**Cardiovascular Risk Prediction**

**Project Type - Classification Algorithm**

**Contribution - Individual**

**Navneet Keshri**

**Data Science Trainee**

**AlmaBetter, Bangalore**

**Abstract:**

The most common cause of death worldwide is coronary heart disease, which is a type of cardiovascular disease (CVD). If it is discovered or diagnosed early, the odds of survival are good. Heart disease is even referred to as a "silent killer" because it kills people without showing any obvious symptoms. Using machine learning (ML) algorithms, the document discusses a comparative method for classifying coronary heart disease datasets. Several classification models based on machine learning were developed and tested in this study. SMOTE was used to deal with unbalanced classes in the dataset, and the feature selection method was used to look at how two different classes affected performance metrics.

Keywords - Explanatory Data Analysis, Classification Algorithms, Machine Learning, etc.

**Introduction:**

Over the past few years, the burden of cardiovascular disease has been rising rapidly all over the world. To accurately predict the overall risk and identify the most significant risk factors for heart disease, numerous studies have been carried out. Heart disease is even referred to as a "silent killer" because it kills people without showing any obvious symptoms. When high-risk patients are diagnosed with heart disease early, it is easier to make lifestyle changes, which in turn lowers the risk of complications. The vast amount of data generated by the healthcare industry can be effectively analyzed and predicted using machine learning. The purpose of this project is to use a machine learning algorithm to classify patient data and determine whether or not a patient has heart disease to predict the likelihood of developing the condition in the future.

A group of conditions affecting the heart and blood vessels is known as cardiovascular diseases. They consist of heart disease, which affects the blood vessels that supply the heart muscle. The majority of the time, a blockage that prevents blood from flowing to the heart or brain is to blame for heart attacks and strokes, which are typically sudden events. A build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain is the most common cause of this.

In this regard, machine learning techniques can be beneficial. Although there are many different types of heart disease, there are a few fundamental risk factors that all play a role in whether or not a person will eventually develop the condition. This method can be very well adapted for the prediction of heart disease because it collects data from a variety of sources, categorizes it under appropriate headings, and then analyses it to extract the desired data.

**Problem Statement:**

The greatest obstacle facing the medical industry is accurately predicting and diagnosing heart disease. Heart diseases are influenced by numerous factors.

Heart disease is even referred to as a "silent killer" because it kills people without showing any obvious symptoms. When high-risk patients are diagnosed with heart disease early, it is easier to make lifestyle changes, which in turn lowers the risk of complications. Based on the way people currently live, machine learning can help predict the likelihood of heart disease in the coming years. We must evaluate various classification algorithms and recommend the most accurate method for predicting coronary heart disease risk.

The data comes from an ongoing cardiovascular study of Framingham, Massachusetts, residents. The goal of the classification is to figure out if the patient will get coronary heart disease (CHD) in ten years. Patient-related data can be found in the dataset.

Over 3390 records and 17 attributes are present in total. Which of these is the target variable? A potential risk factor is present in each characteristic. Medical, behavioral, and demographic risk factors exist.

**Data Description:**

* **age:** Age of the patient (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
* **education:** level of education from 1 to 4 (Ordinal Variable)
* **sex:** male or female ("M" or "F")
* **is\_smoking:** whether or not the patient is a current smoker ("YES" or "NO")
* **cigsPerDay:** 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.)
* **BPMeds:** whether or not the patient was on blood pressure medication (Nominal)
* **prevalentStroke:** whether or not the patient had previously had a stroke (Nominal)
* **prevalentHyp:** whether or not the patient was hypertensive (Nominal)
* **diabetes:** whether or not the patient had diabetes (Nominal)
* **totChol:** total cholesterol level (Continuous)
* **sysBP:** systolic blood pressure (Continuous)
* **diaBP:** diastolic blood pressure (Continuous)
* **BMI:** Body Mass Index (Continuous)
* **heartRate:** heart rate (Continuous - In medical research, variables such as heart rate though, in fact, discrete, are considered continuous because of the large number of possible values.)
* **glucose:** glucose level (Continuous)
* **TenYearCHD:** (binary: “1”, means “Yes”, “0” means “No”)

