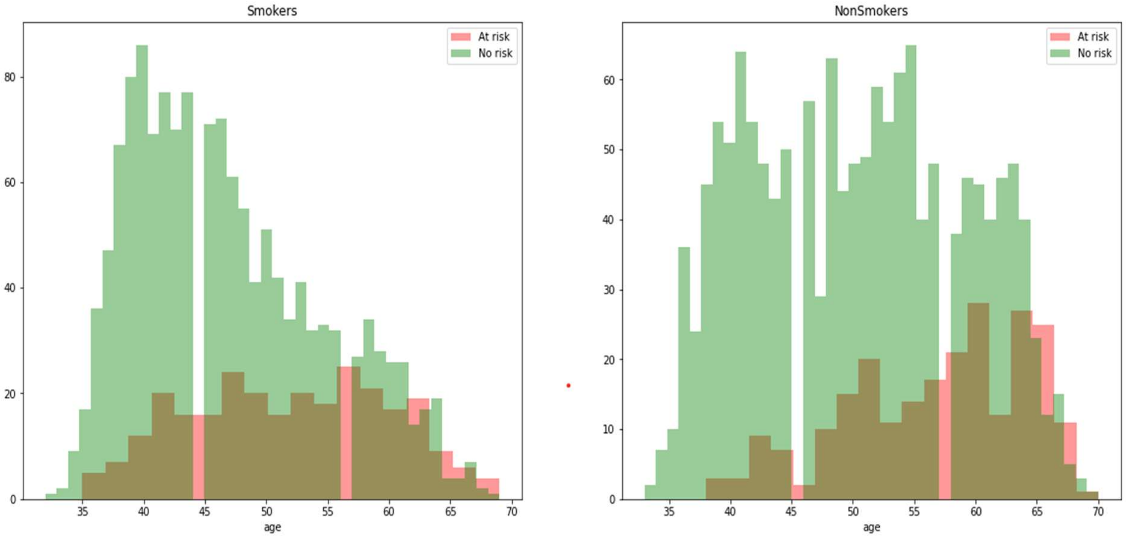


- **The number of males and females which are at risk of CHD is equal.**
- **As the number of males is high, the number of males who are not at risk is higher than female.**

