Coronary Heart Disease Prediction

Parth Sharma-Cohort Amsot

Abstract:

Coronary heart disease, which is a form of cardiovascular disease (CVD), is the leading cause of death worldwide. The odds of survival are good if it is found or diagnosed early. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms
The document discusses a comparative approach to the classification of coronary heart disease datasets using machine learning (ML) algorithms. The current study created and tested several machine-learning-based classification models. The dataset was subjected to SMOTE to handle unbalanced classes and feature selection technique in order to assess the impact on two distinct classes on the performance metrics. The results show that K-Nearest Neighbours produced the highest performance score on the original dataset compared to the other algorithms employed. In conclusion, this study suggests that K-Nearest Neighbours on a well-processed and standardized dataset can predict coronary heart disease with greater robustness than the other algorithms.

1. Problem Statement:

Predicting and diagnosing heart disease is the biggest challenge in the medical industry. There are many factors which influence heart diseases. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications. Machine learning can play a vital and accurate role in predicting chances of heart disease in coming potential years based upon the current way of living.

We need to test different classification algorithms and suggest the one that could predict the risk of Coronary Heart Disease the best.

2. Introduction:

The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict the chance of future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

3. Steps and Methods:

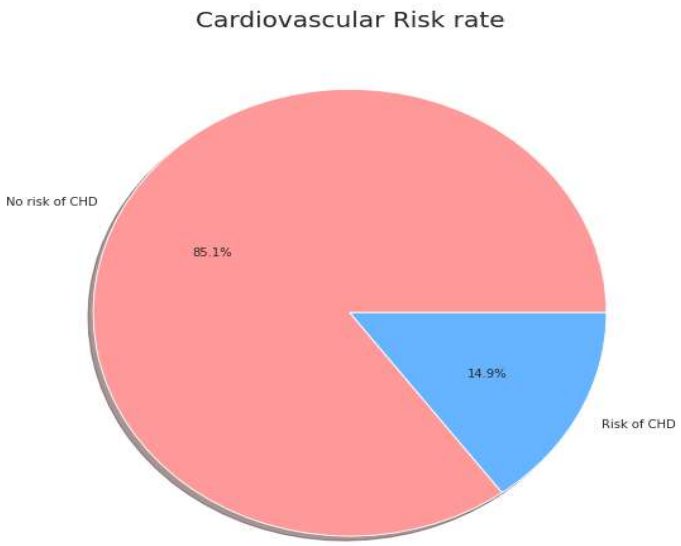
1. Dataset:

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 will develop coronary heart disease in the next ten years (CHD). The dataset contains information about the patients. There are over 3390 records and 17 attributes in total. Out of which one is our target variable. Each characteristic is a potential risk factor. There are demographic, behavioral, and medical risk factors.

TABLE I

Category	Description
Demographic	*Sex*: male or female (Nominal) *Age*: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) *Education:* Educational background of the patient ranked from 1 to 4 (continuous)
Behavioral	*Current Smoker*: whether or not the patient is a current smoker (Nominal) *Cigs Per Day*: the number of cigarettes that the person smoked on average in one day. (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
Medical (history)	*BP Meds*: whether or not the patient was on blood pressure medication (Nominal) *Prevalent Stroke*: whether or not the patient had previously had a stroke (Nominal) *Prevalent Hyp*: whether or not the patient was hypertensive (Nominal) *Diabetes*: whether or not the patient had diabetes (Nominal)
Medical (current)	*Tot Chol*: total cholesterol level (Continuous) Sys BP: systolic blood pressure (Continuous) *Dia BP*: diastolic blood pressure (Continuous) *BMI*: Body Mass Index (Continuous) *Heart Rate*: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.) *Glucose*: glucose level (Continuous)
Predict variable (desired target)	10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)



The target data which is a 10-year risk of coronary heart disease CHD has 85.1% class of 0 and 14.9% class of 1. Therefore, this makes the dataset highly imbalance.

2. Approach:

The aim is to develop classification models for predicting coronary heart disease. The purpose of this project is to evaluate the performance of various classification models developed using various machine learning algorithms. Thus proposed approach is specified as follows. First, we studied the acquired data, the parameters, and attributes and the final measures from the dataset. The data is examined for possible missing values

representations and the range of each attribute. The next stage in our practice is pre-processing of the data. The categorical features are encoded, missing values and outliers are identified and eliminated from the data. This stage is followed by feature engineering of the data to be used in model development.

