**Cardiovascular Risk Prediction**

**Puneet Suthar**

**Data Science Enthusiast**

**AlmaBetter, Bangalore**

**Abstract:**

Cardiovascular diseases, also known as CVDs, are responsible for the highest number of fatalities worldwide, resulting in an estimated 17.9 million deaths annually. This group of disorders affects the heart and blood vessels, including coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other ailments. Heart attacks and strokes account for over 80% of all CVD-related deaths, and a significant portion of these fatalities occur prematurely in individuals under the age of 70. Unhealthy diet, physical inactivity, tobacco usage, and excessive alcohol consumption are the primary behavioral risk factors for heart disease and stroke. These risk factors may cause high blood pressure, high blood glucose, high blood lipids, overweight, and obesity in individuals.

1. **Problem Statement**

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 3,000 records and 17 attributes.

1. **Dataset 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)

1. **Steps involved**
   1. Performing EDA (exploratory data analysis).
   2. Observation and conclusions from the data .
   3. Getting the data ready for Model training.
   4. Model Selection by Evaluating metrics
   5. Training the model.
   6. Deployment of model as a web app.
2. **Performing EDA (exploratory data analysis)**
3. Exploring head and tail of the data to get insights on the given data.
4. Looking for null values, No. of zeros, duplicates, no. of unique in every column, it help us to make a guideline for feature engineering section before major EDA.
5. Dropping unwanted “id” column.
6. In Major EDA we performed below experimemts:
   1. Ploted a distribution plot of numeric data to get the idea of the skewness.
   2. Box plot to get a look on outliers.
   3. Regression plot to see the effect of feature on target column.
   4. Some bar plots between categorical columns and target column
   5. A correlation heat map to understand the correlation and multicollinearity.
   6. A pair plot to have a broad outlook to the data feature dependency.
7. **Observations and conclusions from the data** 
   1. There is a high imbalance in the dataset, in “TenYearCHD” column for 0 we have 2879 rows and for 1 we have 511 rows.
   2. The data has a significant number of null values, particularly in columns such as glucose, education, BPMeds, and total cholesterol. Since the dataset is person-specific and values vary between individuals, removing any rows containing null values is the most logical option. But already we don’t have much data, so dropping rows will not good for our dataset. Therefore we are using techniques like KNN Imputer to impute the null values, they may not be entirely accurate but preserve our data and help in better prediciton.
   3. Slightly more males are suffering from CHD than females.
   4. Both Women and Men lying in Age group of 50-52 have high risk of heart disease.
   5. Men lying in age group 40-42 are at risk.
   6. Men having age more than 65 are also at risk.
   7. Risk is High in same age group despite they are Smokers or not.
   8. The percentage of people who have CHD is almost equal between smokers and non smokers.
   9. No. of citrates per day is high among youngsters and it decreases as age increases but after 65 it increases again.
   10. The percentage of people who have CHD is higher among the diabetic, and those with prevalent hypertension as compared to those who don’t have similar morbidities.
   11. A larger percentage of the people who have CHD are on blood pressure medication.
   12. Another interesting trend I checked for was the distribution of the ages of the people who had CHD and the number of the sick generally increased with age with the peak being at 63 years old.
8. **Getting the data ready for training**
   1. Performed lable encoding in to columns (sex, is\_smoking, education ).
   2. Perform train test split.
   3. To deal with this imbalances we are using SMOTE on the training set. This is a type of data augmentation for the minority class and is referred to as the **Synthetic Minority Oversampling Technique**, or SMOTE for short. The approach is effective because new synthetic examples from the minority class are created that are plausible, that is, are relatively close in feature space to existing examples from the minority class.
   4. Now the data is ready for training.
9. **Model Selection by Evaluating metrics**
   1. Some models are chosen based on the data and the scatter plots.
   2. **Logistic Regression:**

Logistic regression is a type of generalized linear model used for binary classification. It models the log odds of the outcome as a linear function of predictor variables, and applies a logistic function to obtain the predicted probability of the positive class. It is trained using maximum likelihood estimation and evaluated using metrics such as accuracy and precision. It can also be extended to handle multi-class classification.

* 1. **Decision Tree Classifier:**

A decision tree classifier is a type of supervised learning algorithm used for classification tasks. It works by recursively partitioning the input space into regions, with each partition corresponding to a different class. The algorithm selects the features that best split the data based on some impurity criterion, such as entropy or Gini index. This process creates a tree-like structure, where each internal node represents a decision based on a feature, and each leaf node represents a class label. The quality of the tree is typically assessed using metrics such as accuracy, precision, recall, and F1 score. The main advantages of decision tree classifiers are their interpretability, computational efficiency, and ability to handle mixed data types. However, they can be prone to overfitting and instability in the presence of noise and outliers.

* 1. **Random Forest Classifier:**

A random forest classifier is an ensemble learning method that combines multiple decision trees to improve classification performance. It works by constructing a set of decision trees based on random subsets of the input data and features, and then aggregating their predictions to obtain a final classification. The individual trees are trained using bootstrapped samples of the data, and each node of each tree is split based on the best feature among a random subset of features. The quality of the forest is typically evaluated using metrics such as accuracy, precision, recall, and F1 score. Random forests have several advantages over individual decision trees, including reduced overfitting, improved generalization, and robustness to noise and outliers. However, they can be computationally expensive and less interpretable than single decision trees.

