

Cardiovascular Risk Prediction

Kirtesh Verma, Pravin Bejjo, and

Sahil Pardeshi

Data Science Trainees,

Almabetter

❖ INTRODUCTION

Coronary heart disease is the term that describes what happens when your heart's blood supply is blocked or interrupted by a build-up of fatty substances in the coronary arteries. Over time, the walls of your arteries can become furred up with fatty deposits. This process is known as atherosclerosis and the fatty deposits are called atheroma. The most important behavioural risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioural risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. A heart attack occurs when the heart muscle does not have enough blood or oxygen, such as when a blood clot develops from plaque in one of the coronary arteries. The formation of a blood clot is called coronary thrombosis. This clot, if it is big enough, can completely stop the supply of blood to the heart in that blood vessel.

❖ 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 dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. 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.

❖ DATASET DISCRIPTION

- **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)

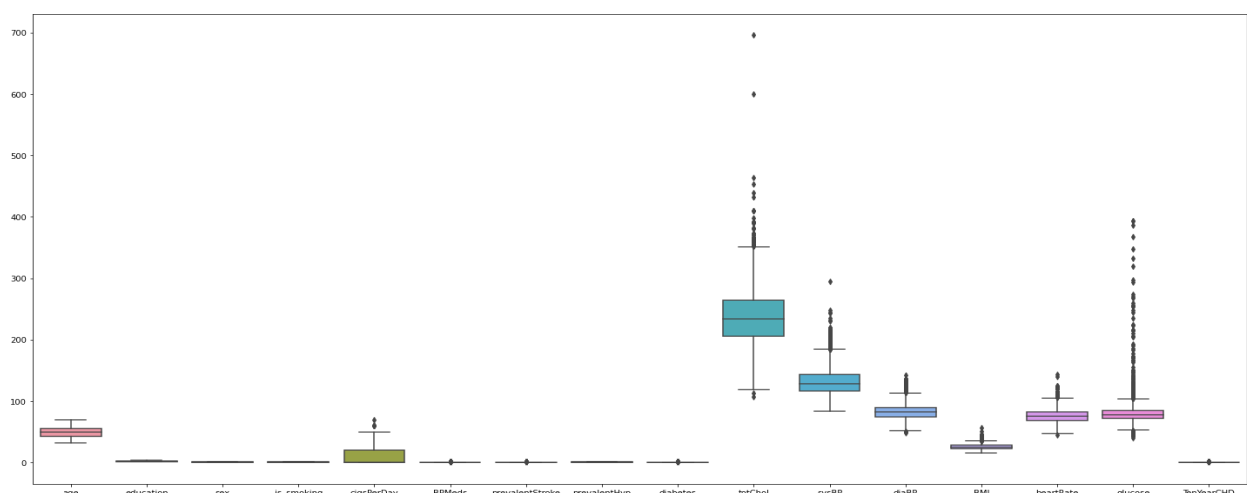
Predict variable (desired target)

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

❖ STEPS INVOLVED

I. EDA (Exploratory data analysis) and DATA PREPROCESSING

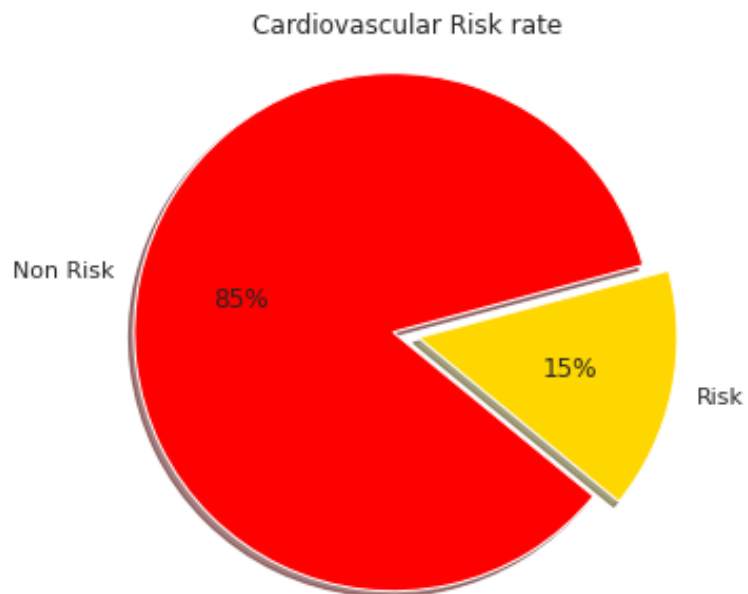
- 1) Exploring head and tail of the data to get insights on the given data.
- 2) Checking the Null values or missing values are present in the dataset or not.
- 3) Check the duplicate values.
- 4) Check if any of the NaN values belongs to the minority class.
- 5) Treating with outliers, some of the outliers are important for our dataset so, we cannot remove them. This would further affect our machine learning model.