The feature engineering stage, which is the data transformation, covers both selecting the relevant features or attributes and completely modifying the data points. After this stage, various classification models are built using various machine learning algorithms.
Finally, the classification metrics are calculated that shows the performance of the models, each metric has its say depending on which we decide which model solves our business problem.

3. Data Pre-processing and EDA:

Data pre-processing began with the visualization of raw data using descriptive statistics tables, skewness, and other descriptions such as min, max, percentile values, and mean. It also includes the identification and removal of missing values, as well as the encoding of categorical values. The missing values in cigsPerDay, totChol, sysBP, diaBP, BMI, glucose, heartRate, were substituted with the mean values of each column. Also, the missing values of BP Meds which is categorical and education (ordinal with range 1-4) were removed from the dataset.
Let's have a look at the missing values present in our data set.

```
Missing Data Count :
age                0
education         87
sex                0
is_smoking         0
cigsPerDay        22
BPMeds            44
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol           38
sysBP              0
diaBP              0
BMI               14
heartRate          1
glucose          304
TenYearCHD         0
dtype: int64
```

Fig: Shows the number of missing values found in each feature or attribute.

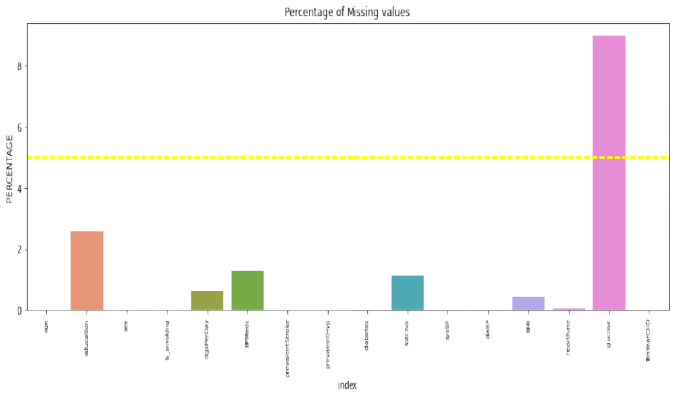


Fig: Shows the percentage of missing values found in each feature or attribute.

Replacing missing values. The cleaning of the data starts with the replacing of missing values using various techniques. In the project **ITERATIVE IMPUTATION** technique was used to fill the missing values where the percentage of missing values where greater than 5%. Replace missing values in education, cigsPerDay, totalChol, BMI, glucose, and heartRate. We have decided to directly drop all the null values since the percentages of missing values was lesser than 5% and rest will be imputed using iterative imputation method.

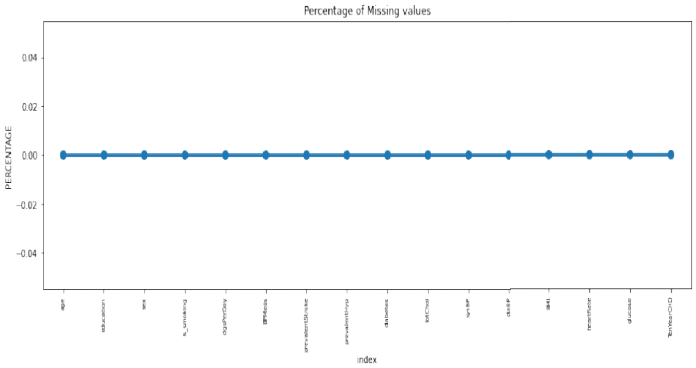
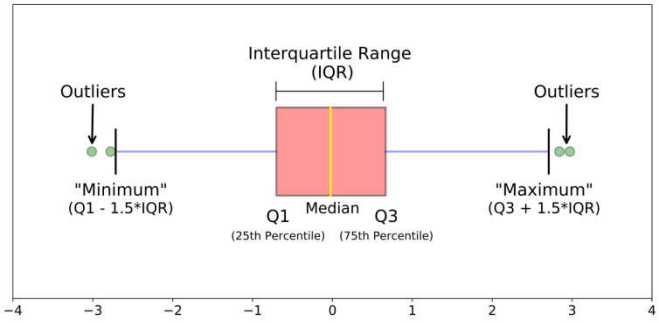


Fig: Shows the percentage of missing values in the in each feature

Further, outliers are removed from the data using the equations below.



