Cardiovascular Risk Prediction

By Mayank Ghai

Data Science Trainee

AlmaBetter, Bangalore

Abstract:

Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age.

The most important behavioral risk factors for heart disease and stroke are unhealthy diet, physical inactivity, tobacco use, and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These "intermediate risk factors" can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure, and other complications.

Cessation of tobacco use, reduction of salt in the diet, eating more fruit and vegetables, regular physical activity, and avoiding harmful use of alcohol have been shown to reduce the risk of cardiovascular disease. Health policies that create conducive environments for making healthy choices affordable and available are essential for motivating people to adopt and sustain healthy behaviors.

Keywords: EDA, Correlation, DecisionTree Random Forest, Classification, Prediction.

Problem Statement:

Visiting hospitals for regular check-ups it is almost always seen that they encourage people to get special check-ups to identify if they are at risk of heart diseases. Heart diseases have unfortunately become very common. It may be due to various reasons such as lifestyle, work pressure, lack of exercise, etc. In this project, we will be working on predicting the 10-year risk of **Coronary Heart Disease** (CHD). We are given a set of variables that impact heart diseases. These variables are related to demographic, past, and current medical history.

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). 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.

Introduction:

Cardiovascular disease (**CVD**) is a class of diseases that involve the heart or blood vessels. CVD includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack).

Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, and peripheral artery disease, thromboembolic disease, and venous thrombosis.

The underlying mechanisms vary depending on

the disease. It is estimated that dietary risk factors are associated with 53% of CVD deaths. Coronary artery disease, stroke, and peripheral artery disease
involve atherosclerosis. his may be caused by high blood pressure, smoking, diabetes mellitus, lack of exercise, obesity, high blood cholesterol, poor diet,
excessive alcohol consumption, he may be caused by high blood
pressure, smoking, diabetes mellitus, lack of exercise, obesity, high blood cholesterol, poor diet, excessive alcohol consumption,
It is estimated that up to 90% of CVD may be preventable. Prevention of CVD involves improving risk factors through healthy eating, exercise, avoidance of tobacco smoke ,and limiting alcohol intake. Treating risk factors, such as high blood pressure, blood lipids and diabetes is also beneficial. Treating people who have strep throat with antibiotics can decrease the risk of rheumatic heart disease. The use of aspirin in people, who are otherwise healthy, is of unclear benefit.

Approach:

The approach followed here is to first check the sanctity of the data and then understand the features involved. The events followed were in our approach:

- **Understanding the business problem and the datasets**
- **Data cleaning and preprocessing-** Finding null values and imputing themwith appropriate values.
  Converting categorical values into appropriate data types and merging the datasets provided to get a final dataset to work upon.
- **Exploratory data analysis-** of categorical and continuous variablesagainst our target variable.
- **Data manipulation-** Feature selection and engineering, feature scaling, outlier detection and treatment, and encoding of categorical features.
- **Feature Selection –** Using the chi-square test score we are going to select those features that are going to affect the predictions and hence only these features will be there in the dataset that we are going to use in the modeling To know which model will be best for the prediction we have split the data into Test-Train data and with its help we checked the parameter that is used to choose the best model among the given predictive models.

Understanding the Data:

The first step involved understanding the data and getting answers to some basic questions like; Whatis the data about? How many rows or observationsare there in it? How many features are there in it? What are the data types? Are there any missing values? And anything that could be relevant and useful to our investigation. Let's just understand the dataset first and the terms involved before proceeding further.
Our dataset consists of 3390 rows and 17 columns, which have been explained below.

Let's define the features involved:

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)

Behavioural:
• **is_smoking:** whether or not the patient is a current s moker ("YES" or "NO")
• **Cigs Per Day:** the number of cigarettes that the pers on smoked on average in one day
.(can be considered continuous as one can have any n umber of cigarettes, even half a cigarette.)

Medical (history):

• **BP Meds:** whether or not the patient was on blo od pressure medication (Nominal)
• **Prevalent Stroke:** whether or not the patient ha d previously had a stroke (Nominal)
• **Prevalent Hyp:** whether or not the patient was hypertensive (Nominal)
• **Diabetes:** whether or not the patient had diabete s (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 thought discr ete, are considered continuous because of a 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

Data Cleaning and Preprocessing:

Handling missing values is an important skill in the data analysis process. If there are very few missing values compared to the size of the dataset, we may choose to drop rows that have

missing values. Otherwise, it is better to replacethem with appropriate values.

It is necessary to check and handle these values before feeding them to the models, to obtain good insights into what the data is trying to say and to make great characterization and predictionswhich will in turn help improve the business's growth.

The historical records dataset had no null values .

```
[11] data.isnull().sum()

age                 0
education          87
sex                 0
is_smoking          0
cigs_per_day       22
bp_meds            44
prevalent_stroke    0
prevalent_hyp       0
diabetes            0
total_cholesterol  38
systolic_bp         0
diastolic_bp        0
bmi                14
heart_rate          1
glucose           304
ten_year_chd        0
dtype: int64
```

The dataset had some nulls values in the following columns:

- Education
- Cigs per day
- Bp med
- Total cholesterol
- Bmi
- Heart rate
- Glucose

As we start with a limited set of rows of 3390, we must try to fill the null values strategically instead of dropping them.

Exploratory Data Analysis:

Exploratory data analysis is a crucial part of data analysis. It involves exploring and analyzing the dataset given to find out patterns, trends, and conclusions to make better decisions related to the data, often using statistical graphics and otherdata visualization tools to summarize the results. The visualization tools involved in the investigation are python libraries- matplotlib andseaborn.

The goal here is to explore the relationships of different variables with 'Sales' to see what factors might be contributing to the high and low sales numbers.

Approach:

There are two kinds of features in the dataset: Categorical and Non-Categorical Variables.

Categorical- A categorical variable is a variablethat can take on one of a limited, and usually fixed, number of possible values putting a particular category to the observation.

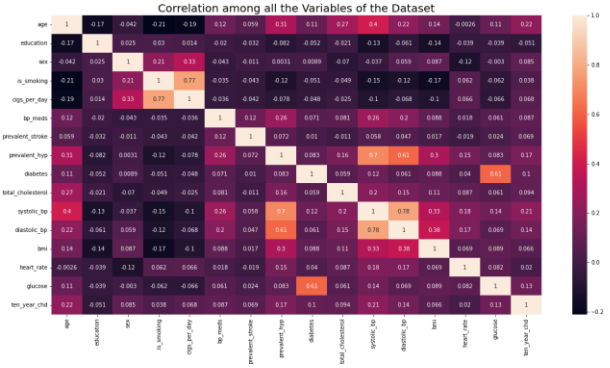
Non-Categorical- A noncategorical or continuous variable is a variable whose value isobtained by measuring, i.e., one which can take on an uncountable set of values.

Both of them are analyzed separately. Categorical data is usually analyzed through count plots and bar plots by the target variable and that is what is done here too. On the other hand, Numeric or Continuous variables were analyzed through distribution plots, box plots, and scatterplots to get useful insights.

Correlation:

Correlation is a statistical term used to measure the degree to which two variables move about each other. A perfect positive correlation means that the correlation coefficientis exactly 1. This implies that as one variable move, either up or down, the other moves in the same direction. A perfect negative correlationmeans that two variables move in opposite directions, while a zero correlation implies no linear relationship at all.

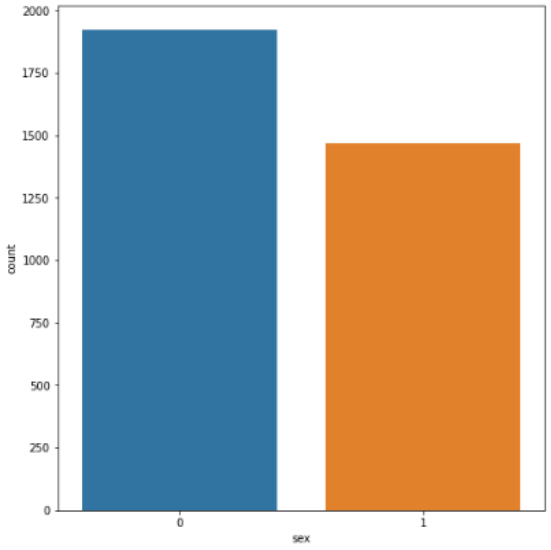
By checking the correlation the factors affecting sales can be figured out.