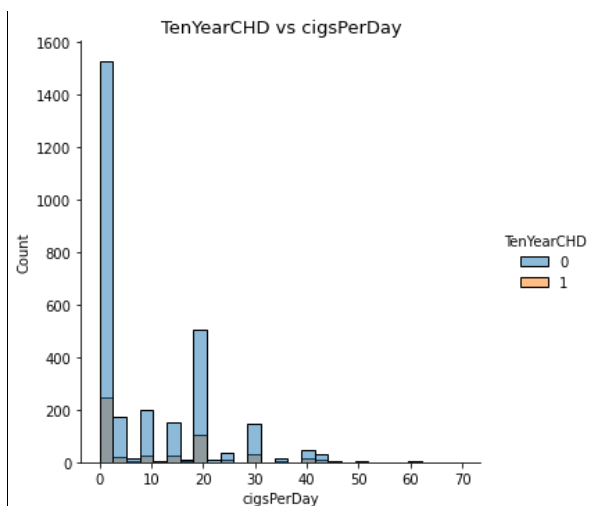
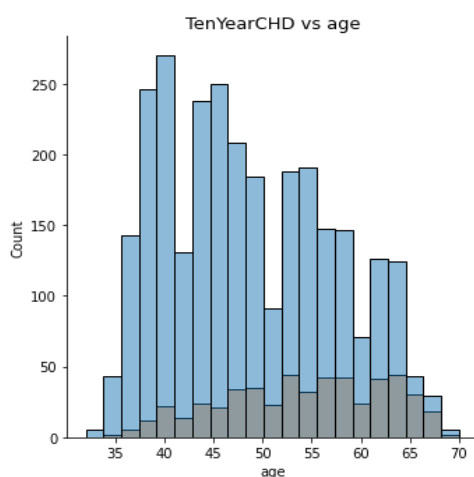
II. Drawing conclusion from the data