Where Q1, Q3 represent the first quantile, the third quantile of each attribute.

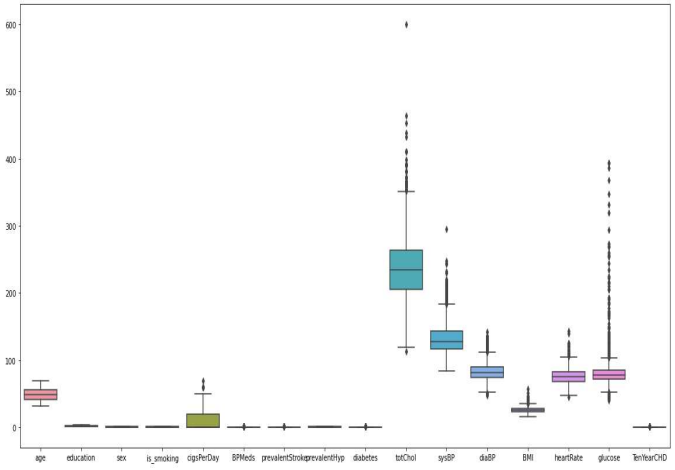


Fig: Shows the outliers present in each feature

The IQR method was used to clean the data frame outliers. From the box plot in Fig above, outliers can be found in the following columns: cigsPerDay, totalChol, sysBP, diaBP, BMI, heartRate, and glucose. Although there are extremes in 'totChol' and sysBP, but the majority of the outliers are close to the upper whisker, which is significant.

Therefore, there are no missing values in our data set. Now, we will be looking at the treated outliers.

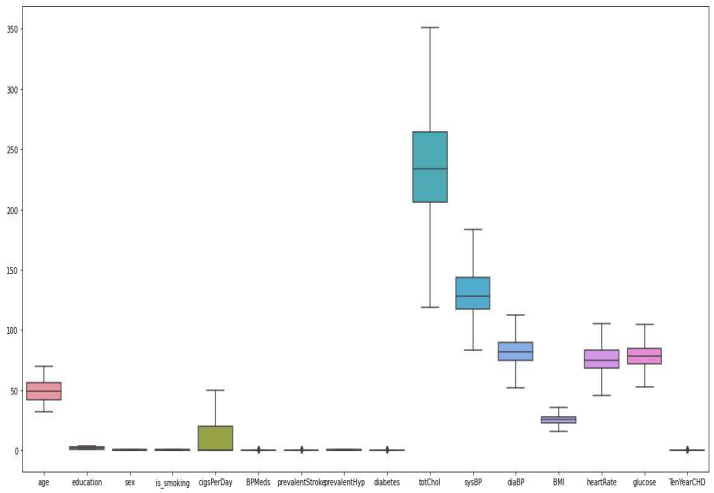


Fig: Shows that there are no outliers present in our dataset

EDA or Exploratory Data Analysis is the critical process of performing the initial investigation on the data to find the anomalies in our data and shape it such that it is useful for taking some insights to solve our purpose.

We will now have a look at some visualisation.

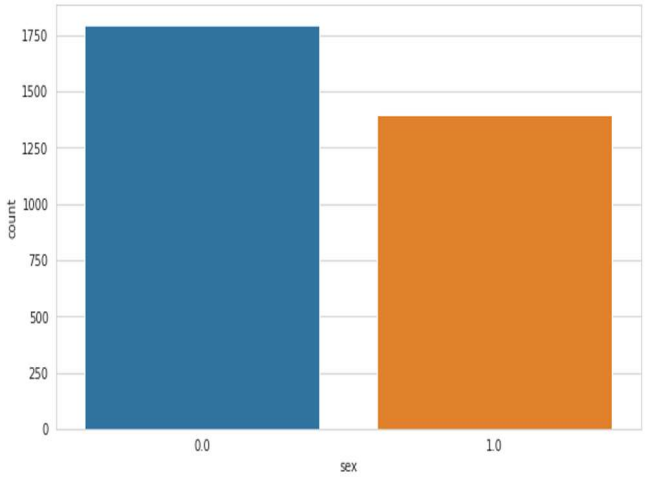


Fig: Proportion of Females(0) and Males(1)