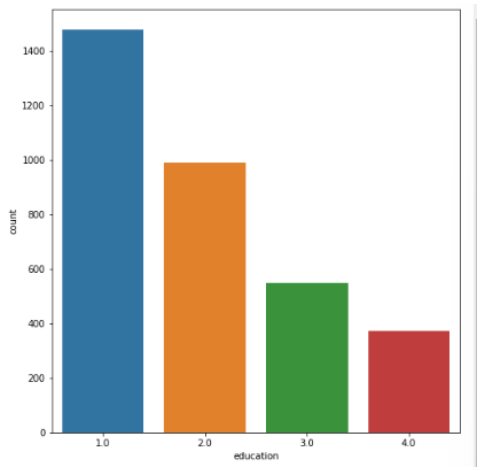


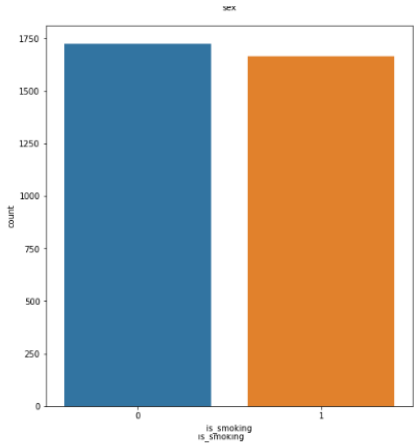
Above is the correlation heatmap for all the continuous variables in the dataset.

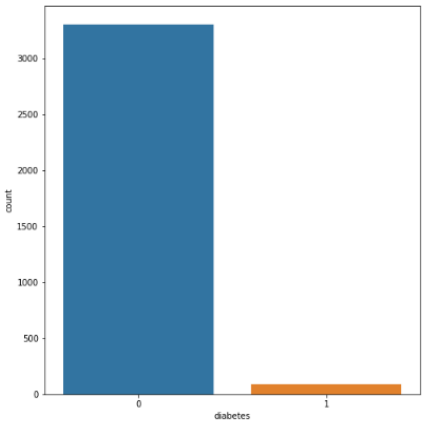
- The variables systolic BP and diastolic BP are highly correlated.
- is_smoking is highly correlated with cigs_per_day High correlation
- prevalent_hyp is highly correlated with systolic bp and diastolic bp
- diabetes is highly correlated with glucose High correlation
- Compared to all the independent data, the correlation coefficient between education and target variable Ten Year CHD is very low and actually negative.

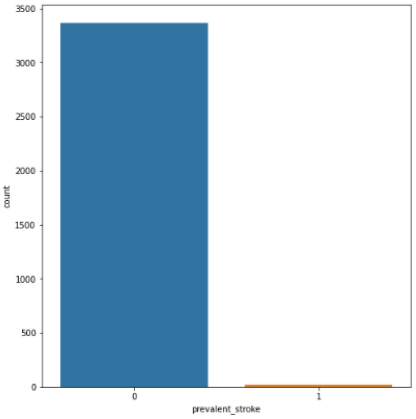
Univariate Analysis

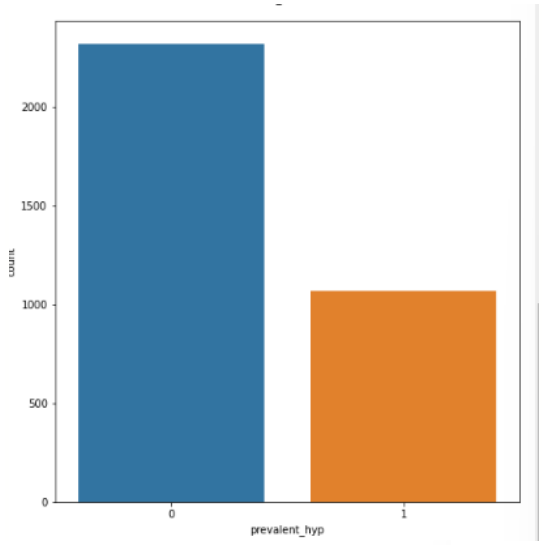


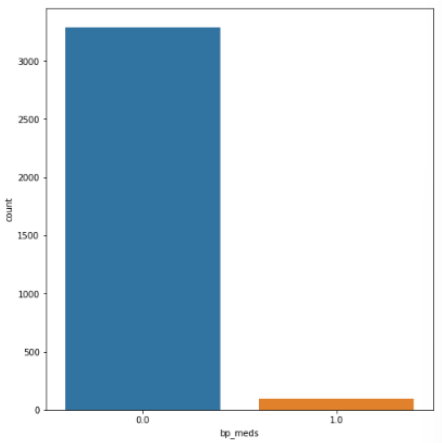








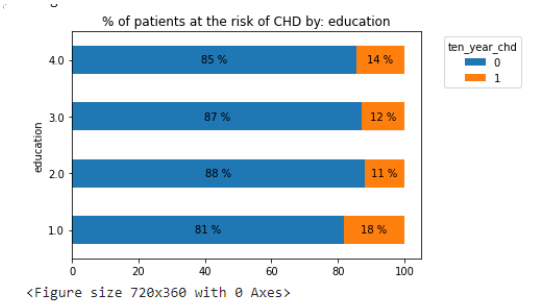




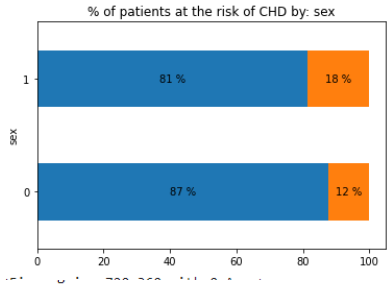
The graphs above show, categorical columns in the form of counterplots, we can understand that:

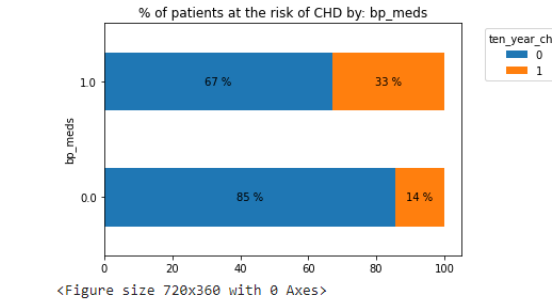
- Females are more in proportion to men by a small margin.
- There are more is_smoking than smokers by a small margin, both are around 1600 each.
- Around 1400 people have an education level of 1, and almost 400 people have an education level of 4. The levels are not defined.
- More than 3000 people are not on BP medication
- Only a small number of people have suffered a stroke previously.
- Around 1000 people were hypertensive.
- A large number (> 3000) of the people do not have diabetes.

Analyzing the relationship between the dependent variable and categorical independent variables