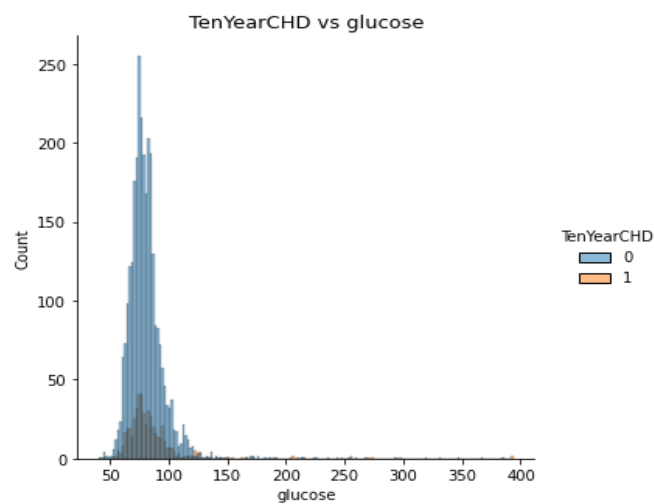
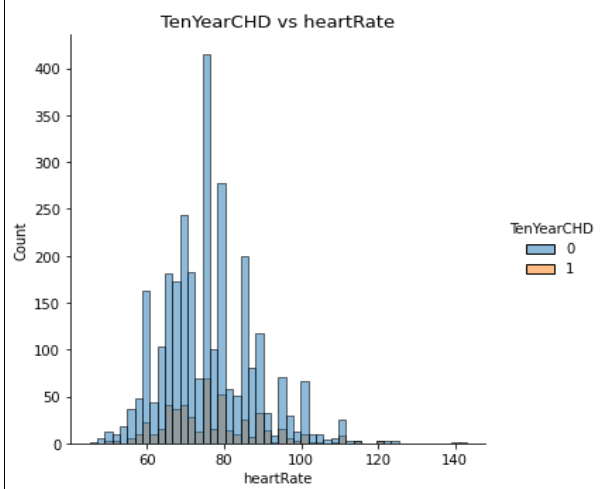
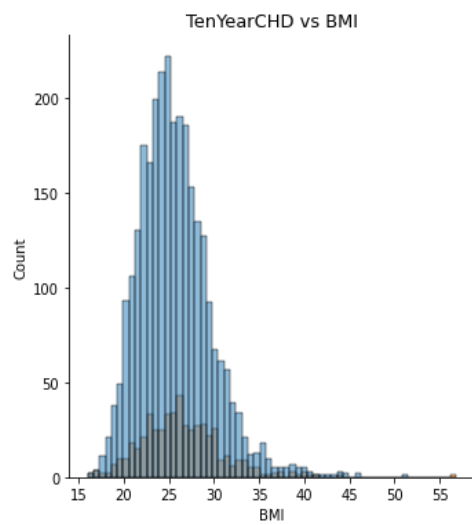
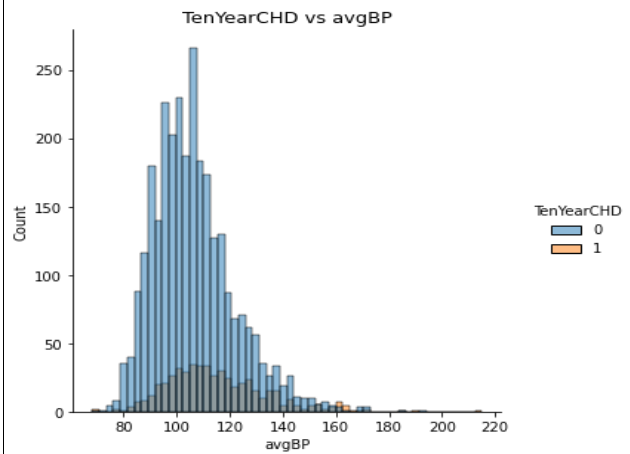
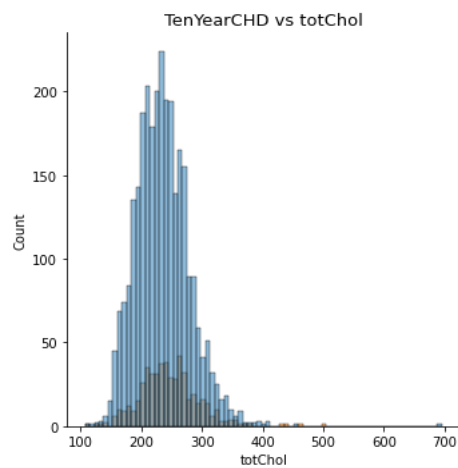
Plotting necessary graphs which provides relevant information on our data like:

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



2) EDA on continuous features.