## **Exploratory Data Analysis:**

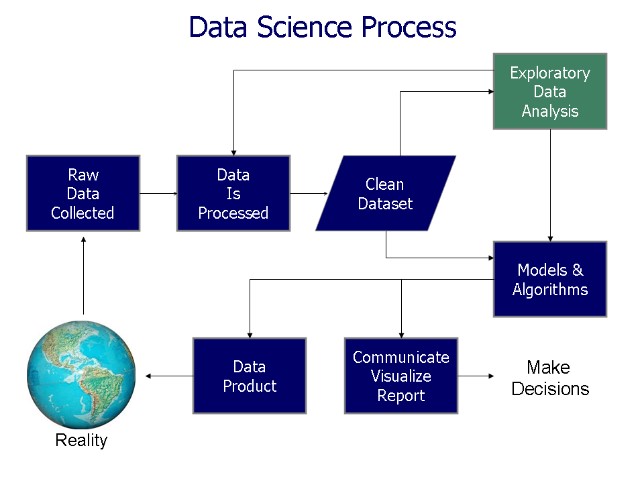
## Exploratory data analysis (EDA) is a method in statistics that uses statistical graphics and other data visualization techniques to summarize the main characteristics of data sets. A measurable model can be utilized or not, yet basically, EDA is for seeing everything that the information can say to us past the proper demonstration and in this manner contrasts conventional speculation testing. We were able to determine various aspects and relationships between the target and the independent variables with the assistance of EDA.

## **Observation:**

## In our case, most of the categorical features are of the binary type, so the values are 0 and 1 (with a few exceptions). In terms of the numerical features, each one has a unique set of continuous and range values.

## For numerical features, we can see that the majority of distributions are right-skewed. The distributions of totChol (total cholesterol) and BMI are roughly comparable. The distribution of glucose is highly skewed to the right. It demonstrates that glucose has many outliers.

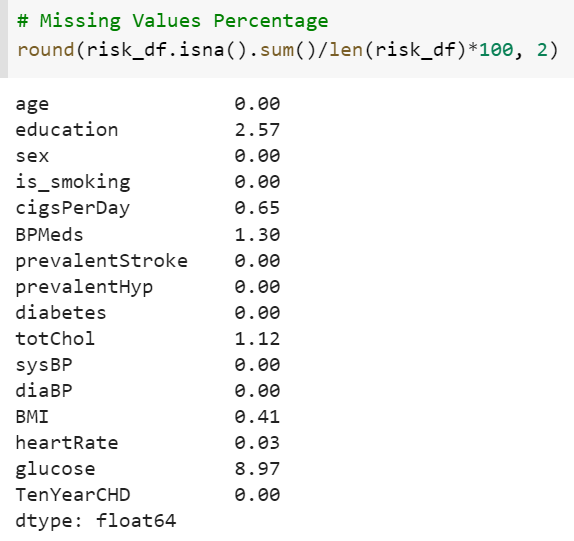
* Outliers are visible in the 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', and 'glucose' columns.
* In the 'education' column majority of the count falls under category 1, with fewer falling under category 4.
* When compared to male patients, female patients are greater in numbers.
* The proportion of non-smokers and smokers is nearly identical.
* Majority of patients do not take blood pressure medication.
* Fewer patients had suffered a stroke previously.
* Lower proportion of patients had hypertension in the past.
* Diabetes was present in fewer patients.
* Lower number of patients had a risk of coronary heart disease (CHD) over 10 years. (Target variable - Imbalance data)
* There is a clear relation between age and heart disease, with the increase in age the chance of heart disease increases. There is no significant relationship between heart disease and heart rate.



## **Data Cleaning:**

The first step in data pre-processing was to use descriptive statistics tables, skewness, and other descriptions like min, max, percentile values, and mean to show the raw data. It also includes textual data preprocessing for clustering and the detection and elimination of missing values.