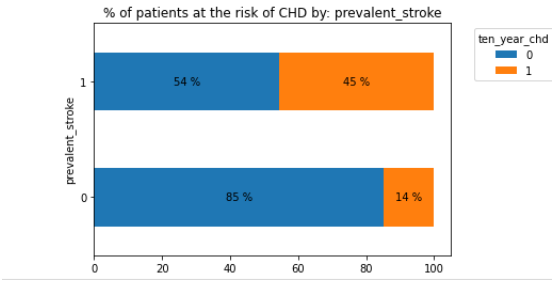


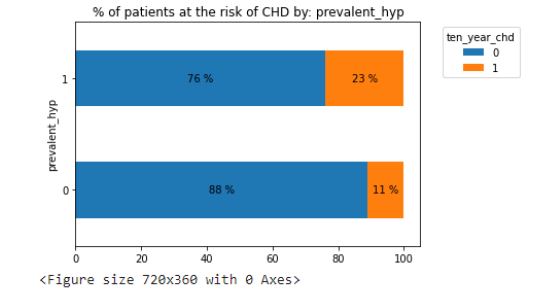
<Figure size 720x360 with 0 Axes>



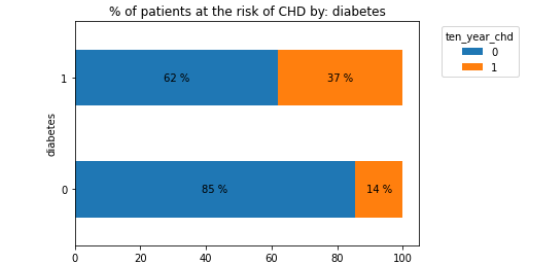


<Figure size 720x360 with 0 Axes>





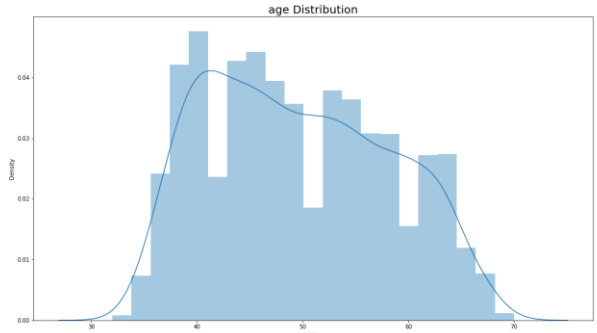
<Figure size 720x360 with 0 Axes>

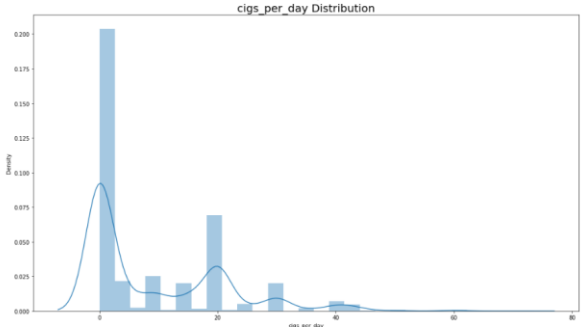


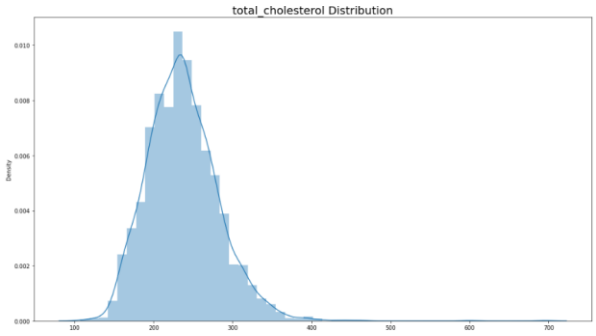
Observation:

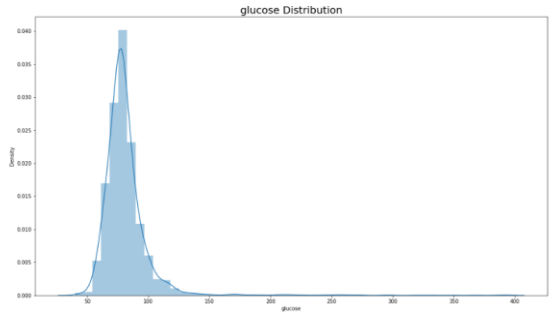
- 18%, 11%, 12%, 14% of the patients belonging to the education level 1, 2, 3, 4 respectively were eventually diagnosed with CHD.
- Male patients have significantly higher risk of CHD (18%) than female patients (12%)
- Patients who smoke have significantly higher risk of CHD (16%) than patients who don't smoke (13%)
- Patients who take BP medicines have significantly higher risk of CHD (33%) than other patients (14%)
- Patients who had experienced a stroke in their life have significantly higher risk of CHD (45%) than other patients (14%)
- Hypertensive patients have significantly higher risk of CHD (23%) than other patients (11%)
- Diabetic patients have significantly higher risk of CHD (37%) than other patients (14%)

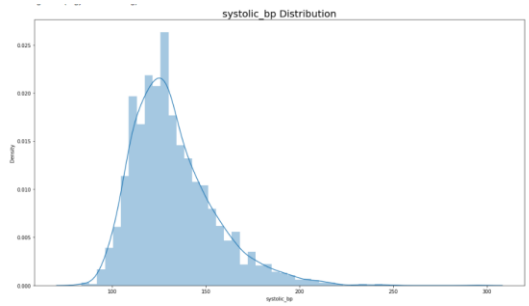
Numerical Features

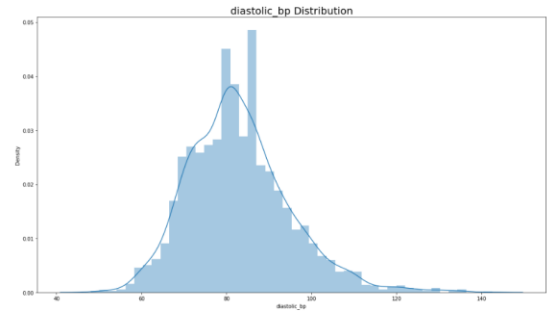


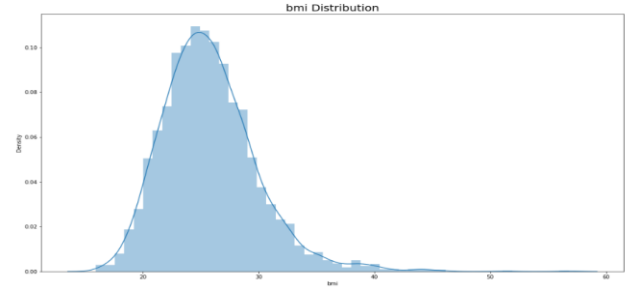


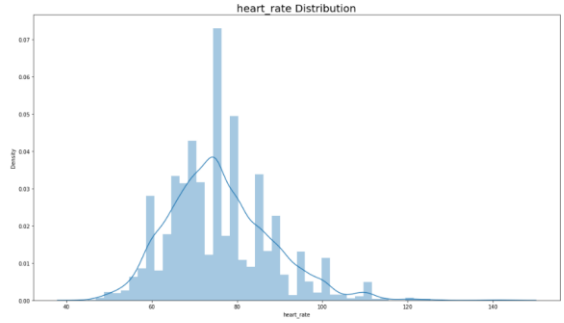












As seen from the above histograms, we can understand that :

- Total Cholesterol, systolic bp, diastolic BP and BMI have an uniform distribution, and the rest are unevenly distributed
- cigs_per_Day has a highly uneven distribution with the most data present in 0
- cigs_Per_Day and systolic bp show quite a bit and slight right skewness respectively.
- Age ranges from 35 years to 70 years and is almost normally distributed, with most people belonging to an age group of 40.
- Cigarettes smoked per day on an average are mainly 0, but 20 cigarettes a day are also prevalent.
- cigs_per_day has 1633 (50.3%) zeros
- Cholesterol ranges from 100 to 700, with most belonging to 150 to 350.
- Systolic BP ranges mainly from 100 to 200.
- Diastolic BP ranges mainly from 60 to 120.
- BMI ranges mainly from 16 to 40.
- Heart rate ranges from 40 to 110 and most occurrences are around 75.
- Glucose ranges mainly from 50 to 125, rest seem like outliers with extreme numbers but cannot be ignored as these numbers can cause a risk of heart disease.

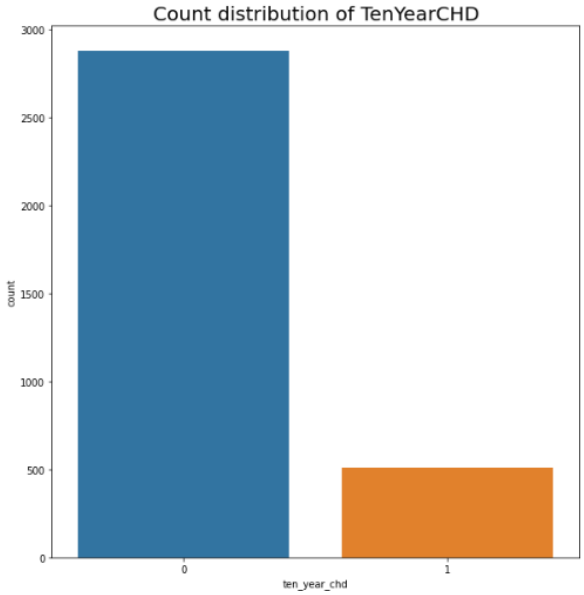
For the same numerical features:

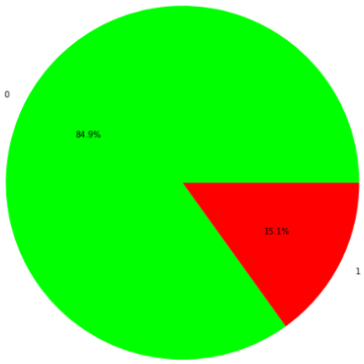
Cigs_Per_Day has uneven distribution although most of the data is concentrated on 0