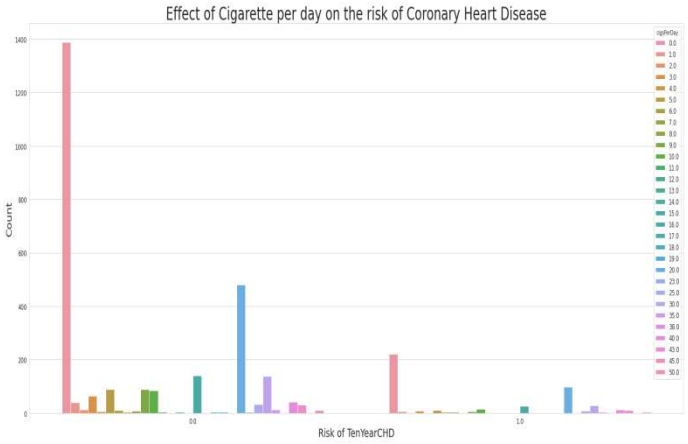


Fig: Shows the effect of number of cigarettes on the risk of CHD

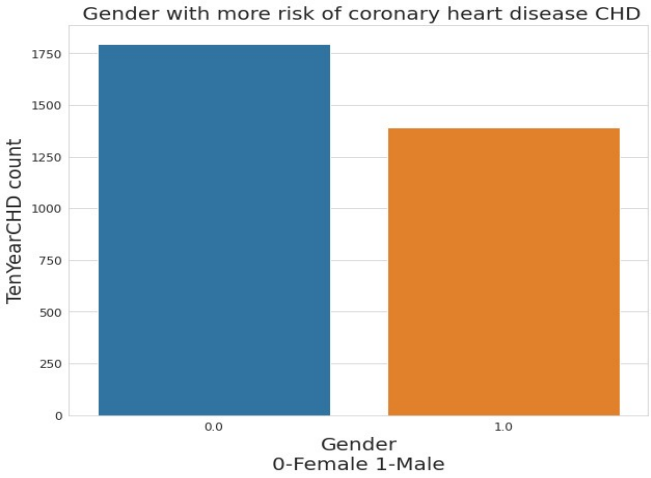


Fig: Shows the gender more prone to the risk of CHD

4. Correlation Matrix:

The correlation matrix illustrates how the features are related to one another or to the target variable. The correlation heat map is shown in Fig below

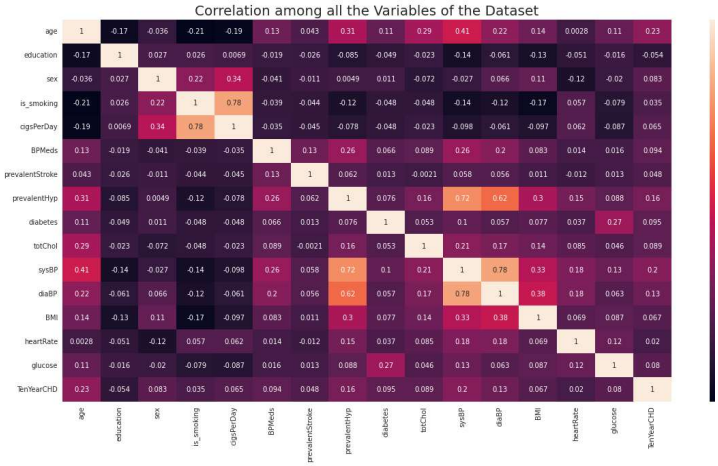


Fig: Correlation Matrix

From the above correlation matrix we can see that:

1. Highest correlation exists between systolic BP and diastolic BP.
2. Systolic and Diastolic BP shows a high correlation with hypertension.
3. Variables such as age, prevalent hypertension, systolic BP, diastolic BP, influence the risk of heart disease mainly.
4. All the variables have a positive correlation with the dependent variable, except for education.
5. Systolic BP and age have a positive correlation.