### 6) Are smokers at more risk of coronary heart?

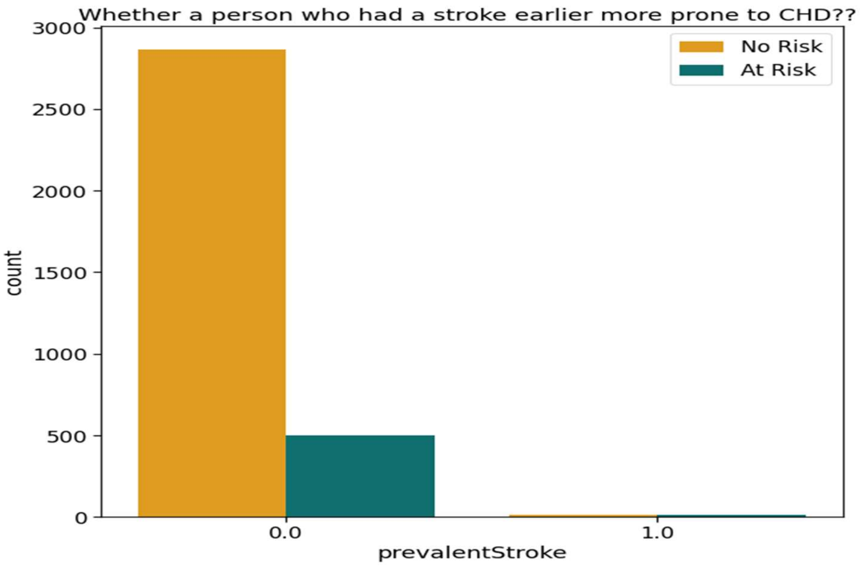


### 7) Age & Smoking v/s Risk



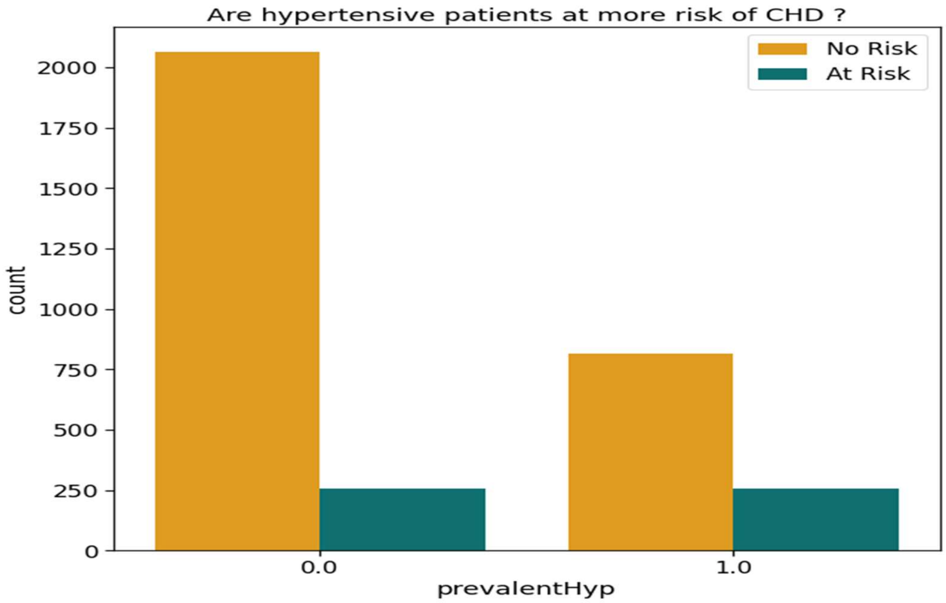
- **Age clearly plays an important role irrespective of smoking or not, which is clearly evident from the above two plots.**

**8)** Whether a person who had a stroke earlier more prone to CHD?



- **The person who previously had a heart stroke are more at risk to CHD than those who did not.**

**9)** Are hypertensive patients at more risk of coronary heart disease??



- **Out of all the people who are not Hypertensive, the number of people getting CHD is very less.**
- **People who are hypertensive has more chances of getting CHD.**

## Approach

We checked the Outliers and correlation matrix to overcome the noise in the dataset. Also, data was balanced using the SMOTE method and scaled by Standard Scaler transformation. As the Coronary Heart Diseases dataset defines the classification problem. We decided to train the models such as Logistic regression, K-nearest Neighbors, Decision Tree Classifier, Support Vector Machine, Random forest & Gradient boosting. Also, we used Hyperparameter Tuning for improvement in the model fitting to understand the better results of the model as well as the metrics.

## Data Modelling

After the data preparation is completed it is ready for the purpose of analysis. Only numerical valued features are taken into consideration. The data were combined and labelled as X and y as independent and dependent variables respectively.

## Splitting the dataset

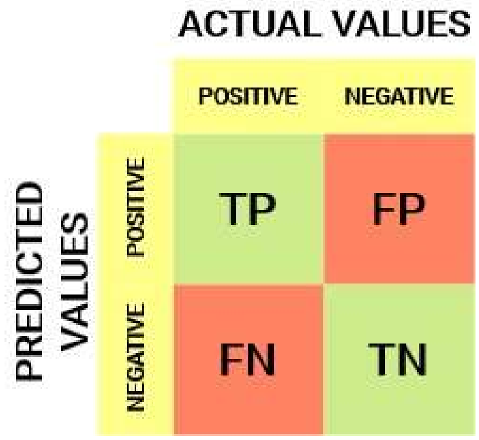
The train_test_split was imported from the sklearn. model selection. The data is now divided into 70% and 30% as train and test splits respectively. 70% of the data is taken for training the model and 30% is for a test and the random state was taken as 0.

## Scaling the data

We have used the **Standard Scaler method** to scale the dataset.