The majority portions of the following columns lie in the range:

- Total_Cholesterol: 150 to 300
- Systolic_BP: 100 to 150
- Diastolic_BP: 60 to 100
- BMI: 20 to 30
- Heart_Rate: 50 to 100
- glucose: 50 to 150

Target Variable



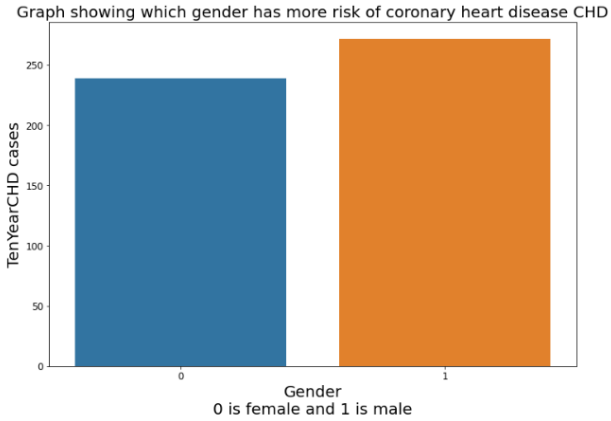


The distribution is highly imbalanced. As in, the number of negative cases outweighs the number of positive cases. This would lead to a class imbalance problem while fitting our models. Therefore, this problem needs to be addressed and taken care of.

Bivariate Analysis

Relationship between education and cigs_Per_Day

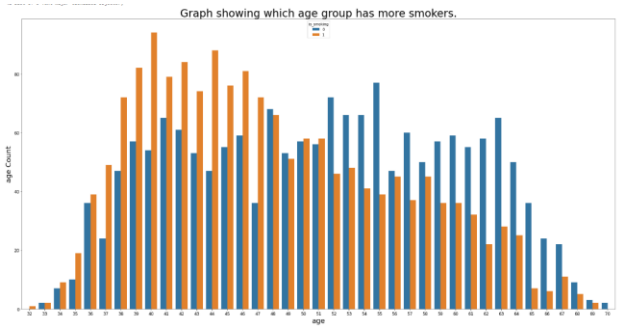
Which gender has more risk of coronary heart disease CHD



Observation:

According to this dataset, males have shown a slightly higher risk of coronary heart disease Ten Year CHD

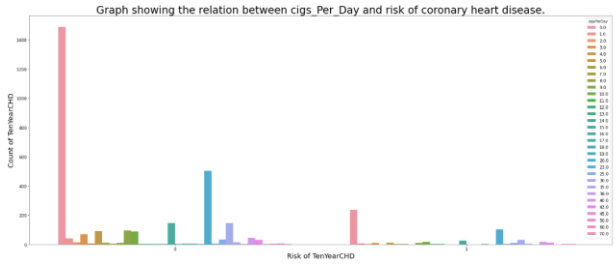
Graph showing which age group has more smokers



Observation:

- Mid-age groups ranging from the age of 38 - 46 have more smokers.
- No smokers observed below the age of 32
- maximum age for an smoker is 69

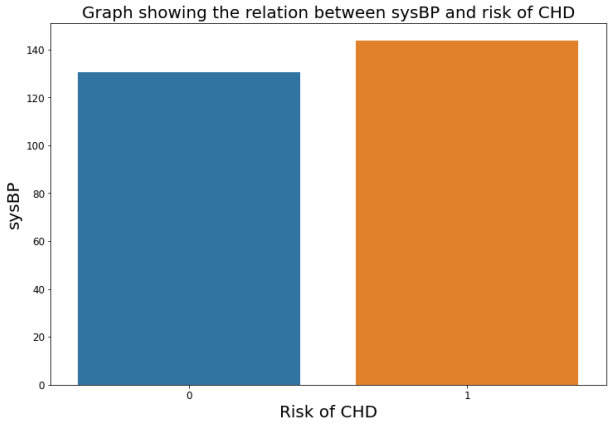
Relation between cigs_per_Day and risk of coronary heart disease



Observation:

- Low cigs_Per_Day comes with a lower risk of CHD.
- Those who don't smoke, i.e., with a cigs_Per_Day of 0.0 has a really low risk of contracting the disease
- Although that is the case, low cigs_Per_Day doesn't guarantee a much lower risk of CHD

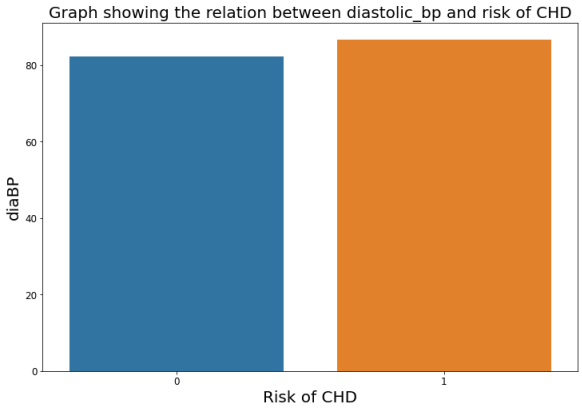
Relation between systolic BP and risk of CHD.



Observation:

- Minor relation of higher risk of Ten Year CHD found with higher systolic BP
- Majority of people with systolic_BP ranging from 72 - 130 have a lower chance of contracting the disease.

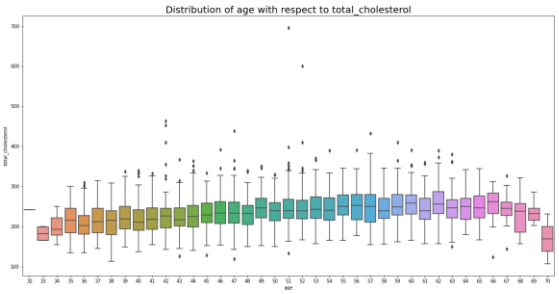
Relation between diastolic_BP and risk of CHD



Observation:

- Minor relation found between higher risk of Ten Year CHD with higher diastolic BP similar to the previous one
- Majority of people with diastolic BP ranging up to 80.0 have a a lower chance of contracting the disease.

Relation between age and total_Cholesterol

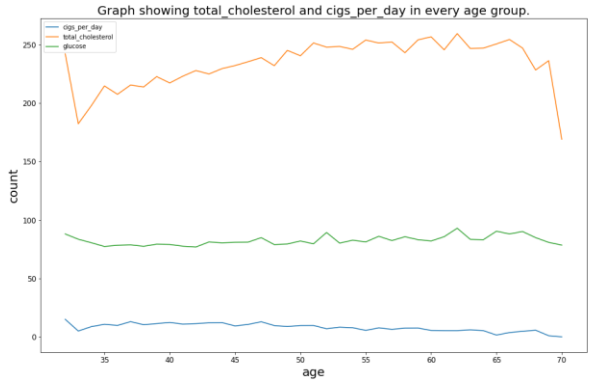


Observation:

The boxplots are shifted in an upward manner suggesting that aged people have more cholesterol (bad cholesterol in general).

Multivariate Analysis

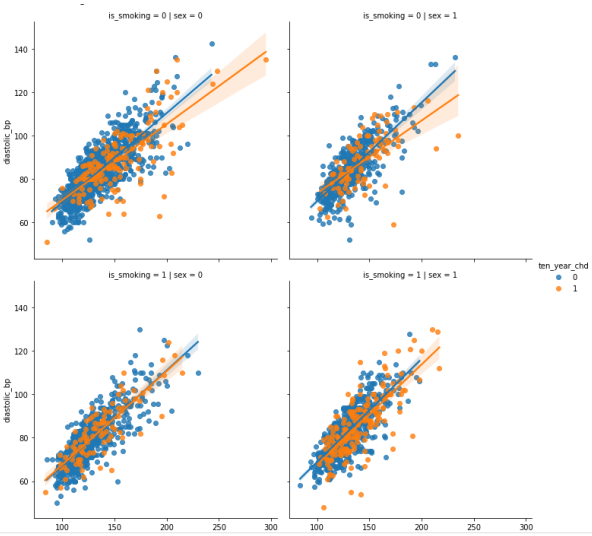
Relationship between age and cigs_Per_Day, total_Cholesterol, glucose.



