**ABSTRACT:**

A lengthy, ongoing study of Framingham, Massachusetts, citizens' cardiovascular health is known as the Framingham Heart Study.With 5,209 adult Framingham volunteers at the study's outset, it has now recruited its third generation of participants.The epidemiology of hypertensive or arteriosclerotic cardiovascular disease remained mostly unknown prior to the study.This longitudinal study served as the foundation for much of the knowledge that is now widely known about heart disease, including the impact of diet, exercise, and common drugs like aspirin.It is an initiative of the National Heart, Lung, and Blood Institute in partnership with Boston University (from 1971).The project is staffed by a variety of medical specialists from Greater Boston's hospitals and academic institutions.

**INTRODUCTION:**

Globally, heart disease is the leading cause of morbidity and mortality, killing more people each year than any other disease combined.

The WHO estimates that 17.9 million deaths worldwide from heart disease occurred in 2016, accounting for 31% of all fatalities.

More than 75 percent of these fatalities occurred in developing and middle-income nations.

Coronary heart disease, sometimes known as a heart attack, is by far the most prevalent and lethal of all heart conditions.

For instance, it is believed that someone in the United States experiences a heart attack every 40 seconds, and that there are roughly 805,000 heart attacks there each year (CDC 2019).

**PROBLEM STATEMENT:**

The results come from an ongoing cardiovascular study of people living in the Massachusetts town of Framingham.

▪ The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

▪ The dataset contains information on the patients.There are 17 qualities and more than 3,000 records in it.

**DATA DESCRIPTION:**

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 smoker.

• 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 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

| **MISSING VALUES TREATMENT:**  When there is no value stored for the variable in an observation, it results in missing data, also known as missing values.  The conclusions that can be derived from the data can be significantly impacted by missing data, which is a typical occurrence.  Further to that, we discovered missing observations in seven columns, which we then addressed with the column's corresponding median value. |  |
| --- | --- |

**EXPLORATORY DATA ANALYSIS:**

We were able to understand how various features in our dataset affect the target variable thanks to the exploratory data analysis we conducted on our train dataset.

**HANDLING OUTLIERS:**

An outlier is a data point that significantly deviates from the rest of the results. In our data, an outlier is an observation that deviates from the norm or is anomalous relative to other observations. For any feature contained in the datasets, the box plot may be easily created in order to identify outliers. Points that fall outside of the range of values are considered outliers. With respect to its median values, we have dealt with the outlier that is distant from the upper and lower boundaries (or 50th percentile).

**CLEANING AND MANIPULATING**

**THE DATASET:**

Data manipulation is the process of changing data in order to organise and make it more readable. locating and eliminating redundant values from datasets. Consider some categorical features and use value counts to provide us with some insight into that column. When a label encoding is specified in two columns, labels are transformed into numeric form to make them machine-readable.

**UNIVARIATE ANALYSIS:**

When one variable is the subject of an analysis, it is known as a univariate analysis, and its properties are intended to be discovered and identified. The most fundamental analysis method is this one, and it is frequently employed in many other kinds of research. At a glance, we can determine if the data we utilise is normally distributed, left- or right-handed, or if there are outliers, for example. Knowing the mean median from a data set, as well as the size of the distribution and concentration.

**BIVARIATE ANALYSIS:**

In a bi-variate analysis, the link between two independent variables and one dependent variable is determined. using association and dissociation tests in bi-variate analysis with a predetermined level of significance. Any combination of continuous and categorical variables can be subjected to bi-variate analysis.

**MULTICOLLINEARITY:**

* Multiple independent variables in a model that are correlating from a statistical concept known as multicollinearity.
* If a pair of variables have a correlation coefficient of +/- 1.0, they are said to be perfectly collinear.
* Less trustworthy statistical judgments will be the result of multicollinearity among independent variables.
* When creating multiple regression models that incorporate two or more variables, it is preferable to employ independent variables that are neither repetitious nor associated.
* Due to greater standard errors, multicollinearity in data collection can produce less trustworthy conclusions.
* It is not advisable to build models on top of the highly dependent variables that Sys bp, Dia bp, BMI, Glucose, Heart Rate, Age, and Total Cholesterol pose.

**VARIANCE INFLATION FACTOR:**

The intensity of multicollinearity in regression analysis is gauged by the variance inflation factor (VIF).

It is a statistical concept that describes how collinearity causes a regression coefficient's variance to grow. We often take a VIF score of less than 5 into account.

**V.I.F = 1/1-R^2**

After considering the features with VIF score less than 10 we are finally left with the following features: age, education, sex, cigsperday, prevalentHYP, BPmeds, diabetes, prevalentstroke.