**Null Value Treatment:** Replace the values that are missing for education, cigs per day, total cholesterol, BMI, glucose, and heart rate. The remaining features will be imputed using the mean, median, and mode imputation methods for those with less than 5% missing values.

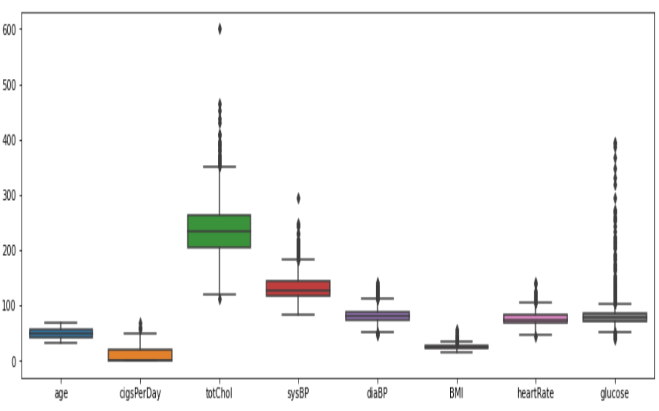


Typically, we use other records to replace these null values. However, the entries in this dataset are person-specific. The values vary from person to person, and the dataset is related to the medical field in this particular instance. Consequently, removing rows with any null value is the most logical choice we have for dealing with such values.

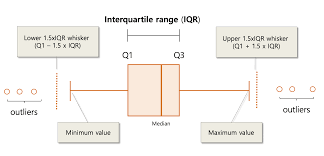
We cannot take any risks with this prediction, so if we attempt to impute null values using advanced methods, it may affect the outcome because the values will be incorrect.

In the healthcare industry, every piece of data is crucial. Because of this, we came up with a solution by setting a threshold value. If a feature has less than 5% null values, we decide to drop those rows, and the remaining rows are imputing, which will affect prediction but not significantly.

**Outliers Treatment:** Since we have limited datapoint hence we are not simply removing the outlier instead of that we are using the clipping method.



**Clipping Method:** In this method, we set a cap on our outliers data, which means that if a value is higher than or lower than a certain threshold, all values will be considered outliers. This method replaces values that fall outside of a specified range with either the minimum or maximum value within that range.



The outliers in the data frame were removed using the IQR technique. Outliers can be identified in the following columns from the box plot in the preceding figure:

Cigarette Smoking Per Day, Total Chol, SysBP, DiaBP, BMI, Heart Rate, and Glucose Although "totChol and sysBP have extremes, the majority of outliers are close to the upper whisker, which is significant."

**Feature Engineering:**

Feature engineering is the process of creating new features from existing ones to improve the performance of a machine-learning model. This involves transforming raw data into a more useful and informative form, by either creating new features from the existing data or selecting only the most relevant features from the raw data.

The goal of feature engineering is to extract relevant information from the raw data and represent it in a way that can be easily understood by the machine learning model. The success of a machine learning model depends heavily on the quality of the features used as inputs, so feature engineering plays an important role in model performance.

* **Encoding:**

Encoding is a technique in feature engineering that is used to convert categorical variables into numerical values that can be used by machine learning algorithms.

There are several encoding techniques, including:

One-hot encoding: creates a binary column for each unique category, with a value of 1 indicating the presence of the category and 0 indicating the absence.

Label encoding: assigns a unique integer value to each category.

Ordinal encoding: assigns an ordered integer value to each category based on the natural ordering of the categories.

Count encoding: replaces a categorical value with the number of times it appears in the dataset.