Observation:

- There is a minor relation between total Cholesterol and glucose.
- Total Cholesterol has a steep, linear, and inverse graph for lower ranges of age
- Cigs_Per_Day has a fairly parallel relationship with age

Distribution of systolic_BP vs diastolic_ BP concerning is_smoking and sex attributes



Observation:

The above graph plots the relationship between systolic blood pressure and diastolic blood pressure for patients based on their gender and whether they are current smokers or not and plots the best fit line

Data Manipulation:

Data manipulation involves manipulating and changing our dataset before feeding it to various regression machine learning models. This involves keeping important features, outlier treatment, feature scaling, and creating dummyvariables if necessary.

Outlier Detection:

In statistics, an outlier is a data point that differs significantly from other observations. Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.

Z-score is a statistical measure that tells you how far a data point is from the rest of the dataset. In amore technical term, Z-score tells how many standard deviations away a given observation is from the mean.

z = (x-mean)/standard deviation

All other outliers are required but the one in total cholesterol and systolic bp columns of our dataset should be removed.

A new feature Pulse pressure will be derived from the systolic bp and diastolic Bp which will give the pulse pressure of a person respectively and the rest of the two features will be dropped from the dataset to reduce multicollinearity.

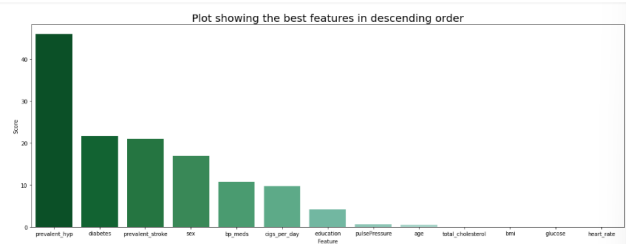
To reduce the skewness of all the continuous variables available in the dataset we have applied a suitable transformation to remove the skewness of that columns. We aim to reduce the magnitude of skew in the variables since they will impact our prediction model. The goal here is to bring the skew for all continuous variables within the range [-0.5, 0.5] i.e. almost normally distributed.

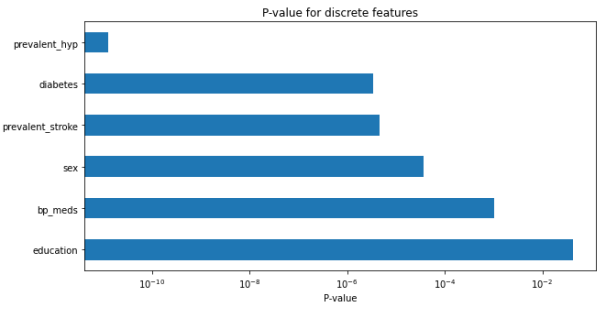
We can log transform all continuous variables to reduce skew and bring it within the permissible range except for the glucose variable because this feature is extremely skewed to the right, hence we can inverse transform it to reduce the skew.

Attribute	Original skew	Transformation Used	Skew After Transformation
Age	0.103528	Log10	-0.015053
Cigarettes Per Day	0.476592	Log10	0.290367
Total Cholesterol	0.407456	Log10	0.012596
BMI	0.661379	Log10	0.369286
Heart Rate	0.410993	Log10	0.165867
Glucose	3.892094	Inverse	-0.323851
Pulse Pressure	1.412381	Log10	0.3541744

Feature Selection:

Feature selection is a crucial step in any machine learning model, as properly selected features will help us to make and choose a model that will give better predictions. We will use the chi-square test and plot P values for discrete features.to select the best features among the available ones and make a dataset from these features and pass down the various models to check the performance of all the models.





Observation:

- Since the prevalent hypertension column (prevalent _hyp) has the small-valuable, we can say that it is the most im portant feature (among the categorical ind ependent variables) which determines the outcome of the dependent variable.
- The education feature has the highest p-value, which indicates that it is the least important feature (among categorical independent variables).

Modeling:

Feature Scaling:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is done to prevent the biased nature of machine learning algorithms towards features with greater values and scale. The two techniquesare:
Normalization: is a scaling technique in which values are shifted and rescaled so that they end upranging between 0 and 1. It is also known as
Min-Max scaling. [0,1]

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization: is another scaling technique where the values are centered around the mean with a unit standard deviation. This means thatthe mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
[-1,1]

$$X' = \frac{X - \mu}{\sigma}$$

Here we have applied a max min scaler to normalize the data before giving the values to the model to make valuable predictions.

We divide the dataset into training and test sub-datasets for predictive modeling.

Factors affecting choosing the model:

Determining which algorithm to use depends on many factors like the problem statement and thekind of output you want, the type and size of the data, the available computational time, the number of features, and observations in the data, to name a few.

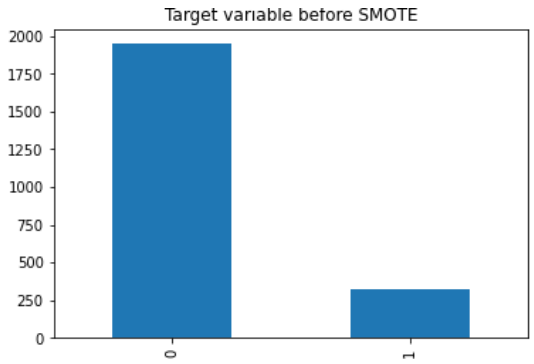
Train-Test Split:

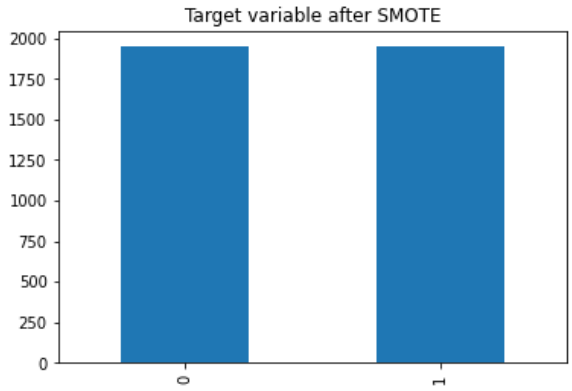
In machine learning, train/test split splits the data randomly, as there's no dependence from one observation to the other. That's not the casewith time series data. Here, it's important to usevalues at the rear of the dataset for testing and everything else for training.

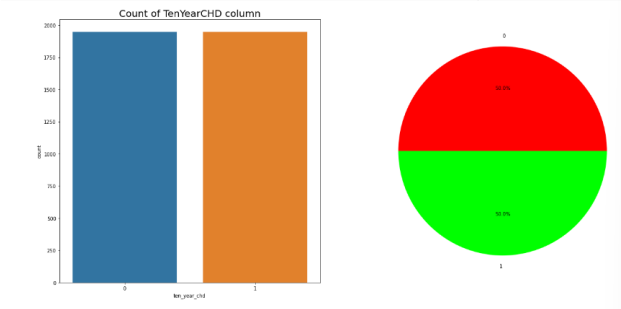
The **30% of Data** was kept as a testing set and the rest of the historical data was used in thetraining set.

Sampling:

- Since we are dealing with unbalanced data, ie, only ~15% of the patients were diagnosed with coronary heart disease, we oversample the training dataset using SMOTE (Synthetic Minority Oversampling Technique).
- This ensures that the model has trained equally on all kinds of results, and it is not biased to one particular result.







We have successfully oversampled the minority class using SMOTE. Now the model we build will be able to learn from both classes without any bias.

Machine Learning Model –

Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.
Building a model by learning the patterns of historical data with some relationship between data to make a data-driven prediction.

Types of Machine Learning
• Supervised Learning
• Unsupervised Learning
• Reinforcement Learning

Supervised learning
In a supervised learning model, the algorithm learns on a labelled dataset, to generate reasonably predictions for the response to new data.

(Forecasting outcome of new data)
• Regression
• Classification

Machine Learning models can be understood as a program that has been trained to find patterns within new data and make predictions. These models are represented as a mathematical function that takes requests in the form of input data, makes predictions on input data, and then provides an output in response. First, these models are trained over a set of data, and then they are provided an algorithm to reason over data, extract the pattern from feed data and learn from those data. Once these models get trained, they can be used to predict the unseen dataset.

Parameters that we need to check to evaluate the machine learning model that we are going to use. These parameters tell us about the model accuracy on training data, but it is also important to get a genuine and approximate result on unseen data otherwise Model is of no use.