5. Feature Engineering:

Feature engineering involves feature transformation and feature selection.
Since, above correlation matrix shows that there is a good relation between the sysBP and diaBP. We added a new feature in replacement of above to reduce the multicollinearity and that feature is **Pulse Pressure.**

Pulse Pressure = sysBP - diaBP

Feature selection is also known as attribute selection is a process of extracting the most relevant features from the dataset .This section of the study entails both the selection of relevant features from the group of attributes and Chi-Score was used to select the best feature.
Feature Selection feature extraction or selection technique which employs Select Best feature selection strategy. It computes between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable

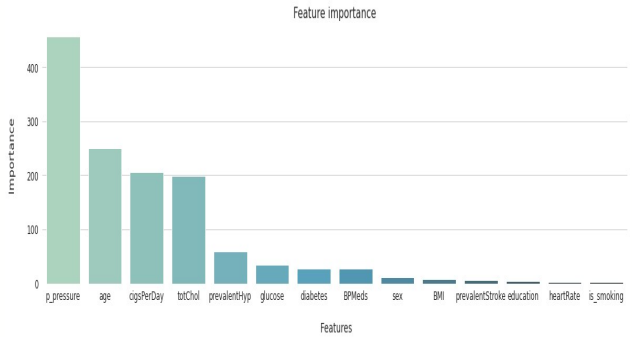


Fig: Shows the gender more prone to the risk of CHD

	Independent Feature	Chi_Score
13	p_pressure	457.250519
0	age	249.931510
4	cigsPerDay	205.859597
9	totChol	199.719662
7	prevalentHyp	58.674438
12	glucose	34.243363
8	diabetes	27.931583
5	BPMeds	27.187058
2	sex	12.252065
10	BMI	8.294096
6	prevalentStroke	7.219641
1	education	4.826685
11	heartRate	2.207399
3	is_smoking	1.960362

6. Treating the Class-Imbalance:

Under Sampling and Over Sampling :

1. Under Sampling: In under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate.

2. Over Sampling: In Over Sampling, dataset balance is done by increasing the size of the scarce samples. This process is considered when the amount of data is inadequate.

Here, we will be using SMOTE for oversampling in order to treat the imbalance in the target variable.

Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class. Simply adding duplicate records of minority class often don't add any new information to the model. In SMOTE new instances are synthesi zed from the existing data.

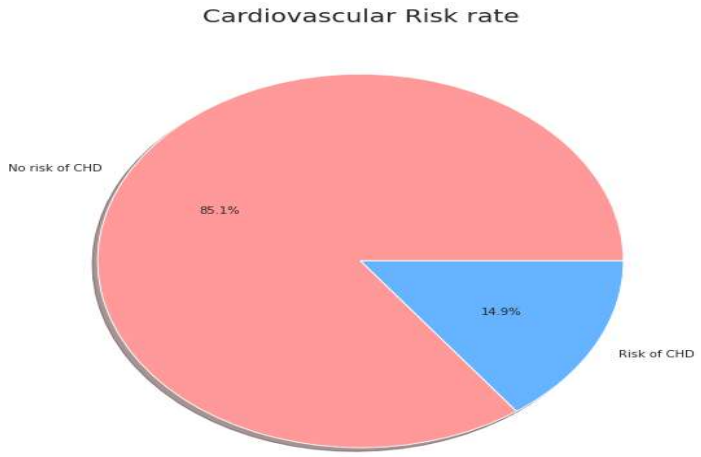


Fig: Class-Imbalance

After treating the class-imbalance the results we got are:

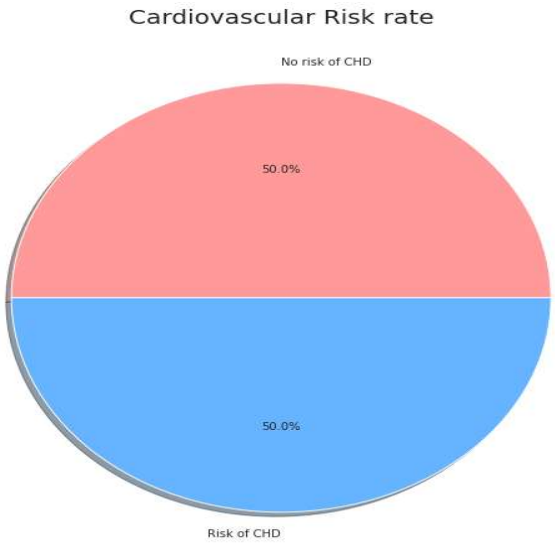


Fig: Balanced-Class

7. Machine Learning Models :

The final stage entails the development of a model using various machine learning algorithms. This project's algorithms include:

1. Logistic Regressor.
2. K-Nearest Neighbour Classifier.
3. Decision Tree Classifier.
4. Random Forest Classifier.
5. Support Vector Machines.
6. Light Gradient Boosting Machine With Grid Search CV.
7. Random Forest Classifier with Randomised Search CV.

