Cardiovascular Risk Prediction

Hark Pun,
Data Science Trainee,
AlmaBetter, Bangalore.

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 to assess the impact on two distinct classes on the performance metrics.

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 analyzing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

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 best that could predict the risk of coronary heart disease.

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.

2. 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)

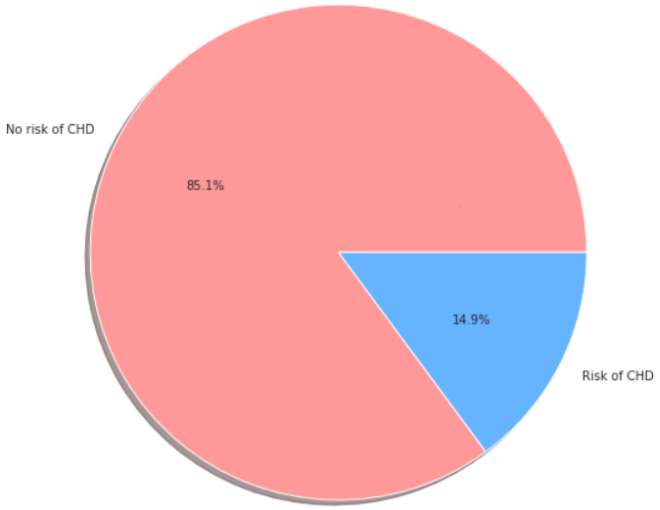
- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
- Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
- Diabetes: whether or not the patient had diabetes (Nominal)

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



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.

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.

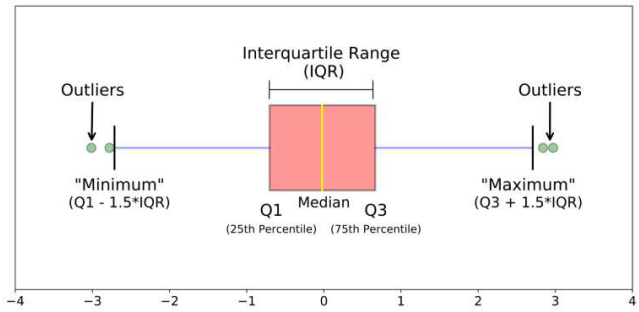
Let’s have a look at the missing values present in our data set.

```
# counting missing values
df.isna().sum()
```