Evaluation Metrics:

There are several model evaluation metricsto choose from but since our dataset was highly imbalanced, it is critical to understand which metric should be evaluated to understand the model performance.

- **Accuracy**- Accuracy simply measures how oftenthe classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions. Accuracy is useful when the target class is well balanced but is not a good choice for the unbalanced classes, because if the model poorly predicts every observation of the majority class, we are going to get pretty high accuracy.
- **Confusion Matrix** - It is a performance measurement criteria for the machine learning classification problems where we get a table witha combination of predicted and actual values.
- **Precision** - Precision for a label is defined as the number of true positives divided by the number ofpredicted positives.
- **Recall** - Recall for a label is defined as the number of true positives divided by the total number of actual positives. Recall explains how many of the actual positive cases we were able to predict correctly with our model.
- **F1 Score** - It's the harmonic mean of Precision and Recall. It is maximum when Precision is equal to Recall.
- **AUC ROC** - The Area Under the Curve (AUC)is the measure of the ability of a classifier to distinguish between classes. When AUC is 0.5,the classifier is not able to distinguish between the classes and when it's closer to 1, the better it becomes at distinguishing them.

Evaluation metrics that we are using :

- Since the data we are dealing with is unbalanced, accuracy may not be the best evaluation metric to evaluate the model performance.
- Also, since we are dealing with data related to healthcare, False Negatives are of higher concern than False Positive
- In other words, it doesn't matter whether we raise a false alarm but the actual positive cases should not go undetected
- Considering these points in mind, it is decided that we use **Recall** as the model evaluation metric.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

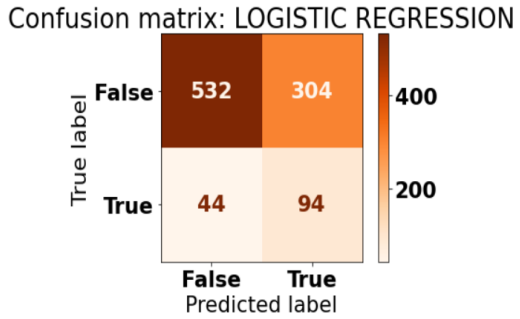
So among all the above metrics, which metric should be prioritized in comparing the performance of our various models? That's the major question here as we have a multiclass classification problem, where the problem statement just asks us to track and classify between ignored, read, and acknowledged classes, we can not decide here what we want toprioritize in terms of classification, we just wantto correctly classify and characterize accordingly.

When we have a high-class imbalance, we'll choose the F1 score because a high F1 score considers both precision and recall. To get a high F1, both false positives and false negativesmust be low. The F1 score depends on how highly imbalanced our dataset is!

Logistic Regression:

Logistic Regression is a classification algorithm that predicts the probability of an outcome that can have only two values. Multinomial logistic regression is an extension of logistic regression that adds native support for multi-class classification problems. Instead, the multinomial logistic regression The algorithm is a model that involves changing the loss function to cross-entropy loss and predicting probability distribution to a multinomial probability distribution to natively support **multi-class classification problems.**

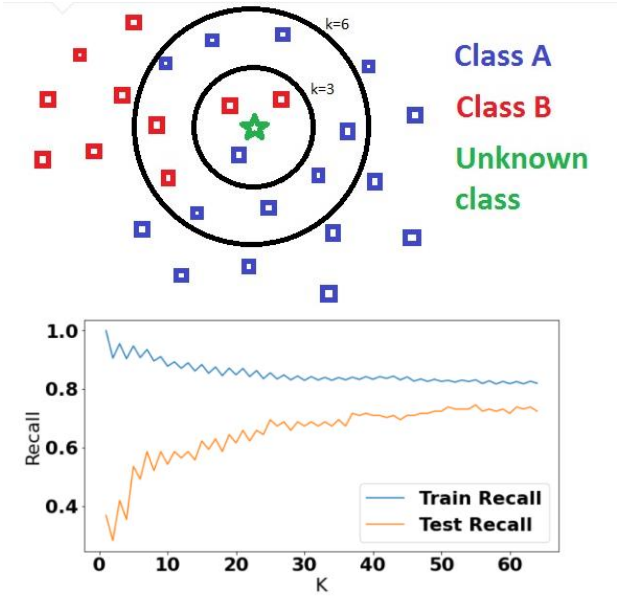
	precision	recall	f1-score	support
0	0.92	0.64	0.75	836
1	0.24	0.68	0.35	138
accuracy			0.64	974
macro avg	0.58	0.66	0.55	974
weighted avg	0.83	0.64	0.70	974



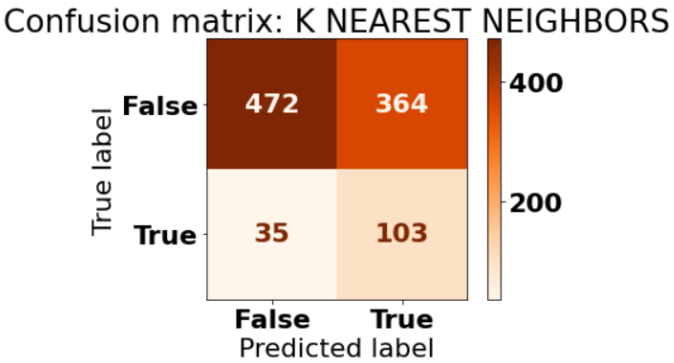
Model_Name	Precision_train	Precision_test	accuracy_score_train	accuracy_score_test	roc_auc_score_train	roc_auc_score_test	recall_train	recall_test	f1_s
LOGISTIC REGRESSION	0.680653	0.236181	0.669235	0.64271	0.669235	0.658762	0.695942	0.681159	

K-Nearest Neighbors Regression:

KNN regression is a lazy learning and non-parametric method that uses 'feature similarity' to predict the values of new data points. It calculates the distance between test data and each row of training data with the help of any of the methods namely: Euclidean, Manhattan, or Hamming distance. Assign the points which are nearest to it and approximate the result with the mean or mode values of its neighbors for regression or calcification respectively. The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation.
The method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables.
.



	precision	recall	f1-score	support
0	0.93	0.56	0.70	836
1	0.22	0.75	0.34	138
accuracy			0.59	974
macro avg	0.58	0.66	0.52	974
weighted avg	0.83	0.59	0.65	974



Model_Name	Precision_train	Precision_test	accuracy_score_train	accuracy_score_test	roc_auc_score_train	roc_auc_score_test	recall_train	recall_test	f1_score_train	f1_score_test
K NEAREST NEIGHBORS	0.671085	0.220557	0.712121	0.590349	0.712121	0.655485	0.832049	0.746377	0.742949	0.340496

Naive Bayes:

Naive Bayes classifiers are a collection of

classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal

contribution to the outcome.

About our dataset, this concept can be understood as:

- We assume that no pair of features are dependent. For example, the temperature being 'Hot' has nothing to do with the humidity or the outlook being 'Rainy' does not affect the winds. Hence, the features are assumed to be **independent**.
- Secondly, each feature is given the same weight(or importance). For example, knowing the only temperature and humidity alone can't predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing **equally** to the outcome.

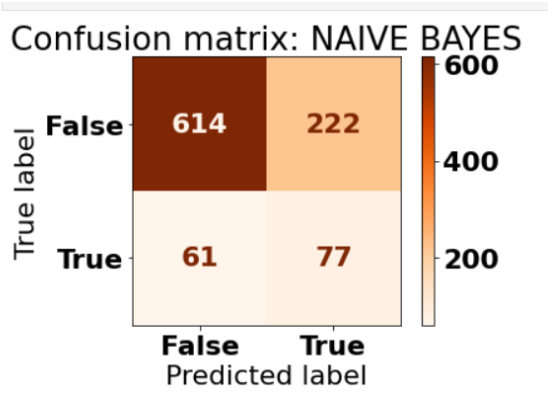
Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation

where A and B are events and P(B) ≠ 0.

