**Cardiovascular Risk Prediction**

**By Harshavardhan Shete**

‘Lifestyle’ is one of the main components which determines our current as well as future health status. In today’s world, many non-communicable diseases like diabetes, hypertension, cardiovascular diseases, etc. have a close connection with our lifestyle. Thus, there are many kinds of gadgets which help you to get a healthy lifestyle like ‘Fitbands’ which tell you how many steps have you walked in a day, what is your heart rate, how many calories did you burn, etc. The data collected for all these parameters become an important source of information along with few other data points that can be utilized to check the risk of any of the diseases that have a close connection with lifestyle. And, once the risk is identified, people can make required changes in their lifestyle, can take necessary precautions, and can be saved.

**Problem Statement:**

Framingham Heart Study is an ongoing study on heart-related risks of residents of Framingham, Massachusetts. Here we are given with dataset having 3390 residents’ data about their demographic, behavioral components, medical history, and current medical data.

Our objective is to classify the resident weather he/she will have heart risk in the next ten years or not, with the help of gathered information. This is a binary classification task.

The dataset contains 3390 rows and 17 columns.

The following are the columns in the dataset:

• 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)

• is\_smoking: whether or not the patient is a current smoker ("YES" or "NO")

• 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.)

• 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)

• 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 the large number of possible values.)

• Glucose: glucose level (Continuous)

**Solutioning**

The solution of the problem involved the following steps:

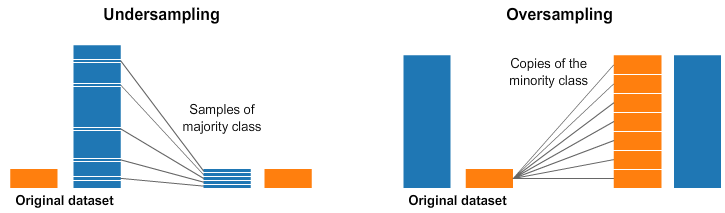
1. Exploratory Data Analysis
2. Handling the imbalanced dataset
3. Model training and performance comparison
4. **Exploratory Data Analysis**

Exploratory Data Analysis (EDA) as the name suggests, is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. It also helps to understand the relationship between the variables (if any) and it will be useful for feature engineering. It helps to understand data well before making any assumptions, to identify obvious errors, as well as better understand patterns within data, detect outliers, anomalous events, find interesting relations among the variables.

Exploration and visualizations performed with Covid-19 Tweet dataset:

1. Target variable class proportion
2. Age-wise count of people with CHD risk
3. Gender count and gender-wise risk prevalence
4. Education level and risk prevalence
5. Count of smokers and non-smokers, smoking habit vs. risk prevalence
6. Cigarettes per day and risk prevalence
7. Count of people with and without BP medicine, BP medicines vs. risk prevalence
8. Count of people with and without stroke history, stroke history vs. risk prevalence
9. Count of people with and without hypertension history, hypertension history vs. risk prevalence
10. Count of people with and without diabetes history, diabetes history vs. risk prevalence
11. Cholesterol level vs. CHD risk prevalence
12. Systolic and diastolic BP level vs. CHD risk prevalence
13. BMI vs. CHD risk prevalence
14. Heart Rate vs. CHD risk prevalence
15. Glucose level vs. CHD risk prevalence
16. Correlation matrix to understand the level of correlation of target variable & independent variables as well as correlation within independent variables
17. **Handling Imbalanced Dataset**

In the case of binary classification problems, if a disproportionately higher number of records belong to one class and another class has very fewer records then it is an example of an imbalanced dataset. Machine learning algorithms tend to produce unsatisfactory classifiers when faced with an imbalanced dataset. Thus, it is important to deal with imbalanced datasets before training models. Broadly the imbalanced datasets can be handled in two ways- algorithm level approach (using appropriate evaluation metrics, using k-fold cross-validation, designing an appropriate model) and data level approach (resampling techniques like random-under sampling, random oversampling, cluster oversampling, informed oversampling).



I have utilized oversampling technique to deal with an imbalanced dataset. In order to reduce overfitting during oversampling, synthetic sampling is used. SMOTE (Synthetic Minority Over Sampling Technique) is one of the popular algorithms which varies attributes of the observations to create a new synthetic sample.

1. **Model training and performance metrics**

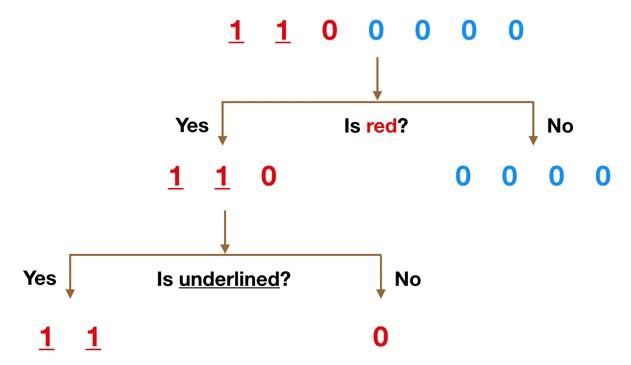
Various binary classification algorithms were tried with given data and they were compared using certain performance parameters to choose the best one. The algorithms tried are- Logistic regression, Decision tree classifier, Support vector machine classifier, K-nearest neighbor classifier and the performance parameters used were accuracy, precision, recall, and F1 score.