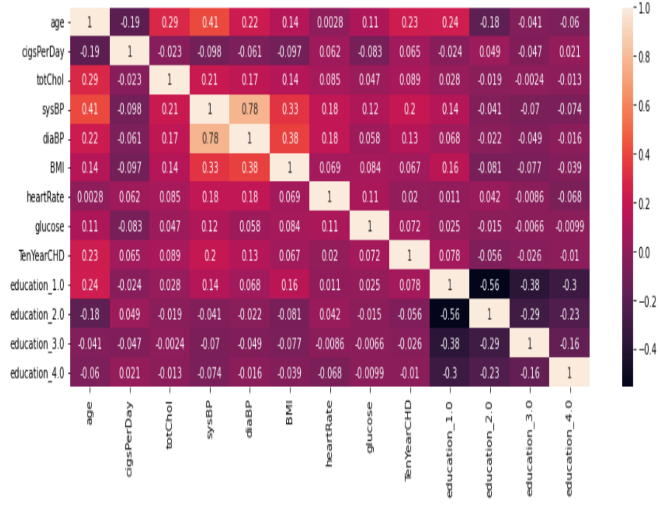
Every categorical feature except 'education' is binary and already encoded with 1 and 0 values hence we do not need to use One Hot Encoding for any column but we can use One Hot Encoding for education.

We're using an algorithm that treats the categorical variables as unordered, such as decision trees or random forests, hence one-hot encoding can be more effective in representing the categorical variables.

* **Correlation Matrix:**

The correlation coefficient is a numerical measure of the strength and direction of a linear relationship between two variables. In other words, it measures the extent to which changes in one variable are associated with changes in the other variable. The correlation coefficient ranges from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation.

A correlation coefficient is an important tool in data analysis and machine learning, as it can help to identify relationships between variables and can be used in feature selection techniques to remove highly correlated features, which can reduce overfitting and improve the performance of the model.

****

The following is what we can see from the correlation matrix above:

* The strongest correlation exists between daily cigarette consumption and smoking.
* The systolic BP and diastolic BP correlation are highest.
* Hypertension is strongly correlated with both systolic and diastolic blood pressure.
* The risk of heart disease is primarily influenced by age, prevalent hypertension, systolic and diastolic blood pressure, and other variables.
* Except for education, there is a positive correlation between all variables and the dependent variable.

We added a new feature in

replacement of the above to reduce the

multicollinearity and that feature is Pulse

Pressure.

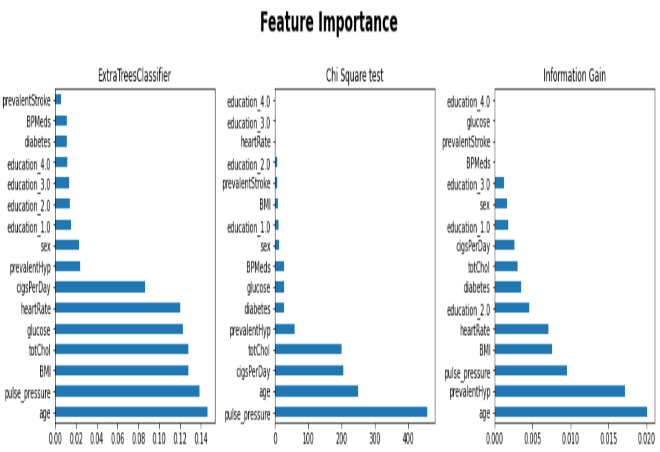
pulse\_pressure = sysBP – diaBP

* **Feature Selection:**

This section of the study involves both the selection of relevant features from the group of attributes and feature selection, which is the process of extracting the most relevant features from the dataset.

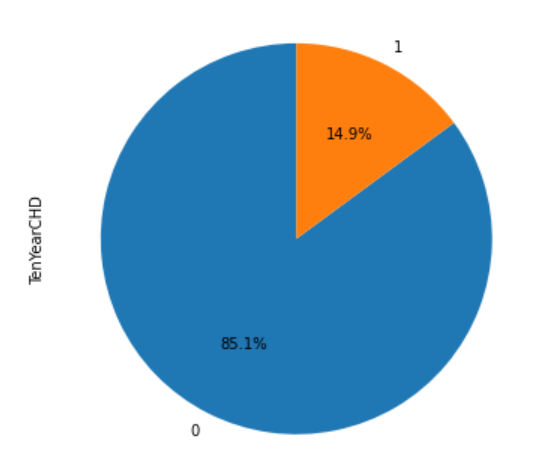
We utilized a variety of feature selection techniques.