## Metrics used

- **Classification Report:** A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model.
- **Accuracy:** the proportion of total dataset instances that were correctly predicted out of the total instances
- **Recall (sensitivity):** the proportion of the predicted positive dataset instances out of the actual positive instances
  sensitivity=true positives/ (true positives+false negatives)
- **F1 score:** a composite harmonic mean (average of reciprocals) that combines both precision and recall. For this, we first measure the precision, the ability of the model to identify only the relevant dataset instances
  precision=true positives/ (true positives+false positives)
  The F1 score is estimated as
  F1=2×(precision×recall)/(precision+recall)
- **Confusion Matrix:** A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.



|  |  | ACTUAL VALUES |  |
|---|---|---|---|
|  |  | POSITIVE | NEGATIVE |
| PREDICTED VALUES | POSITIVE | TP | FP |
|  | NEGATIVE | FN | TN |

- True Positive (TP) -The predicted value matches the actual value. The actual value was positive and the model predicted a positive value
- True Negative (TN) -The predicted value matches the actual value. The actual value was negative and the model predicted a negative value
- False Positive (FP) – Type 1 error: -The predicted value was falsely predicted. The actual value was negative but the model predicted a positive value
- False Negative (FN) – Type 2 error: -The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value

## Model implementation

- **Logistic Regression** -A logistic regression is a type of statistical procedure. It is used to refer specifically to the problem in which the dependent variable is binary, that is the number of available categories is two, while the problem with more than two categories is referred to as multi logistic regression.

- **K-nearest Neighbors -** The K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets.

- **Support Vector Machine** –

Support Vector Machine (SVM) is a classification technique used for the classification of linear as well as non-linear data. SVM is the margin based classifier. It selects the maximum margin. This model is further used to perform classification of testing data.

- **Decision Tree Classifier -**Decision tree build classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

- **Random Forest -** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. [1][2] Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

- **Gradient Boosting -** The term gradient boosting consists of two sub-terms, gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term 'gradient' is related here. Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc.

Gradient Boosting Machine (GBM) builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

- ➔ **ROC Curve:** An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:
- ➔ True Positive Rate-True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows: **TPR=TP / (TP+FN)**
- ➔ False Positive Rate-False Positive Rate (FPR) is defined as follows:
  **FPR =FP / (FP+TN)**

## ➢ Challenges Faced

- Pre-processing the data was one of the challenges we faced which includes handling missing values and treating with outliers and filling the missing values with simpleimputer and KNNimputer
- As data is imbalanced, to balance the data in an appropriate way was a bit tricky. Under sampling the majority class Over sampling the minority class SMOTE Synthetic Minority Over Sampling Technique Reduces overfitting during oversampling Synthetic Sampling is used.
- Using the appropriate metrics for comparison of the implemented machine learning algorithms.

## ❖ Conclusion

We are finally at the conclusion of our project!

During the time of our analysis, we initially did EDA on all the features of our dataset. We first analysed the dataset find the null values treat with the outliers, fill the NaN values with KNNimputer and simpleimputer. Check the majority and minority set of the target variable then check if any of the NaN values belong to the minority class. We do feature engineering and one hot encoding for sex and is_smoking feature. Next, we analysed categorical variable and discrete variable. We do bivariate analysis and multivariate analysis. After the EDA we split the data into training and testing set and standardised the data with standard scaler. The given data is imbalanced data so we balanced the data with the help of SMOTE technique.

Next, we implemented 6 machine learning algorithms logistic regression, decision tree, random forest, gradient boosting, support vector machine, and k-nearest neigbours. We did hyperparameter tuning to improve our model performance.

The results of our evaluation are:

| | Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| 7 | XGBoost | 0.931134 | 0.957055 | 0.902778 | 0.929124 |
| 6 | Random Forest | 0.901042 | 0.920000 | 0.878472 | 0.898757 |
| 4 | KNN with hyperparameter tuning | 0.835069 | 0.934028 | 0.779710 | 0.849921 |
| 2 | SVM with hyperparameter tuning | 0.801505 | 0.834491 | 0.782845 | 0.807843 |
| 3 | K Nearest Neighbour | 0.774884 | 0.894676 | 0.721755 | 0.798966 |
| 0 | Decision Tree | 0.741898 | 0.744731 | 0.736111 | 0.740396 |
| 1 | Support Vector Machines | 0.726273 | 0.737269 | 0.721404 | 0.729250 |
| 5 | Logistic Regression | 0.662616 | 0.653935 | 0.665489 | 0.659661 |

- **Number of people belonging to middle age group are highest whereas number of people belonging to young age group are lowest**
- **Male and female both are equally prone to CHD**
- **Number of male smoker is higher than female smokers.**
- **People who suffered previously from a heart attack have high chances of getting CHD.**
- **XGBoost performed the best among all other models with highest accuracy and f1 score**
- **heartrate is the most important feature in predicting the CHD followed by totChol and glucose.**