1. Logistic Regression:

Logistic regression is the first basic go-to algorithms when dealing with binary classification problems. It is an extension of the linear regression model for classification problems. Logistic Regression is a popular algorithm as it converts the values of the log of odds which can range from -inf to +inf to a range between 0 and 1. Since logistic functions output the probability of occurrence of an event, they can be applied to many real-life scenarios therefore these models are very popular.

1. Decision tree classifier

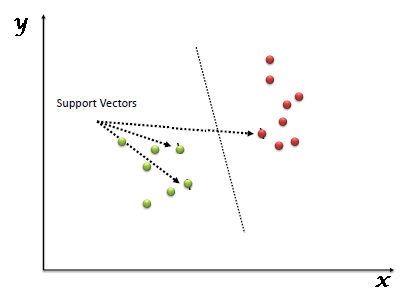
A decision tree classifier is a systematic approach for binary classification. It poses a set of questions to the dataset (related to its attributes/features). The decision tree classification algorithm can be visualized on a binary tree. On the root and each of the internal nodes, a question is posed and the data on that node is further split into separate records that have different characteristics. The leaves of the tree refer to the classes in which the dataset is split.



An image depicting the functioning of a Decision tree classifier.

1. Support vector machine classifier

“Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).



Support vectors belonging to two different classes.

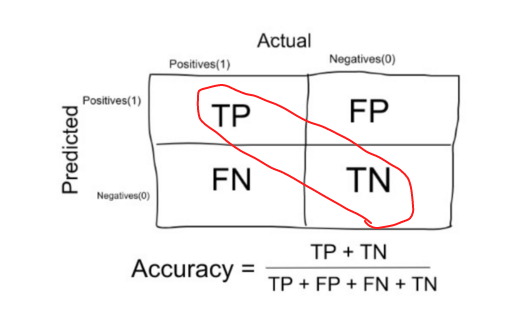
1. K-nearest neighbor classifier

KNN or k-nearest neighbors is the simplest classification algorithm. This classification algorithm does not depend on the structure of the data. Whenever a new example is encountered, its k nearest neighbors from the training data are examined. Distance between two examples can be the Euclidean distance between their feature vectors. The majority class among the k nearest neighbors is taken to be the class for the encountered example.

**Performance Metrics:**

1. Accuracy:

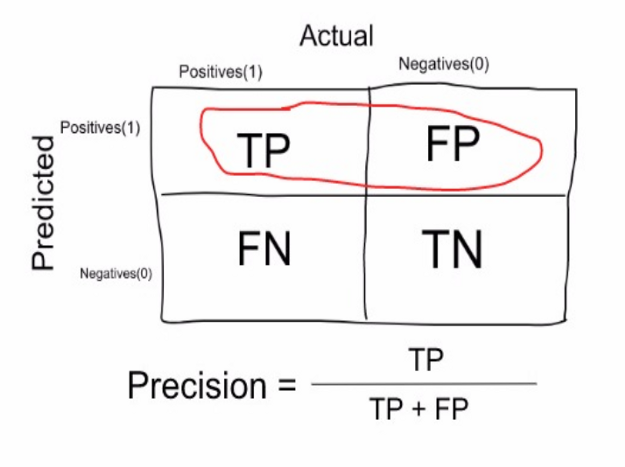
Accuracy in classification problems is the number of correct predictions made by the model over all kinds of predictions made. Accuracy is a good measure when the target variable classes in the data are nearly balanced. Accuracy should never be used as a measure when the target variable classes in the data are a majority of one class.



Accuracy calculation

1. Precision:

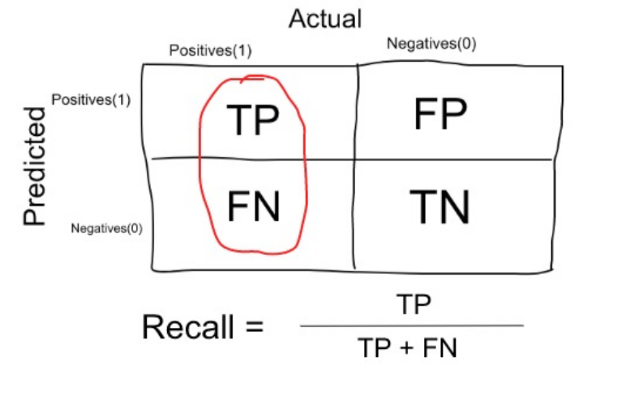
It is the number of correct positive results divided by the number of positive results predicted by the classifier.



Precision Calculation

1. Recall or Sensitivity:

It is the number of correct positive results divided by the number of *all*relevant samples (all samples that should have been identified as positive).



Recall calculation

1. F1 Score:

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

F1 Score= 2\* (Precision\*Recall)/ (Precision + Recall)

**Final verdict:**

**After comparing the results of various classifiers for the given classification problem, it is concluded that K Nearest Neighbour Classifier has performed best with 0.87 accuracy amongst all the models tried.**

**Sources**

* Towardsdatascience.com
* Ibm.com
* Analyticsvidhya.com
* Geeksforgeeks.org