1. Logistic Regression:

Logistic regression is also a supervised learning classification algorithm that is used to solve both classification and regression problems. In classification problems, the target variable may be in a binary or discrete format either 0 or 1. Logistic regression algorithm works on the sigmoid function, so the categorical variable results as 0 or 1, Yes or No, True or False, etc. It is a predictive analysis algorithm that works on mathematical functions. Logistic regression uses a sigmoid function or logistic function which is a complex cost function. The sigmoid functions return the value between 0 and 1. If the value less than 0.5 then it is considered as 0 and greater than 0.5 it is considered as 1. Thus to build a model using logistic regression sigmoid function is required. There are three main types of logistic regression:

1) Binomial: The target variable can have only 2 possibilities either "0" or "1" which may represent "win" or "loss", "pass" or "fail", "true" or "false", etc.

2) Multinomial: Here, the target variable can have 3 or more possibilities that are not ordered which means it has no measure in quantity like "disease A" or "disease B" or "disease C".

3) Ordinal: In this case, the target variables deal with ordered categories. For example, a test score can be categorized as: "poor", "average", "good", and "excellent".

Here, each category can be given a score like 0, 1, 2, and 3.

$$(x) = 1/(1+e\ X)$$

2. K-Nearest Neighbours:

KNN also called Knearest neighbour is a supervised machine learning algorithm that can be used for clas sification and regression problem. K nearest neighb our is non-parametric i.e. It does not make any assumptions for underlying data assumptions.Here t he algorithm classifies a input or unseen data set on the basis of characterstics shared by the nearest data points.

3. Decision Tree:

Decision Tree algorithm is also a supervised learning technique, mostly preferred for solving Classification problems but can be used for both classification and regression problems. The decision tree is a tree-structured classifier, where branches represent the decision rules which are used to make any decision and have multiple branches, internal nodes represent the features of a dataset, and each leaf node represents the outcome of the decisions and does not contain any further branches. It is a graphical representation for getting all the possible solutions to a problem/decision based on given constraints. The decisions or the analysis are performed based on features of the given dataset. A decision tree simply asks a question and based on the answer, it further splits the tree into sub trees.

4. Random Forest:

Random Forest classifier is a supervised learning technique in machine learning. It can be used to solve both Classification and Regression problems in machine learning. It is based on the process of combining multiple classifiers to solve a complex problem and to improve the performance of the model, which is known as ensemble learning. Random Forest consists of several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Rather than relying on a single decision tree, the random forest acquires the prediction from each tree, and based on the majority of votes for predictions, it predicts the final output. The higher number of trees in the forest leads to better accuracy and also prevents the problem of over fitting. The final output is taken by using the majority voting classifier for a classification problem while in the case of a regression problem the final output is the mean of all the outputs.

5. Support Vector Machine:

Support vector machine (SVM) is a supervised learning algorithm that is used to analyze data. It is used to resolve classification and regression problems. An SVM model is a delineation of the examples as points in space, mapped so that the examples of the discrete categories are divided by a clear gap. The points are separated by a plane which is known as a hyper plane. A set of training data is given to it to mark them as belonging to either one of two categories; an SVM training algorithm then builds a model that assigns new examples of the same space are mapped and then predicts to which category they belong, making it a non-probabilistic binary linear classifier. SVM can be of two types:

- **Linear SVM:**

Linear SVM is used for data that is linearly separable. It means if a dataset can be segregated into two different classes by using a single straight line, then such data is labelled as linearly separable data, and the classifier is used called a Linear SVM classifier.

- **Non-linear SVM**:

Non-Linear SVM is used for data that cannot be separated linearly, which means if a dataset cannot be sorted by using a straight line, then such data is referred to as non-linear data and the classifier used is called a Non-linear SVM classifier.