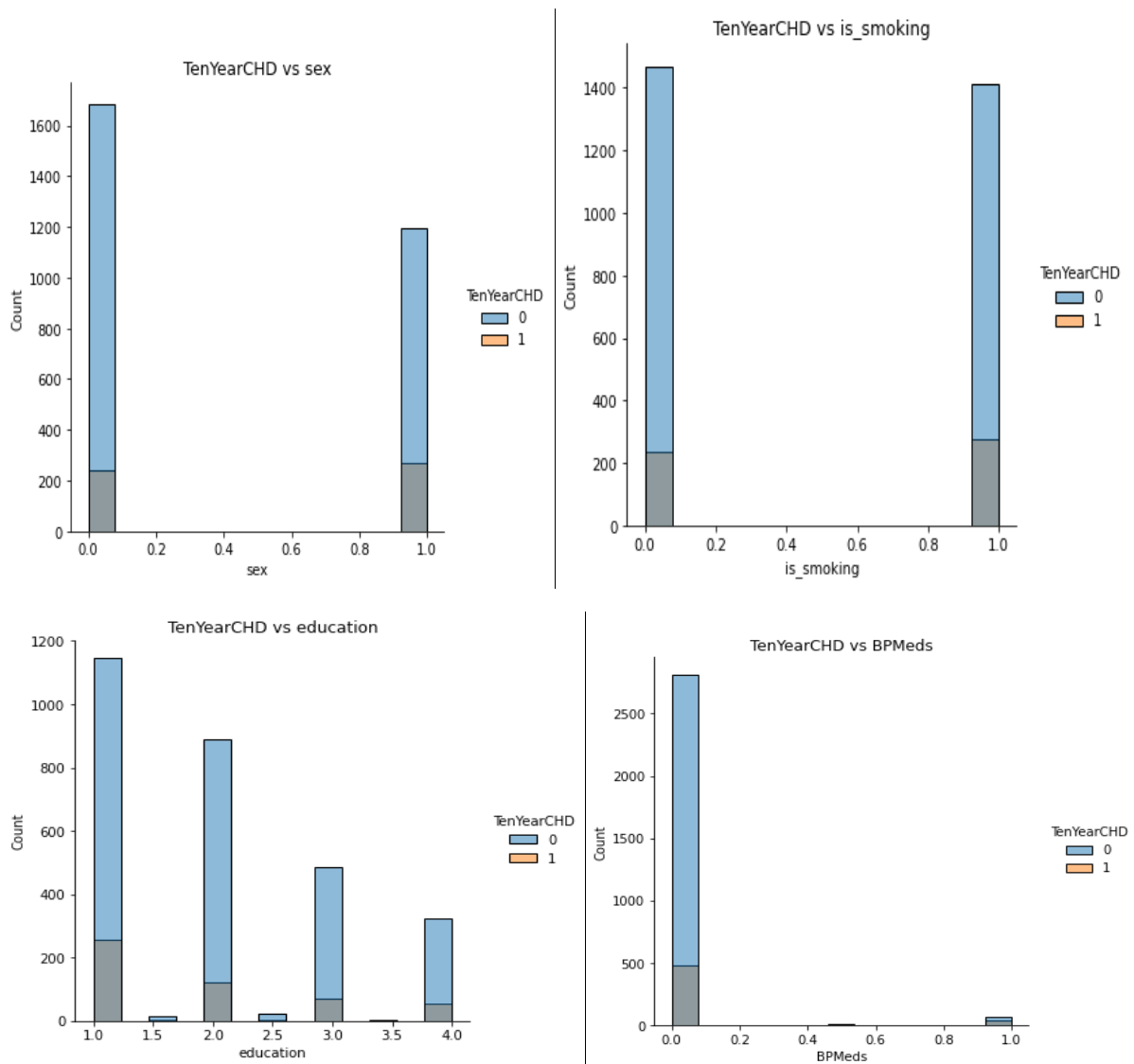


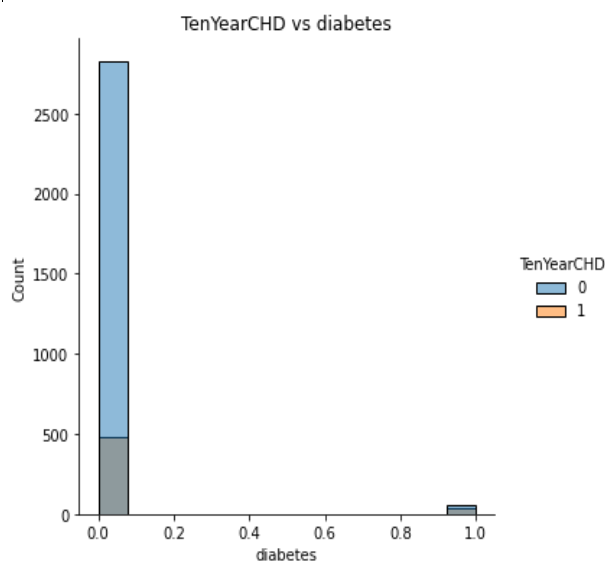
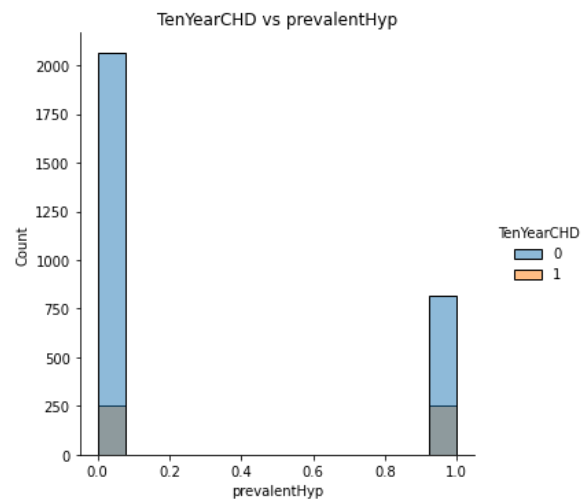
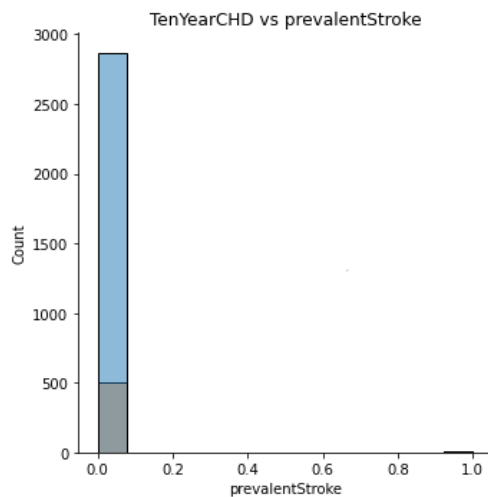
- Glucose slightly shows a bit of right skewness. But since we are concerned about people with cardiovascular issues, we

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.