**MODEL BUILDING**

## PREREQUISITES:

**CLASS IMBALANCED ISSUE:**

Using a dataset of patients, we must determine whether a person will eventually develop a cardiovascular disease based on the given features or symptoms.But there's a problem, Only 15% of people have this condition, making it relatively unusual.

## THE METRIC TRAP:

The metrics we use to measure the performance of our model are one of the main problems when working with unbalanced datasets.It can be misleading to use simple measures like an accuracy score.Without conducting any feature analysis in a dataset with extremely unbalanced classes, the classifier will always "predict" the most prevalent class, which, despite having a high accuracy rate, is obviously not the correct one. Therefore, we must address the disparity between the classes.There are various approaches to taking on this. Let’s see how:

## RANDOM OVER-SAMPLING:

Increasing the number of copies of the minority class is the definition of oversampling.When you don't have a lot of data at your disposal, oversampling may be a wise decision.Under-sampling has drawbacks, including the potential for overfitting and inadequate test set generalization.

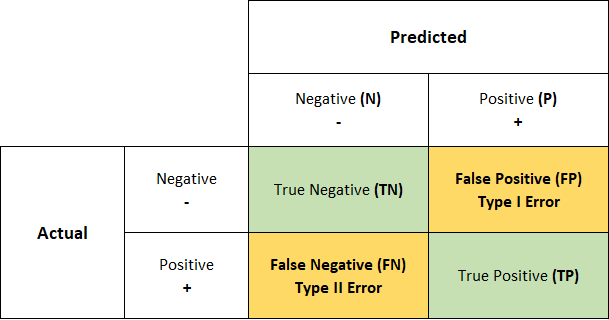
## SMOTE-TOMEK:

The ability of SMOTE to create artificial data for the minority class and the capability of Tomek Links to eliminate data from the majority class that is identified as Tomek links, or samples of data from the majority class that are most similar to the minority class data, are combined in this method.Only the train data was oversampled; the test data was not duplicated artificially.

The oversampling strategy will introduce data points close to current data points belonging to the same class, and this may result in an inaccurate representation of your test data. Your model may work well (may get a better score in Val) but in the future after deploying, it may not work better so while training, validate with imbalanced data only.

## 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.



## AREA UNDER CURVE –RECEIVER OPERATING CHARACTERISTICS (AUC ROC)

## 

## An indicator of performance for classification issues at different threshold levels is the AUC-ROC curve.The AUC reflects the level or measure of separability, and the ROC is a probability curve.

## It reveals how well the model can differ across classes.The model is more accurate at classifying 0 classes as 0, and classifying 1 class as 1, the higher the AUC. By analogy, the model is more effective at differentiating between patients with the condition and those who do not have it the higher the AUC.

## CLASSIFICATION REPORT:

The accuracy of predictions made by a classification algorithm is evaluated using a classification report.How many of the forecasts came true and how many didn't?To be more precise, the metrics of a categorization report are predicted using True Positives, False Positives, True Negatives, and False Negatives.

## METRICS:

* **Accuracy score**: which is the ratio of the number of correct predictions to the total number of input samples. It measures the tendency of an algorithm to classify data correctly.
* **Precision**: Precision is the ratio between the True Positives and all

the Positives. For our problem statement, that would be the measure of patients that we correctly identify as having a heart disease out of all the patients actually having it.Precision and Recall - Precision

* **Recall**: The recall is the measure of our model correctly identifying True Positives. Thus, for all the patients who actually have heart disease, recall tells us how many we correctly identified as having a heart disease.

Recall

* **F1 Score**: There are also a lot of situations where both precision and recall are equally important. For example, for our model, if the doctor informs us that the patients who were incorrectly classified as suffering from heart disease are equally important since they could be indicative of some other ailment, then we would aim for not only a high recall but a high precision as well. In such cases, we use something called F1-score. F1-score is the Harmonic mean of the Precision and Recall:

## FEATURE IMPORTANCE:

It assigns the score of input features based on their importance to predict the output. The more the features will be responsible to predict the output more will be their score.We can use it in both classification and regression problems.

## HYPERPARAMETER TUNING:

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process.

**N\_NEIGHBOUR’S:** Number of neighbours to use by default 5.’

**N\_ESTIMATOR’S**: number of trees in the forest, by default 100.

**MAX\_DEPTH:** It governs the maximum height up to which the trees inside the forest can grow. The default is set to None.

**MIN\_SAMPLES\_SPLIT:** It specifies the minimum number of samples an internal node must hold in order to split into further nodes. However, the default value is set to 2.