6. Light Gradient Boosting Machine With Grid Search CV:

LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.

It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfils the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSS and EFB described below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks

Gradient-based One Side Sampling Technique for LightGBM:

Different data instances have varied roles in the computation of information gain. The instances with larger gradients (i.e., under-trained instances) will contribute more to the information gain. GOSS keeps those instances with large gradients (e.g., larger than a predefined threshold, or among the top percentiles), and only randomly drop those instances with small gradients to retain the accuracy of information gain estimation.

8. PERFORMANCE ANALYSIS:

In this project, various machine learning algorithms like SVM, Decision Tree, Random Forest, Logistic Regression, Support Vector Machines, Light Gradient Boosting Machine With Grid Search CV, Random Forest Classifier with Randomised Search CV are used to predict heart disease.
For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered.

The accuracy for individual algorithms has to measure and whichever algorithm is giving the best accuracy that is considered for the heart disease prediction. For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered.

Accuracy- Accuracy is the ratio of the number of correct predictions to the total number of inputs in the dataset. It is expressed as:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

Precision- It is the ratio of correct positive results to the total number of positive results predicted by the system.

$$\text{Precision(P)} = \frac{\text{TI}}{(\text{TP} +}$$

Recall- It is the ratio of correct positive results to the total number of positive results predicted by the system.

$$\text{Recall(R)} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

F1 score- It is the harmonic mean of Precision and Recall. It measures the test accuracy. The range of this metric is 0 to 1.

$$\text{F1 score} = 2 * \frac{1}{\left(\frac{1}{\text{Precision}}\right)+\left(\frac{1}{\text{Recall}}\right)} = \frac{2PR}{(P+R)}$$

Confusion Matrix- It gives us a matrix as output and gives the total performance of the system.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Where
TP: True positive
FP: False Positive
FN: False Negative
TN: True Negative

4. Results:

Here, we have to note to make that depending upon the nature of our problem Recall is more important to us rather than accuracy. Because, we do not want any case where the patient has a risk of CHD and it is classified as there is no risk.
Therefore, we need to select that model that is robust in classifying the classes more accurately.
First, we will have a look at the confusion matrix of all the models that has been used in the project.