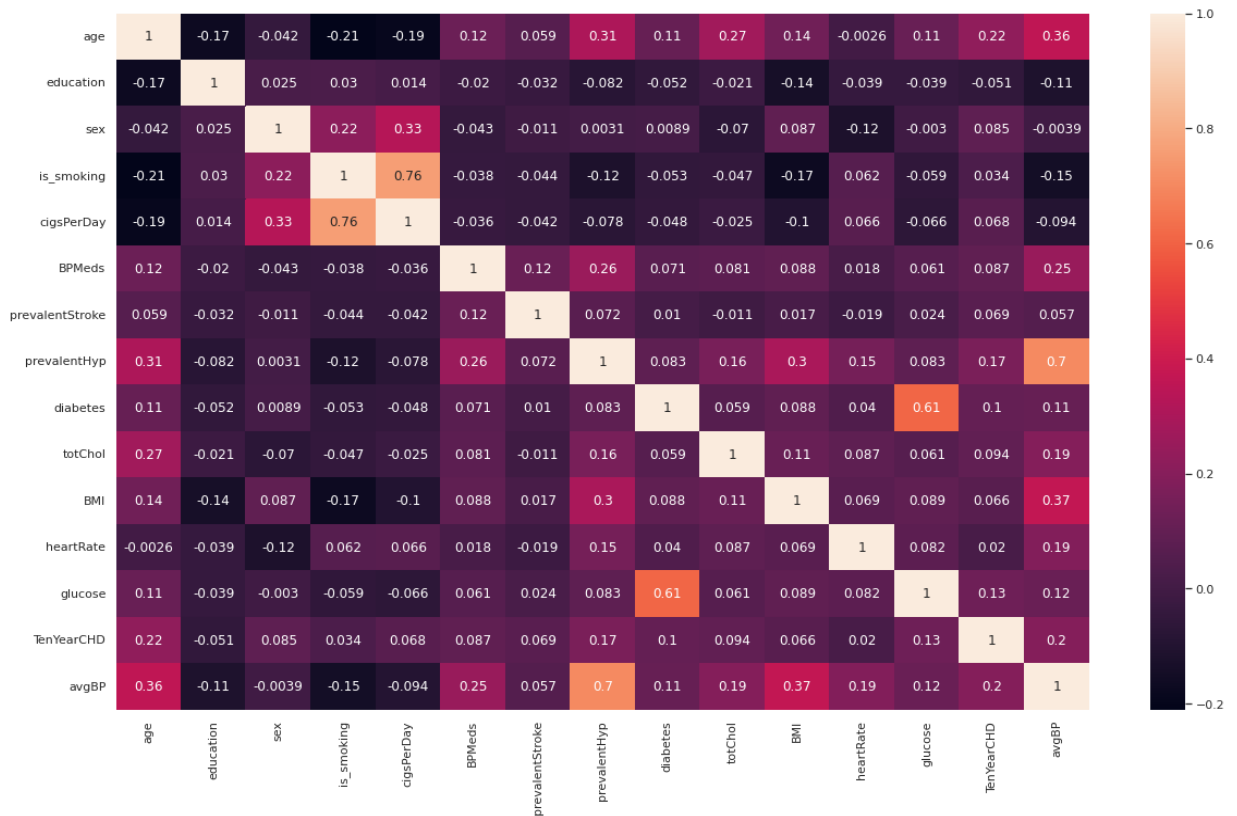
3) EDA on discrete features



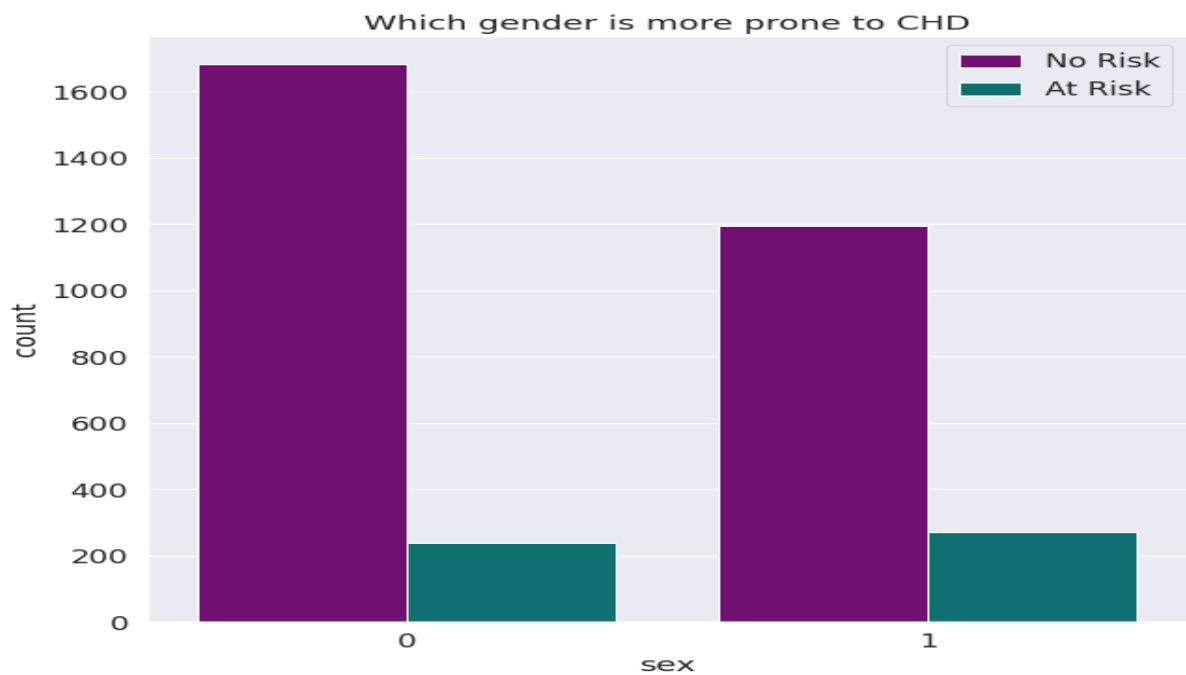


- Education is important, as it is evident that if people are aware, they take care and precautions in order to avoid the risk of CHD.