- Basically, we are trying to find the probability of event A, that event B is true. Event B is also termed as **evidence**.
- P(A) is the **priority** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

	precision	recall	f1-score	support
0	0.91	0.73	0.81	836
1	0.26	0.56	0.35	138
accuracy			0.71	974
macro avg	0.58	0.65	0.58	974
weighted avg	0.82	0.71	0.75	974



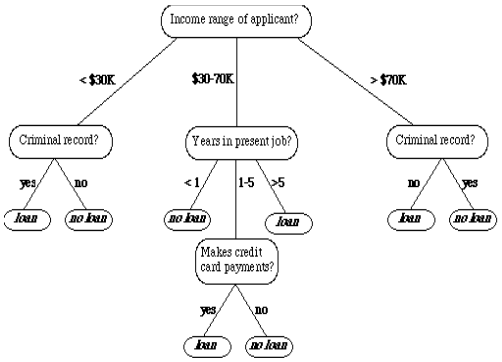
Model_Name	Precision_train	Precision_test	accuracy_score_train	accuracy_score_test	roc_auc_score_train	roc_auc_score_test	recall_train	recall_test	f1_score_train	f1_score_test
NAIVE BAYES	0.67118	0.257525	0.636364	0.709446	0.636364	0.64621	0.524669	0.557971	0.595197	0.352403

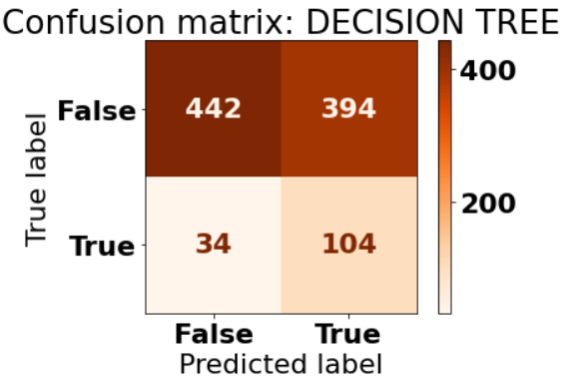
Decision Tree:

A decision tree is a type of supervised learning algorithm that can be used in classification as well as regressor problems. The input to a decision tree can be both continuous as well as categorical. The decision tree works on an if-then statement. A decision tree tries to solve a problem by using tree representation (Node and Leaf)

- Assumptions while creating a decision tree: Initially all the training set is considered as a root
- Feature values are preferred to be categorical, if continuous then they are discretized
- Records are distributed recursively based on attribute values
- Which attributes are considered to be in the root node or internal node is done by using a statistical approach.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches





The results show that a simple decision tree is performing pretty well on the validation set but it has completely overfitted the train set with a sample test score of 0.75. It's better to have a much more generalized model for future data points.

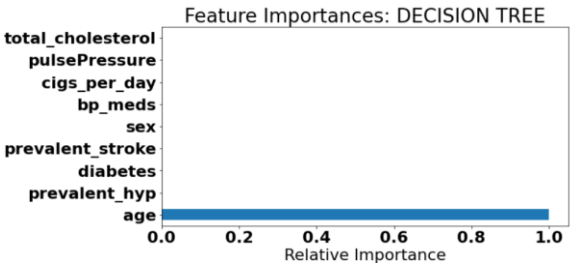
Businesses prefer the model to be interpretable to understand the patterns and strategize accordingly unlike any scientific.

The facility where the results matter much more thaninterpretability.

If interpretability is important then sticking with tree-based algorithms when most of the features are categorical; is beneficial and using tuned Hyperparameters to grow the tree deep enough without overfitting.

Model_Name	Precision_train	Precision_test	accuracy_score_train	accuracy_score_test	roc_auc_score_train	roc_auc_score_test	recall_train	recall_test	f1_score_train	f1_score_test
DECISION TREE	0.628156	0.208835	0.653826	0.560575	0.653826	0.641166	0.75398	0.753623	0.685341	0.327044

	precision	recall	f1-score	support
0	0.93	0.53	0.67	836
1	0.21	0.75	0.33	138
accuracy			0.56	974
macro avg	0.57	0.64	0.50	974
weighted avg	0.83	0.56	0.62	974

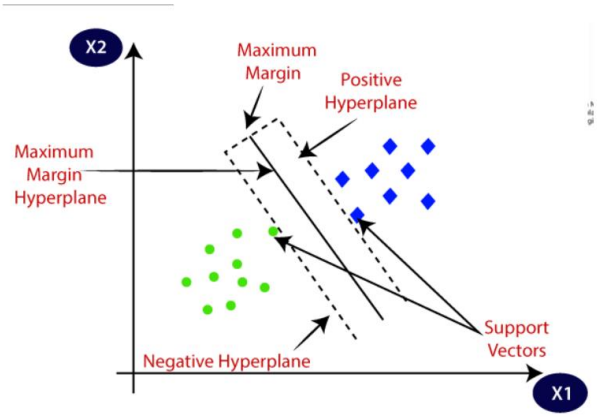


Support Vector Machine Algorithm

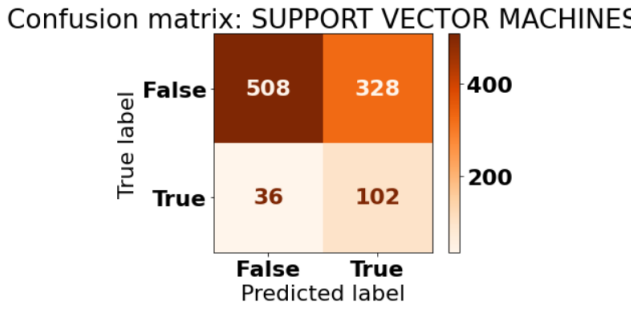
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine. Consider the below diagram in which two different categories are classified using a decision boundary or hyperplane:



	precision	recall	f1-score	support
0	0.93	0.61	0.74	836
1	0.24	0.74	0.36	138
accuracy			0.63	974
macro avg	0.59	0.67	0.55	974
weighted avg	0.84	0.63	0.68	974



Model_Name	Precision_train	Precision_test	accuracy_score_train	accuracy_score_test	roc_auc_score_train	roc_auc_score_test	recall_train	recall_test	f1_score
SUPPORT VECTOR MACHINES	0.650372	0.237209	0.665896	0.626283	0.665896	0.673393	0.717514	0.73913	0.6

Random Forest:

Random forests are an ensemble learning method for classification and regression that operates by constructing a multitude of decisiontrees at training time. For regression tasks, the output of the random forest is the average of theresults given by most trees.

In simple terms, a random forest builds multiple decision trees and merges them to get amore accurate and stable prediction.

Random Forest Hyperparameters:

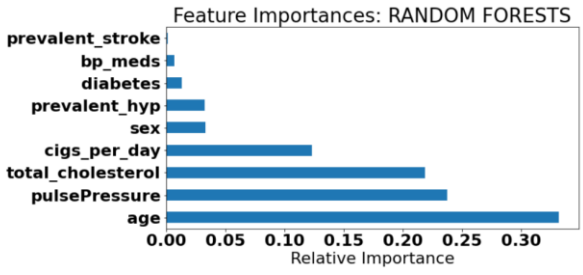
- max_depth- The max_depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node
- min_sample_split- a parameter that tells the decision tree in a random forest the minimum required number of observations in any given node to split it.
  The default value of the minimum_sample_split is assigned to 2.This means that if any terminal node hasmore than two observations and is not a pure node, we can split it further into subnodes.
- max_leaf_nodes- This hyperparameter sets a condition on the splitting of the nodes in the tree and hence restricts the growth of the tree. If after splitting we have more terminal nodes than the specified number of terminal nodes, it willstop the splitting and the tree will not grow further.
- min_samples_leaf- This Random Forest hyperparameter specifies the minimum number of samples that should be present in the leaf node after splitting a node.
- n_estimators- the number of trees
- max_sample (bootstrap sample)-The max_samples hyperparameter determines what fraction of the original dataset is given to any individual tree.
- max_features- This resembles the number of maximum features provided to each tree in a random forest.

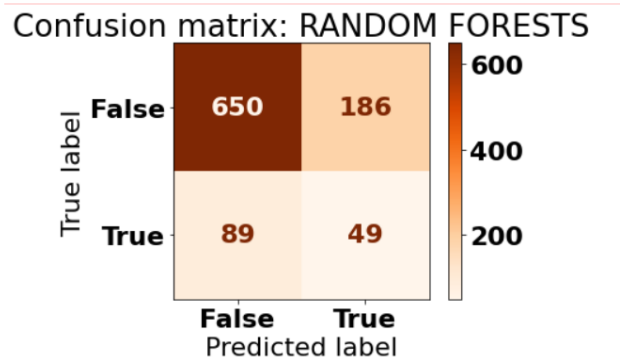
Random Forest Hyperparameter Tuned Model :

Halving Grid search CV is used to perform a random search on hyperparameters. Halving Search CV uses a fit and score method, predict probe, decision func, transform, etc.