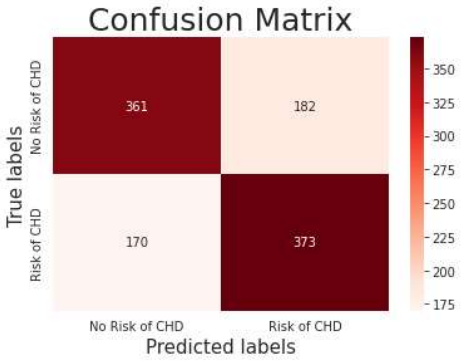


Fig: Confusion Matrix for Logistic Regression

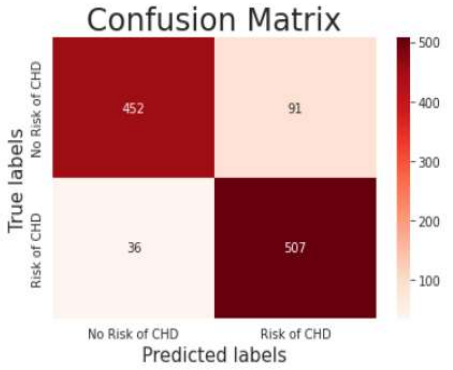


Fig: Confusion Matrix for KNN

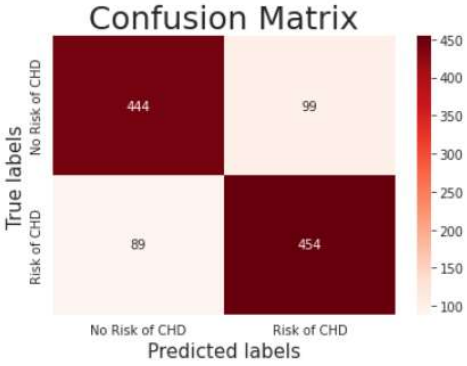


Fig: Confusion Matrix for Decision Tree

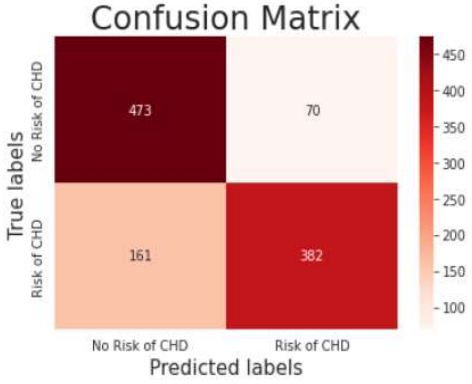


Fig: Confusion Matrix for SVM

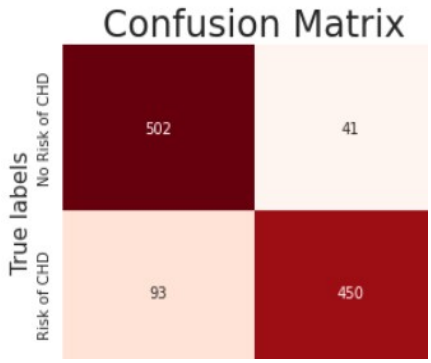


Fig: Confusion Matrix for Random Forest Classifier

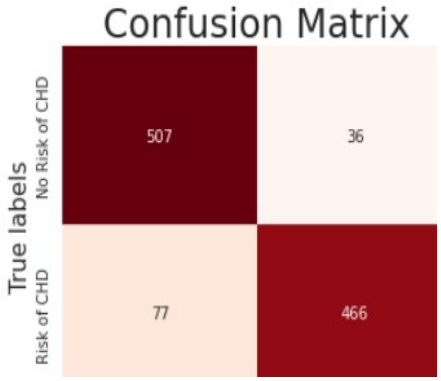


Fig: Confusion Matrix for Random Forest Classifier Using Randomized Search Cv

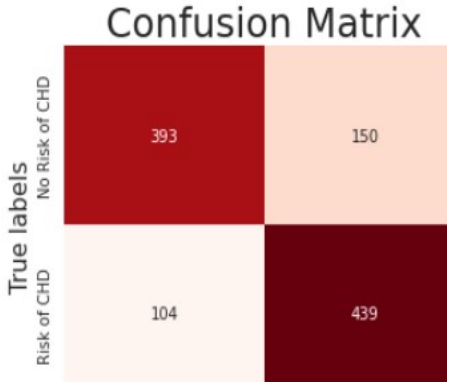


Fig: Confusion Matrix for Light Gradient Boosting Machine Using Grid Search Cv

TABLE: Recall comparison of algorithms

Model	Recall
Logistic Regression	0.68
KNN	0.93
Decision Tree	0.81
Random Forest Classifier	0.82
Support Vector Machine	0.70
LGBM using With Grid Search CV	0.80
Random Forest Classifier Using Randomised Search CV	0.85

	Model	Train Accuracy	Test Accuracy	Train F1-score	Test F1-score	Train Precision	Test Precision	Train Recall	Test
0	Logistic Regresssion	0.669968	0.675875	0.672607	0.679417	0.667271	0.672072	0.678029	0
1	KNN	1.000000	0.883057	1.000000	0.888694	1.000000	0.847826	1.000000	0
2	Decision Tree	1.000000	0.818600	1.000000	0.818433	1.000000	0.819188	1.000000	0
3	Random Forest Classifier	1.000000	0.876611	1.000000	0.870406	1.000000	0.916497	1.000000	0
4	SVM	0.791340	0.787293	0.775964	0.767839	0.837694	0.845133	0.722708	0
5	LGBM With Grid Search CV	0.752418	0.766114	0.765437	0.775618	0.727197	0.745331	0.807923	0

Fig: Classification Metrics Comparison

5. Conclusion:

1. The models that could can be deployed according to our study is KNN with Accuracy-0.883, Precision-0.847, Recall-0.933.
2. Better model can be developed that can predict the risk of coronary heart disease.
3. With the help of the experts we can engineer an extensive amount of variable that could make our prediction more accurate.
4. Hence, we can say that Machine Learning can help save lives of many and help them to switch over to a healthy lifestyle to chop off any health related issue.

6. References:

1. Towards Data Science.
2. Tutorials point.
3. Analytics Vidhya.
4. Wikipedia.
5. Stackover flow.
6. machinelearningmastery.com