- BPMeds, Prevalent stroke, diabetes has very low variance, thus we are unable to come up with any generalized conclusion about the co-relation between their history and prevalence of cardiovascular risk.

1) Correlation heat map

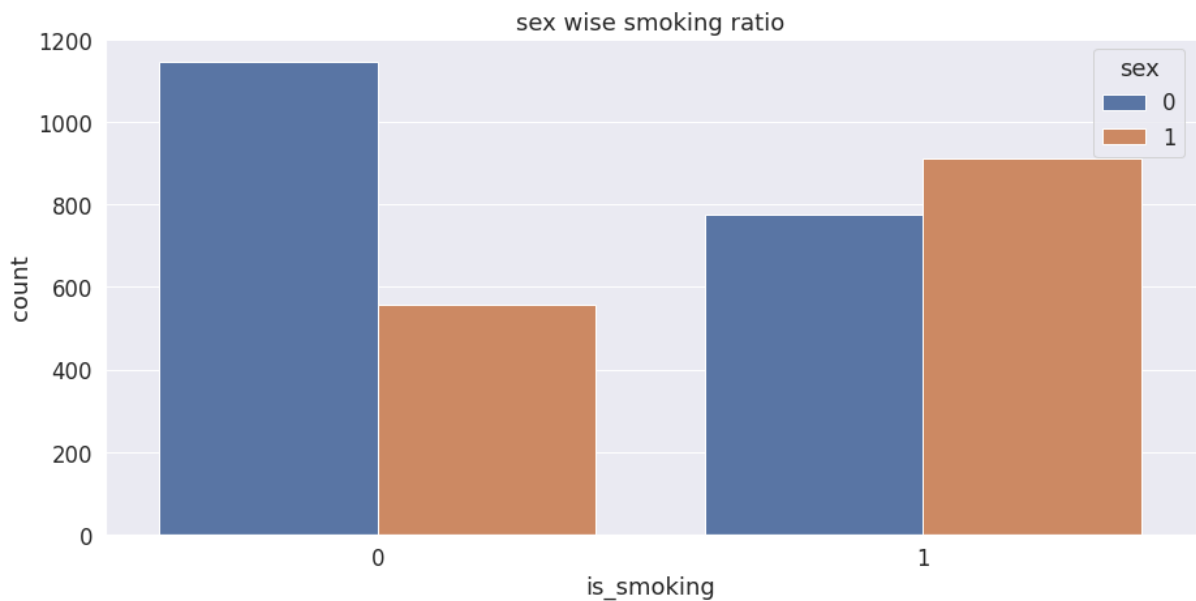


2) Which gender is prone to coronary heart disease?