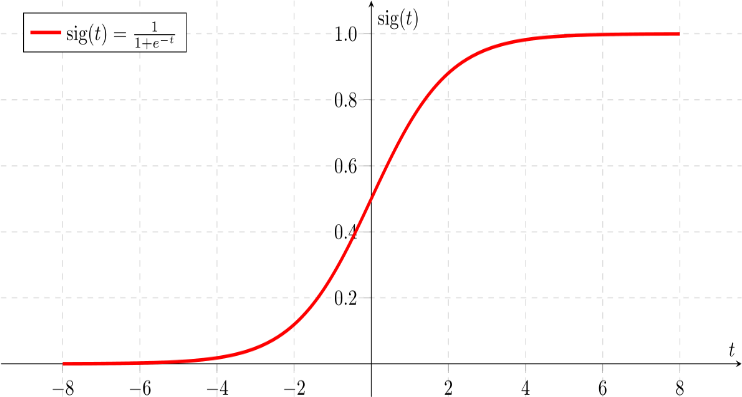
**MIN\_SAMPLES\_LEAF:** It specifies the minimum number of samples that a node must hold after getting split. The default value is set to 1.

**LEARNING RATE**: The learning rate determines the step size at each iteration while your model optimizes toward its objective. A low learning rate makes computation slower, and requires more rounds to achieve the same reduction in residual error as a model with a high learning rate.

**KERNEL:** Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. Some examples are radial basis function, Gaussian, sigmoid and polynomial kernel.

## LOGISTIC REGRESSION:

In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. Rather than providing the exact values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false, etc.

So, logistic regression cost function

## 

## NAIVE BAYES CLASSIFIER:

Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur. Because it presumes that the occurrence of one trait is unrelated to the occurrence of other features, it is known as naive. For instance, if the fruit is recognized as an apple based on its red, spherical, and delicious fruit, form, and flavor. So, without relying on one another, each characteristic helps to recognize it as an apple. As a result of its reliance on Baye's theorem, it is known as Bayes.

Baye's theorem has the following formula:

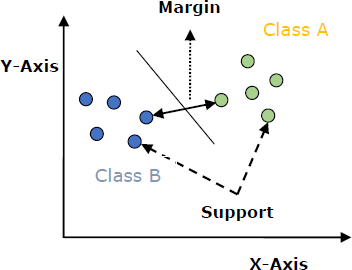
Naïve Bayes Classifier Algorithm

* **P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.
* **P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true
* **P(A) is Prior Probability**: Probability of hypothesis before observing the evidence.
* **P(B) is Marginal Probability**: Probability of Evidence.

## SUPPORT VECTOR CLASSIFIER:

An SVM model is just a hyperplane in multidimensional space that represents several classes. SVM will generate the hyperplane in an iterative manner in order to reduce error.

SVM aims to classify datasets in order to find a maximum marginal hyperplane (MMH).



## RANDOM FOREST CLASSIFIER:

A randomly chosen subset of the training data is used by the Random Forest classifier to generate a collection of decision trees. It generally consists of a number of decision trees (DT) drawn from a randomly chosen subset of the training set, and it then compiles the votes from various DTs to determine the final prediction. In layman's terms, random forest may generate three decision trees using input from a subset of the training set, which is presented as [X1, X2, X3, X4] with corresponding labels as [L1, L2, L3, L4].

## XGBOOST CLASSIFIER:

## 

Decision trees are generated sequentially in this approach.

Weights are significant in XGBoost. Each independent variable is given a weight before being fed into the decision tree that forecasts outcomes. Increased weight is applied to factors that the tree incorrectly anticipated before feeding those variables to the second decision tree. These distinct classifiers/predictors are then combined to produce a robust and accurate model. It can be used to solve problems including regression, classification, ranking, and custom prediction.

## K- NEAREST NEIGHBOUR’S:

* Instance-based learning: Here, we use full training instances to predict output for unknown data rather than learning weights from training data to do so (as in model-based algorithms).
* Lazy learning is when a model isn't taught using practice data in advance and is instead taught when a prediction request is made for a brand-new instance.
* Non-parametric: The mapping function in KNN does not have a preset form.
* The following techniques are used to determine the separation between the two data points: Manhattan distance, Minkowski distance, Euclidean distance, and Hamming distance. Among them, Euclidean is the most widely used.

## 

## CONCLUSION

* A high recall is preferred if we wish to absolutely rule out any scenarios where the patient has a cardiac disease. A high degree of precision is preferred if we wish to avoid treating a patient who has no heart disease.
* If we want to completely rule out all circumstances where the patient has heart illness, a high recall is desired. If we want to avoid treating a patient who doesn't have cardiac disease, high levels of precision are preferable.
* Since we added synthetic data points to the training set to address the extreme class imbalance, the distribution of data across the train and test sets is different, and this difference is what causes the high performance of the models in the train set rather than overfitting.
* Best performance of Models on test data based on evaluation metrics for class 1:
  + Recall - SVC
  + Precision - Naive Bayes Classifier
  + F1 Score - Logistic Regression, XGBoost
  + Accuracy - Naive Bayes Classifier