1. Extra Trees Classifier
2. Chi-Square Test
3. Information Gain



* **Handling Class Imbalance:**

When there are significantly more instances of certain classes than others, the issue of class imbalance typically arises. Class imbalance in the target class is a problem for machine learning models because it can result in biased predictions. That is why we need to balance the target class.



The data set differs significantly. Our data, therefore, lack balance. We will use the Synthetic Minority Oversampling Technique (SMOTE) to resolve this issue.

**SMOTE** (Synthetic Minority Oversampling Technique) works by randomly selecting a minority class point and calculating its k-nearest neighbors. Between the selected point and its neighbors, the synthetic points are added. Continue with the steps until the data is balanced.

**Model Building:**

The creation of a model using a variety of machine-learning algorithms is the final step.

Algorithms used in this project include:

1. Logistic Regression
2. Support Vector Machines
3. K-Nearest Neighbor Classifier
4. Decision Tree Classifier
5. Random Forest Classifier with RandomizedSearchCV
6. Adaptive Boosting Classifier
7. Xtreme Gradient Boosting Classifier with GridSearchCV
8. Light Gradient Boosting Classifier with GridSearchCV

**Performance Analysis:**

The majority and minority of the count in our dataset come from the negative and positive, respectively. Accuracy scores aren't very helpful when dealing with imbalanced data or classes.

It is acceptable to classify a healthy individual as having a 10-year risk of coronary heart disease CHD (false positive) and to conduct additional medical tests; however, it is categorically unacceptable to fail to identify a particular patient or to classify a particular patient as healthy (false negative). As a result, the model's recall score will be the primary focus of our project.

Balanced accuracy is a better metric than F1 when positives and negatives are equally important. When more attention is required, F1 is an excellent scoring metric for imbalanced data.

* **Evaluation Matrices used for the comparison:**

**Accuracy**: Simply put, accuracy is the percentage of times that the classifier correctly predicts. The ratio of the total number of predictions to the number of correct predictions is how accuracy is defined. If a model has a 99 percent accuracy rate, you might think it is doing very well. However, this is not always the case and can be misleading in some situations.

When the target class is well-balanced, accuracy is useful, but it is not a good choice for unbalanced classes.

**Confusion Matrix:** The Confusion Matrix is a performance measurement for classification problems in machine learning in which there can be two or more classes output. It is a table with actual and predicted value combinations. The table that is frequently used to describe the performance of a classification model on a set of test data for which the true values are known is referred to as a confusion matrix. It is extremely helpful for determining the AUC-ROC curves, precision, recall, and accuracy.

**Precision:** Precision explains the percentage of correctly predicted cases that were successful. When False Positives are more of a concern than False Negatives, precision can be useful.

A label's precision is calculated by dividing the number of predicted positives by the number of true positives.

**Recall**: Recall describes the proportion of actual positive cases that our model correctly predicted. When a False Negative is more important than False Positive, this metric is helpful. In medical cases, it matters whether we raise a false alarm or not, but the actual positive cases should not go unnoticed. The number of true positives divided by the total number of actual positives is the definition of recall for a label.

**F1 Score:** This score incorporates both Precision and Recall metrics. When Precision and Recall are equal, it reaches its peak.

The harmonic mean of recall and precision is the F1 Score.