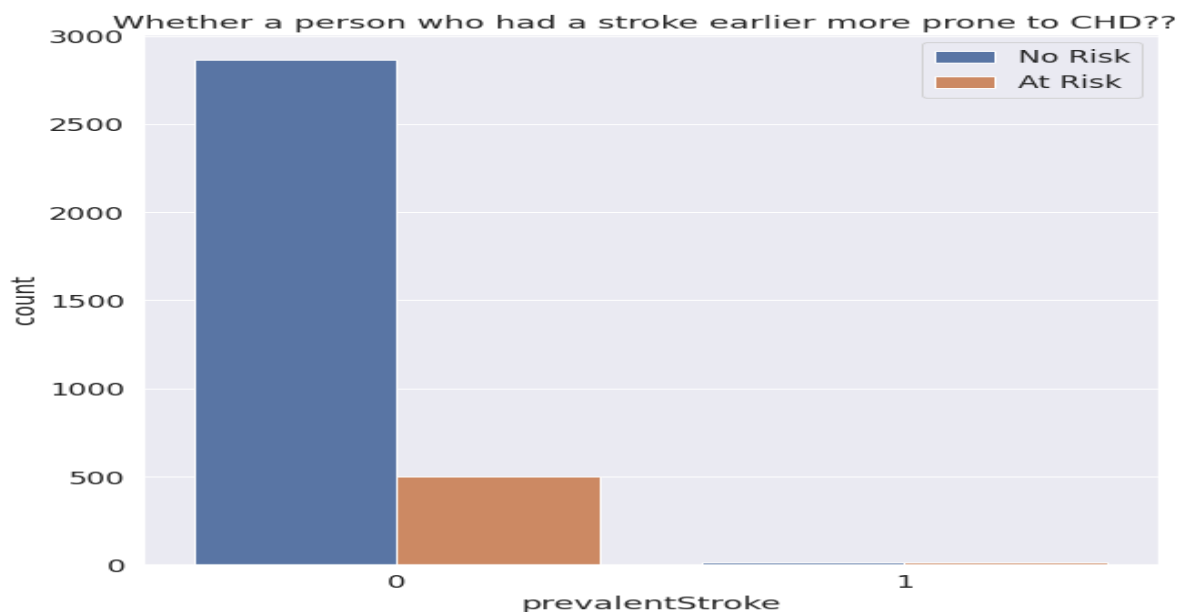


- The number of males and females which are at risk of CHD is equal.
- As the number of males is high, the number of males who are not at risk is higher than female.

3) Are smokers at more risk of coronary heart?



4) 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.
- 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 Neighbours, 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
$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$
- **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
$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

The F1 score is estimated as
$$F1 = \frac{2 \times (\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$$
- **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 Neighbour 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:

$$\text{TPR} = \text{TP} / (\text{TP} + \text{FN})$$

- False Positive Rate-False Positive Rate (FPR) is defined as follows:

$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN})$$

❖ 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.
- 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.

❖ MODELS PERFORMANCE

N	Model	Accuracy	Precision	Recall	F1 score
0	Decision Tree	0.759259	0.852201	0.627315	0.722667
1	Support Vector Machines	0.709491	0.725694	0.702915	0.714123
2	SVM with hyperparameter tuning	0.780093	0.814815	0.761905	0.787472
3	K Nearest Neighbour	0.748264	0.843750	0.708455	0.770206
4	KNN with hyperparameter tuning	0.805556	0.896991	0.758317	0.821845
5	Logistic Regression	0.656250	0.641204	0.661098	0.650999
6	Random Forest	0.879630	0.882284	0.876157	0.879210

❖ CONCLUSION

We performed EDA on the dataset and observed and understood the relationship between the variables. We clean the data and converted into machine readable format by making the necessary changes like label encoding. We fitted different machine learning models so to predict the heart disease.

- **We concluded the following points:**

- 1) Both the males as well as females are equally prone to getting CHD.
- 2) The number of males is slightly higher than that of female.
- 3) The number of males who smoke is more than females.
- 4) Person who smokes have higher chances of suffering from heart related problems.
- 5) People suffering from diabetes, BP, people who had a stroke in the past, people who are hypertensive have higher chances of CHD than those without it.
- 6) Age is an important factor in heart related problems. It increases with age. Hence middle and older age group people suffer from heart disease where as young age group people are least affected by it.
- 7) Random Forest turned out to be the best model in correctly classifying the people who have chances of CHD with an accuracy and F1 score of 88% and precision and recall of 88% and 87% respectively.