Like Grid Search CV, all parameter values are tried out, but the time taken by Halving Search CV takes less time, The number of parameter settings that are tried is given by n_iter.

	precision	recall	f1-score	support
0	0.88	0.78	0.83	836
1	0.21	0.36	0.26	138
accuracy			0.72	974
macro avg	0.54	0.57	0.54	974
weighted avg	0.78	0.72	0.75	974





Model_Name	Precision_train	Precision_test	accuracy_score_train	accuracy_score_test	roc_auc_score_train	roc_auc_score_test	recall_train	recall_test	f1_score_train	f1_score_test
RANDOM FORESTS	0.601211	0.208511	0.560092	0.717659	0.560092	0.566292	0.356959	0.355072	0.447954	0.262735

XG Boost:

XG Boost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. The two reasons to use XG Boost are also the two goals of the project:

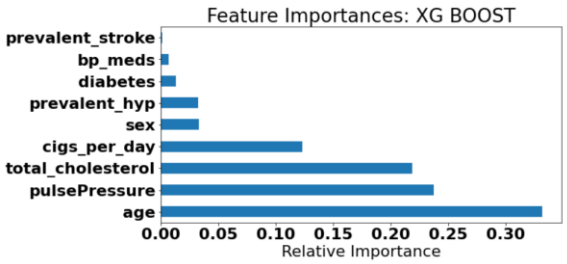
- Execution Speed.
- Model Performance.

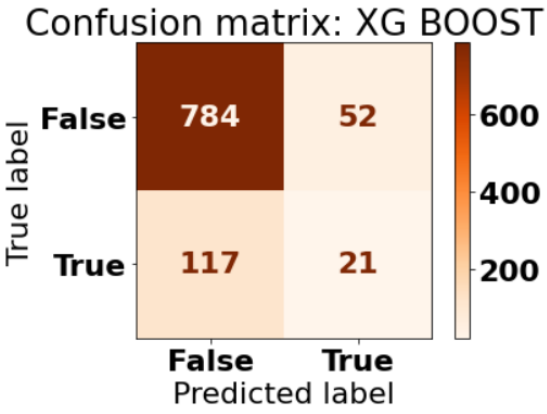
Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new **models.**

XgBoost Hyperparameters:

- max_depth- The max_depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node
- min_sample_split- a parameter that tells the decision tree in a random forest the minimum required number of observations in any given node to split it.
  The default value of the minimum_sample_split is assigned to 2.This means that if any terminal node hasmore than two observations and is not a pure node, we can split it further into subnodes.
- max_leaf_nodes- This hyperparameter sets a condition on the splitting of the nodes in the tree and hence restricts the growth of the tree. If after splitting we have more terminal nodes than the specified number of terminal nodes, it willstop the splitting and the tree will not grow further.
- min_samples_leaf- This Random Forest hyperparameter specifies the minimum number of samples that should be present in the leaf node after splitting a node.
- n_estimators- the number of trees
- max_sample (bootstrap sample)-The max_samples hyperparameter determines what fraction of the original dataset is given to any individual tree.
- max_features- This resembles the number of maximum features provided to each tree in a random forest

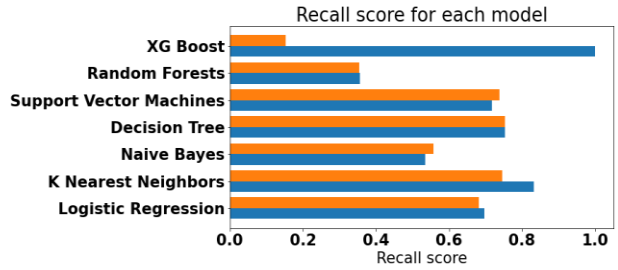
	precision	recall	f1-score	support
0	0.87	0.94	0.90	836
1	0.29	0.15	0.20	138
accuracy			0.83	974
macro avg	0.58	0.54	0.55	974
weighted avg	0.79	0.83	0.80	974





Model Summary

	Model_Name	Precision_train	Precision_test	accuracy_score_train	accuracy_score_test	roc_auc_score_train	roc_auc_score_test	recall_train	recall_test	f1_score_train	f1_score_test
0	LOGISTIC REGRESSION	0.669653	0.226181	0.693235	0.642710	0.695235	0.630762	0.698542	0.691159	0.677839	0.350746
1	K NEAREST NEIGHBORS	0.671085	0.220557	0.712121	0.593549	0.712121	0.655485	0.832149	0.746377	0.742349	0.340496
2	NAIVE BAYES	0.671180	0.257525	0.636364	0.709446	0.636364	0.646210	0.534699	0.557971	0.598197	0.352403
3	DECISION TREE	0.628156	0.208835	0.693826	0.569575	0.693826	0.641166	0.759820	0.765625	0.698341	0.320744
4	SUPPORT VECTOR MACHINES	0.693072	0.237209	0.669896	0.626283	0.669896	0.673593	0.717514	0.759130	0.692295	0.359155
5	RANDOM FORESTS	0.671211	0.208911	0.760592	0.717559	0.760592	0.596250	0.559899	0.559072	0.447954	0.262735
6	XG BOOST	1.000000	0.287671	1.000000	0.825489	1.000000	0.644986	1.000000	0.152174	1.000000	0.199552



The recall of the decision tree is the highest among all the models that we have used here and XG Boost has given us the lowest recall of all.

Conclusion:

Conclusion:

The main objective of risk prediction is to paint an accurate picture of expected fatal issues. Doctors aim to either hit their expected target or exceed it.

When the sales forecast is accurate, operations gosmoothly and future planning for the company's growth is done efficiently.

Healthcare centers use Risk predictors to determine whether in thecoming years the person is susceptable of any health-related problem or not.

Thehealth of an Individual represents the lifestyle one is living and what habits he is following.

The work here predicts whether the person is having any cardiovascular risk or not and compare the results from the models developed.

Some Important conclusion drawn from the analysis are as follow:

- The **Target Value distribution is highly imbalanced**. As in, the number of negative cases outweighs the number of positive cases. This would lead to a class imbalance problem while fitting our models. Therefore, this problem needs to be addressed and taken care of.
- According to this dataset, **males have shown a slightly higher risk of coronary heart disease Ten Year CHD**.
- Since the data we are dealing with is unbalanced, accuracy may not be the best evaluation metric to evaluate the mode
- Also, since we are dealing with data related to healthcare, **False Negatives are of higher concern than False Positive**
- In other words, it doesn't matter whether we raise a **false alarm but the actual positive cases should not go undetected.**
- Considering these points, we decided to use **Recall** as the model evaluation metric.
- It is critical that the model we develop has a high recall score. **It is OK if the model incorrectly identifies a healthy patient as a high-risk patient because it will not result in death, but if a high-risk patient is incorrectly labeled as healthy, it may result in fatality**.
- The **recall** score of the **KNN**, **Decision tree, and SVM are the highest** among all the models we have used here and XG Boost has given us the lowest recall.
- We were able to **create a model with a recall of just 0.75** because of limitations in computational power availability. This indicates that out of 100 individuals with the illness**, our model will be able to classify only 75 as high-risk patients, while the remaining 25 will be misclassified.**

Recommendations:

- With more knowledge from an expert in the field of cardiovascular health, new variables could be developed to enhance the predictions further.
- Better models other than the ones used here could be used to improve predictions.
- This project has provided experience in an important field of healthcare and has clearly illustrated the application of machine learning in this field.

Challenges:

Comprehending the problem statement, and understanding the business implications– understanding the importance of predicting the risk of this disease

- Handling missing values in the dataset, and working with limited availability of data
- Feature engineering–deciding on which features to be dropped/kept/transformed
- Choosing the best visualization to show the trends among different features clearly in the EDA phase
- Choosing the best hyperparameters, which prevents overfitting

References:

- Machine Learning Mastery
- GeeksforGeeks
- Analytics Vidhya Blogs
- Towards Data Science Blogs
- Built-in Data Science Blogs
- Scikit- Learn Org
- Investopedia