**Cardiovascular Risk Prediction**

**Sohail Mohammad, Laveti Syam Sagar, Bisetty Kodanda Nagesh,**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

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 has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients’ information. It includes over 4,000 records and 15 attributes.

Our explore can offer assistance get it what may well be the reason for the classification of such names by highlight choice, information

analysis and forecast with machine learning algorithms taking under consideration past patterns to decide the proper classification.

**1.Problem Statement**

The dataset includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

**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)

**Behavioral:**

**• 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.)

**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, are considered continuous because of the large number of possible values.)

**• Glucose:** glucose level (Continuous) Predict variable (desired target)

**Ten Year CHD:** 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - DV

**2. Data Cleaning**

Prior to EDA, cleaning the data is essential since it will get rid of any ambiguous information that can have an impact on the results.

Education, cigs Per Day, BP Meds, tot Chol, BMI, heart Rate and glucose columns have missing or null values. I have filled these columns for missing values by using KNN imputer.

For columns sex and is smoking, I have changed the categorical value with a numerical value(like yes to 1 and no to 0, and F to 1 and M to 0).

## **3. Exploratory Data Analysis**

Most people smoke between 0 and 10 cigarettes a day.

The risk of coronary heart disease rises with age up to age 63 and then declines beyond that. Men have a better affinity for creating coronary heart malady (CHD) than ladies. Women in differentiate are at a better hazard of stork which regularly happens at more seasoned age. Compared to nonsmokers, smokers have a higher chance of developing coronary heart disease.

## **4. Feature Engineering**

**Impact of heart rate on the target variable**:

People with high heart rates have a high risk of having CHD in the next 10 years.

**Impact of cholesterol level on the target variable**:

People with high cholesterol levels have a high risk of having CHD in the next 10 years.

**Impact of age on the target variable**:

Older people have a high risk of having CHD in the coming 10 years.

**Impact of Body Mass Index on the target variable**:

People with high body mass index have a high risk of having CHD in the next 10 years.

**Impact of systolic blood pressure on the target variable:**

People with high systolic blood pressure have a high risk of having CHD in the next 10 years.

**Impact of Glucose on the target variable:**

People with high Glucose have a high risk of having CHD in the next 10 years.

**5. Data preparation**

The high correlation between :

# 1. Cigs Per Day and is smoking

# 2. Sys BP and Prevalent Hyp

3. Dia BP and Sys BP

Combined Sys Bp and Dia BP to denote a new feature pulse rate.

Dropping Cigs Per Day, is smoking, Sys BP, Prevalent Hyp, and Dia BP columns from the dataset.

Added the columns like pulse pressure, age bucket, and BMI bucket.

The dependent column will be predicted as that is the target variable named “Ten Year CHD.

**6. Steps involved:**

* **Exploratory Data Analysis**

After stacking the dataset we performed this strategy by comparing our target variable that's 10-year risk of future coronary heart infection (CHD) with other autonomous factors. This prepare made a difference us figuring out different perspectives and connections among the target and the autonomous factors. It gave us a better idea of which feature behaves in which manner compared to the target variable.

* **Null values Treatment**

Our dataset contains a huge number of invalid values which might tend to irritate our precision consequently utilized a KNN-imputer to perform lost esteem ascription, and prepared information to evacuate exceptions.

* **Encoding of categorical columns**

We utilized One Hot Encoding to deliver double integrability of and 1 to encode our categorical highlights since categorical highlights that are in string arrange cannot be caught on by the machine and ought to be changed over to numerical organize.

* **Feature Selection**

In these steps we utilized calculations like Extra Tree classifier to check the comes about of each include i.e which include is more vital compared to our demonstrate and which is of less importance.

Next we utilized Chi2 for categorical highlights and ANOVA for numerical highlights to choose the most excellent include which we'll be utilizing advance in our model.

* **Standardization of features**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

* **Fitting different models**

For modelling we tried various classification algorithms like:

1. **Logistic Regression**
2. **SVM Classifier**
3. **Random Forest Classifier**
4. **Grid Search CV**

* **SMOTE boosting for features**

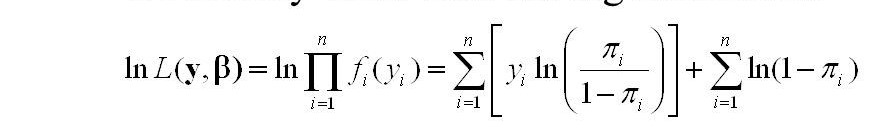
To solve the issue of class imbalance, SMOTE boosting was used to over-sample the minority class observations.

**7.1. Algorithms:**

1. **Logistic Regression:**

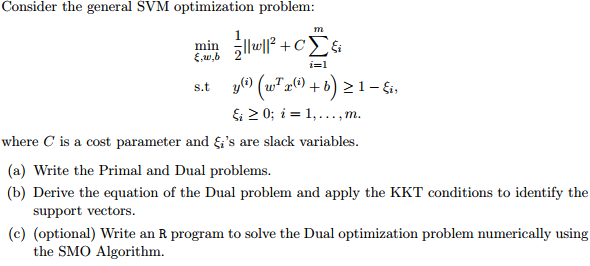
Calculated Relapse is really a classification calculation that was given the title relapse due to the truth that the scientific definition is exceptionally comparable to straight regression. The work utilized in Calculated Relapse is sigmoid work or the calculated work given by: f(x)= 1/1+e ^(-x)

The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic:



.