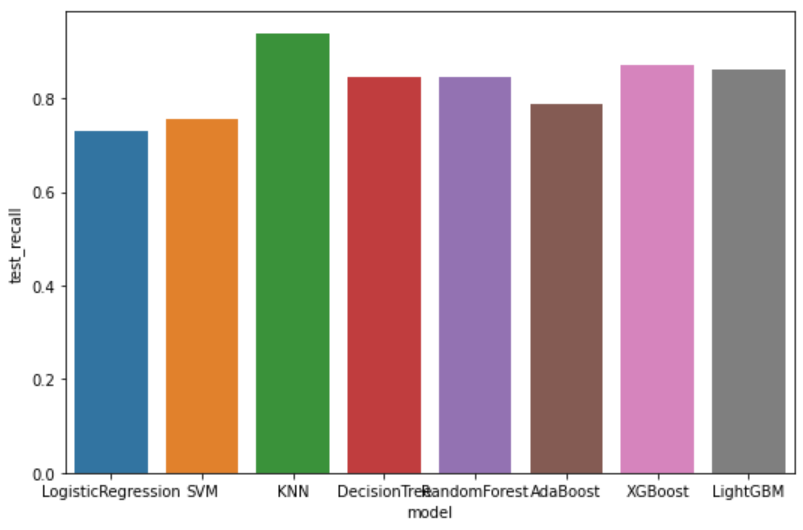
**AUC-ROC:** The Receiver Operator Characteristic (ROC) is a probability curve that separates the "signal" from the "noise" by plotting the TPR (True Positive Rate) against the FPR (False Positive Rate) at various threshold values. The measure of a classifier's ability to differentiate between classes is the Area Under the Curve (AUC). This simply indicates that the classifier can precisely differentiate between all Positive and Negative class points when AUC is equal to 1. The classifier would correctly identify all negatives as positives when the AUC was zero.

**Model Result:**

Here, we must emphasize that recall, rather than accuracy, is more important to us depending on the nature of our problem. because we do not want any situation in which a patient has a risk of coronary heart disease but is deemed to have no risk.

As a result, we must select a robust model that accurately classifies the classes.

First, we'll look at the project's confusion matrix, which includes all of the models used.



We didn't want to mispredict a person's safety when he has the risk of 10 years of CHD, so the final model we chose was KNN based on the best recall score.

**Conclusion:**

In this project, we tackled a classification problem in which we had to classify and predict the 10-year risk of future coronary heart disease (CHD) for patients. The goal of the project was to develop a tool for the early detection and prevention of CHD, addressing a significant public health concern using machine learning techniques.

* There were approximately **3390 records and 16 attributes** in the dataset.
* We started by importing the dataset, and necessary libraries and conducted exploratory data analysis (EDA) to get a clear insight into each feature by separating the dataset into numeric and categoric features. We did Univariate, Bivariate, and even multivariate analyses.
* After that, the outliers and null values were removed from the raw data and treated. Data were transformed to ensure that it was compatible with machine learning models.
* In feature engineering we transformed raw data into a more useful and informative form, by creating new features, **encoding**, and understanding important features. We handled target class imbalance using **SMOTE**.
* Then finally cleaned and scaled data was sent to various models, the metrics were made to evaluate the model, and we tuned the hyperparameters to make sure the right parameters were being passed to the model. To select the final model based on requirements, we checked model\_result.
* When developing a machine learning model, it is generally recommended to track multiple metrics because each one highlights distinct aspects of model performance. We are, however, **focusing more on the Recall score and F1 score** because we are dealing with healthcare data and our data is unbalanced.
* With an **f1-score of 0.907 and a recall score of 0.863** on test data, we have noticed that **LightGBM** Classifier outperforms all other models. It is safe to say that the LightGBM Classifier is the best option for our issue if the f1-score is to be considered.
* Our **highest recall score, 0.938%, came from KNN**.
* The **XGBoost and RandomForestClassifier tree-based algorithms also provided the best approach** to achieving our goal. We were successful in achieving a respective f1-score of 0.904 and 0.893.

The recall score is of the utmost significance in the medical field, where we place a greater emphasis on reducing false negative values because we do not want to mispredict a person's safety when he is at risk. With recall scores of 0.938, 0.870, and 0.863, respectively, KNN, XGB, and LGBM performed the best.

Last but not least, we can select the Final model as our KNN classifier due to its highest recall score. It is acceptable to classify a healthy individual as having a 10-year risk of coronary heart disease CHD (false positive) and to follow up with additional medical tests; however, it is categorically unacceptable to miss identifying a particular patient or to classify a particular patient as healthy (false negative).

**References:**

* <https://www.kaggle.com/>
* <https://stackoverflow.com/>
* <https://towardsdatascience.com/>
* <https://www.analyticsvidhya.com/blog/>
* <https://github.com/>
* <https://www.geeksforgeeks.org/>