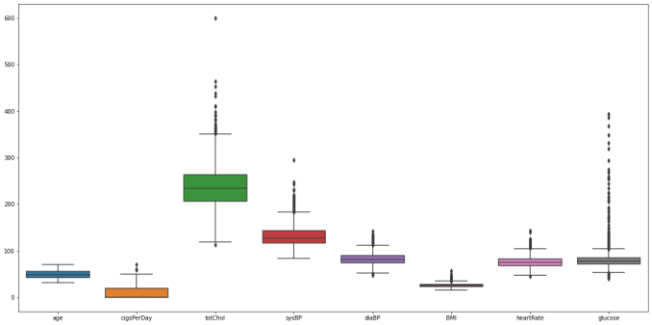
```
age                0
education         87
sex                0
is_smoking         0
cigsPerDay        22
BPMeds            44
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol           38
sysBP              0
diaBP              0
BMI               14
heartRate          1
glucose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

The cleaning of the data starts with the replacing of missing values using various techniques. Replace missing values in education, cigsPerDay, totalChol, BMI, glucose, and heartRate. Whichever feature having less than 5% missing values we have decided to drop them directly and rest will be imputed using mean, median & mode imputation method.

Further, outliers are removed from the data using the **Capping method**.

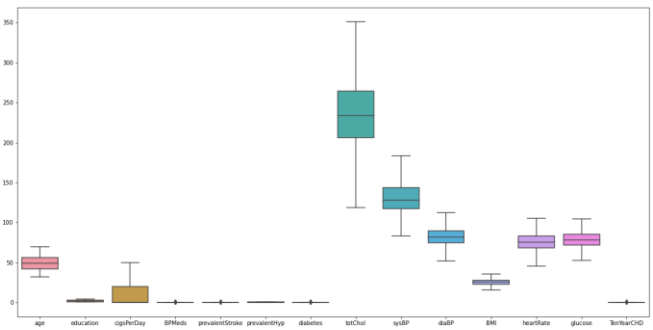


Where Q1, Q3 represent the first quantile, the third quantile of each attribute.



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 '

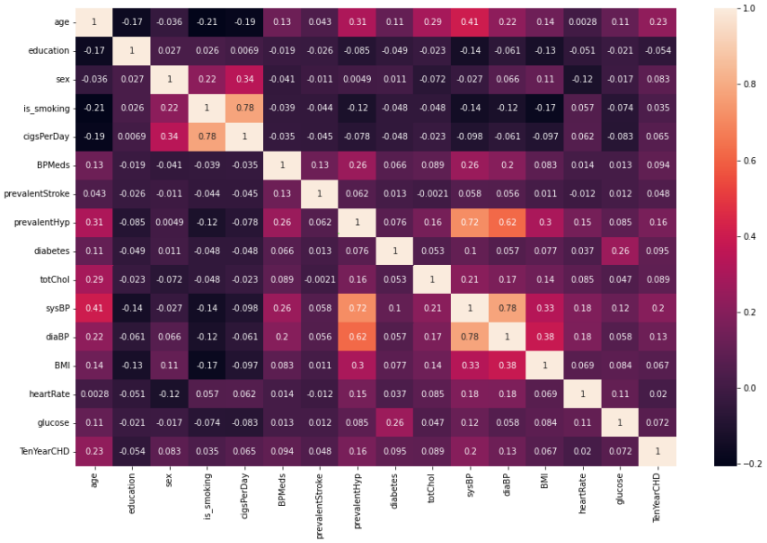
Therefore, there are no missing values in our dataset. Now, we will be looking at the treated outliers.



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.

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

From the above correlation matrix, we can see that:

- Highest correlation exists between is smoking and cigarette per day.
- Highest correlation exists between systolic BP and diastolic BP.
- Systolic and Diastolic BP shows a high correlation with hypertension.
- Variables such as age, prevalent hypertension, systolic BP, diastolic BP, influence the risk of heart disease mainly.
- All the variables have a positive correlation with the dependent variable, except for education.

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 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.
We used different types of Feature Selection methods.

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

In Healthcare industry, every single data is important to make prediction on target variable. In this particular case the dataset is related to medical domain, the entries in this dataset are person specific and the values vary among different individuals and all the features are very much important.
So, we are taking all features to train the model except multicollinearity features.

6. Treating 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.

SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors. Repeat the steps until data is balanced.

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. 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

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 Randomized Search CV, are used to predict heart disease.
For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, f1-score and roc_auc score are considered.

Accuracy- Accuracy is the ratio of the number of correct predictions to the total number of inputs in the dataset.

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

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

$$Precision = \frac{TP}{TP + FP}$$

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

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

F1score-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

Result:

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.

TABLE: Recall Comparison of algorithms

Model	Recall
Logistic Regression	0.681
Support Vector Machine	0.787
KNN	0.954
Decision Tree	0.852
Random Forest using RandomizedSearchCV	0.863
Adaptive Boosting	0.782
XGBoost using GridSearchCV	0.873
LGBM using GridSearchCV	0.866

	model	train_accuracy	test_accuracy	train_precision	test_precision	train_recall	test_recall	train_f1	test_f1	train_roc_auc	test_roc_auc
0	LogisticRegression	0.676	0.687	0.670	0.709	0.677	0.681	0.674	0.695	0.736	0.748
1	SVM	0.763	0.756	0.746	0.756	0.789	0.787	0.767	0.771	0.845	0.824
2	KNN	1.000	0.859	1.000	0.810	1.000	0.954	1.000	0.876	1.000	0.855
3	DecisionTree	1.000	0.825	1.000	0.824	1.000	0.847	1.000	0.835	1.000	0.824
4	RandomForest	0.975	0.876	0.989	0.897	0.959	0.861	0.974	0.879	0.998	0.952
5	AdaBoost	0.816	0.813	0.839	0.849	0.776	0.782	0.806	0.814	0.901	0.895
6	XGBoost	1.000	0.908	1.000	0.947	1.000	0.873	1.000	0.908	1.000	0.959
7	LightGBM	1.000	0.908	1.000	0.953	1.000	0.866	1.000	0.908	1.000	0.957

Fig: Classification Metrics Comparison

Conclusion:

In general, it is good practice to track multiple metrics when developing a machine learning model as each highlight's different aspects of model performance. However, we are dealing with Healthcare data and our data is imbalanced for that particular reason we are more focusing towards the Recall score and F1 score.

- We've noticed that XBG Classifier is the standout performer among all models with an f1-score of 0.908 and recall score of 0.873 on test data. it's safe to say that XGB Classifier provides an optimal solution to our problem.
- In case of Logistic regression, We were able to see the maximum f1-score of 0.695.
- For SVM(Support Vector Machines) Classifier, the f1-score lies around 0.774.
- KNN gave us Highest recall score of 0.954%.

- Out of the tree-based algorithms, LGBM Classifier and Random Forest Classifier was also providing an optimal solution towards achieving our Objective. We were able to achieve an f1-score of 0.908 and 0.884 respectively.

In the medical domain (**more focus towards the reducing False negative values, as we don't want to mis predict a person safe when he has the risk**) here the recall score is the most importance. KNN, XGB, LGBM, Random Forest gave the best recall score 0.954 ,0.873 ,0.866, 0.863 respectively.

Finally, we can **select Final model as KNN Classifier** because it has Highest Recall score and It is okay to classify a healthy person as having 10-year risk of coronary heart disease CHD (false positive) and following up with more medical tests, but it is definitely not okay to miss identifying a dieses patient or classifying a dieses patient as healthy (false negative).

References:

- https://towardsdatascience.com/
- https://www.analyticsvidhya.com/blog/
- https://github.com/rushter/data-science-blogs
- https://www.kaggle.com/
- https://stackoverflow.com/