1. **Support Vector Machine Classifier:**

SVM is utilized for the most part when the information cannot be directly isolated by calculated relapse and the information has clamor. This may be done by isolating the data with a hyperplane at the next arrange dimension. In SVM we utilize the optimization calculation as: 

We use hinge loss to deal with the noise when the data isn’t linearly separable. Kernel functions can be used to map data to higher dimensions when there is inherent non linearity.

1. **Random Forest Classifier:**

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.

1. **Decision Tree**:

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

1. **Grid Search CV**

Grid Search CV is the method of performing hyperparameter tuning in arrange to decide the ideal values for a given show. As said over, the execution of a show altogether depends on the esteem of hyperparameters. Note that there's no way to know in development the leading values for hyperparameters so in a perfect world, we have to be attempt all conceivable values to know the ideal values. Doing this physically seem take a significant sum of time and assets and hence we utilize Grid Search CV to mechanize the tuning of hyperparameters.

**7.2. Model performance:**

Model can be evaluated by various metrics such as:

1. **Confusion Matrix**-

The disarray framework may be a table that summarizes how fruitful the classification model is at foreseeing illustrations having a place to different classes. One pivot of the disarray framework is the name that the demonstrate anticipated, and the other pivot is the real name.

1. **Precision/Recall**-

Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

1. **Accuracy**-

Accuracy is given by the number of correctly classified examples divided by the total number

of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

1. **F1 score-**

**F1-score is a harmonic mean of Precision and Recall**, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall.

**7.3. Hyper parameter tuning:**

Hyperparameters are sets of data that are utilized to control the way of learning an calculation. Their definitions affect parameters of the models, seen as a way of learning, alter from the unused hyperparameters. This set of values influences execution, solidness and elucidation of a show. Each calculation requires a particular hyperparameters lattice that can be balanced agreeing to the trade issue. Hyperparameters change the way a show learns to trigger this preparing calculation after parameters to produce outputs.

We used Grid Search CV, Randomized Look CV and Bayesian Optimization for hyperparameter tuning. This moreover comes about in cross approval and in our case we separated the dataset into distinctive folds. The finest execution enhancement among the three was by Bayesian Optimization.

1. **Grid Search CV-** Grid Search combines a determination of hyperparameters built up by the researcher and runs through all of them to assess the model’s execution. Its advantage is that it may be a basic strategy that will go through all the modified combinations. The greatest impediment is that it navigates a particular locale of the parameter space and cannot get it which development or which locale of the space is vital to optimize the demonstrate.
2. **Randomized Search CV-** In Random Look, the hyperparameters are chosen at random inside a run of values that it can expect. The advantage of this strategy is that there's a more prominent chance of finding locales of the fetched minimization space with more reasonable hyperparameters, since the choice for each cycle is irregular. The drawback of this strategy is that the combination of hyperparameters is past the scientist’s control.
3. **Bayesian Optimization-** Bayesian Hyperparameter optimization may be a exceptionally effective and curiously way to discover great hyperparameters. In this approach, in gullible elucidation way is to use a bolster show to discover the most excellent hyperparameters. A hyperparameter optimization handle based on a probabilistic model, regularly Gaussian Handle, will be utilized to discover information from information watched within the afterward conveyance of the execution of the given models or set of tried hyperparameters.

As it is a Bayesian process at each iteration, the conveyance of the model’s execution in connection to the hyperparameters utilized is assessed and a new probability conveyance is produced. With this dissemination it is conceivable to create a more suitable choice of the set of values that we are going utilize so that our calculation learns within the best conceivable way.

**8. Conclusion:**

Since our point was to lower the false-negative esteem so that patients don't get identified disgracefully and are illustrated to be secure, I utilized the review score as the assessment lattice. The patient's wellbeing may endure significantly as a result of this.

Data were resampled since they weren't adjusted. Tall precision can be accomplished with imbalanced information, be that as it may, in these circumstances, review, accuracy, and F1 score must be considered.

Used a KNN-imputer to perform missing value imputation and handled information to evacuate exceptions. To fathom the issue of lesson awkwardness, Destroyed boosting was utilized to over-sample the minority course perceptions.

Newer elements like pulse pressure, age bucket, and BMI bucket that helped to explain the separation in the Risk were created using the information from EDA.

Due to the parametric relationship within the information, a logistic regression show was implemented and it was successful in accomplishing a Review of 74.5%. Indeed in spite of the fact that the review score for SVM was 81.9 %, SVM isn't an interpretable show, hence I chose an interpretable show for this circumstance.

All measures, including Precision, Recall, Accuracy, and F1 score were evaluated for each model.

Based on this analysis, Logistic regression can identify positive cases with a 74.5% Recall. Using a decision tree, positive cases may be predicted with a review of 49%. With the assistance of Irregular Woodland, positive cases may be anticipated with a 49% Recall. Using a Back Vector Machine, positive cases can be anticipated with an 81.9% Recall. Using Framework Look CV, positive cases can be anticipated with 81.9% Recall.

**References-**

1. Machine Learning Mastery
2. Geeks for Geeks
3. Analytics Vidhya