* 1. **K-Nearest Neighbors:**

K-Nearest Neighbors (KNN) is a supervised learning algorithm used for both classification and regression tasks. It works by finding the k nearest training examples to a new input based on some distance metric, and then predicting the class or value based on a majority vote or weighted average of their labels. The optimal value of k is usually chosen based on cross-validation or other optimization techniques. KNN is a simple and versatile algorithm that can handle non-linear decision boundaries and noisy data. However, it can be sensitive to the choice of distance metric and the curse of dimensionality, and can be computationally expensive for large datasets.

* 1. **Support Vector Machines:**

Support Vector Machines (SVM) is a supervised learning algorithm that seeks to find the best hyperplane that separates the input data into different classes. The algorithm works by maximizing the margin between the hyperplane and the nearest data points, and transforming the input data into a higher-dimensional space using a kernel function to enable non-linear separation. The quality of the SVM is typically evaluated using metrics such as accuracy, precision, recall, and F1 score. SVMs have several advantages over other classification methods, including the ability to handle high-dimensional data, non-linear decision boundaries, and outlier detection. However, they can be sensitive to the choice of kernel function and regularization parameter, and can be computationally expensive for large datasets.

* 1. To find the best suitable model , I defined the a function ModelSelection() which takes a input of x and y data , list of models to be checked, a dictionary contains the hyperparameter, no. of splits, and a scoring method. It gives us the best hyperparameters for the given models and highest precision score for a quick comparison.
  2. But single metric is not enough to get a best model, so a new function is defined Common\_Classification\_metrics() it takes input as a list of models, feature and target data. And gives a common confusion matrix, a common classification report showing metrics (precision, recall, f1-score, support) for all target classes, for all the listed models. And also highlights the min and max values of each section, a ROC curve for all models.
  3. Evaluating metrics :

1. **Accuracy:**

The proportion of correctly classified instances out of the total number of instances. It is a simple and intuitive metric that is widely used to evaluate classification performance. However, it can be misleading in imbalanced datasets where one class dominates the other.

1. **Precision:** The proportion of true positives (correctly predicted positives) out of the total number of predicted positives. It measures the ability of the classifier to avoid false positives, i.e., to correctly identify positive cases without mistakenly labeling negative cases as positive. Precision is important in applications where false positives are costly, such as medical diagnosis or fraud detection.
2. **Recall:** The proportion of true positives out of the total number of actual positives. It measures the ability of the classifier to detect all positive cases, including those that are missed (false negatives). Recall is important in applications where false negatives are costly, such as disease diagnosis or spam filtering.
3. **F1 score:** The harmonic mean of precision and recall, which balances the trade-off between them. It is a popular metric for imbalanced datasets where both precision and recall are important. F1 score is a good overall measure of classification performance and is commonly used to compare different classifiers.
4. **AUC ROC score :**The area under the ROC curve, which is a measure of the classifier's ability to distinguish between positive and negative cases. The AUC ROC score ranges from 0 to 1, where a score of 0.5 corresponds to random guessing, and a score of 1 corresponds to perfect classification. Higher AUC ROC scores indicate better performance, as the classifier is able to achieve higher TPR (True Positive Rate) at lower FPR (False Positive Rate).

The AUC ROC score is useful for evaluating the overall performance of a classifier, especially when the class distribution is imbalanced or the cost of false positives and false negatives is not equal. A high AUC ROC score indicates that the classifier is able to achieve high TPR while maintaining low FPR, which is desirable in our application.

1. **Training the best selected model**
2. We have made significant progress in our machine learning project by training 6 different models on the training dataset. Each of these models has been refined through hyperparameter tuning to enhance its performance and achieve the best possible results. Our objective was to build accurate predictive models that can predict the cardiovascular heart disease in next 10 years based on the patient data.
3. After training and testing each of the models, we found that The Lowest false negatives (Type 2 errors ) are predicted by Logistic Regression model that is 53.
4. Highest precision score for class 0 is 0.32 which is given by Random forest classifier, but it have higher type 2 error.
5. Highest Recall score for class 0 is 0.365 which is given by Logistic regression model. And it have lowest type 2 error also.
6. Logistic regression have highest AUROC 0.708
7. So, we can say that based on the above metrics The logistic regression would be the best model for our use case.
8. We trained our logistic regression model on the training dataset.
9. **Deployment of Selected Model as a web app.**

This is a Python script for a 10-year heart disease prediction application using the Streamlit framework. The application allows users to input personal data such as age, gender, education, smoking habits, blood pressure, and cholesterol levels, and then the application predicts the likelihood of developing heart disease in the next 10 years based on the given data.

The script imports the required libraries and defines a function to calculate the user's age. The user's data is collected through a form, and the input is stored in a Pandas dataframe. The script then applies some preprocessing steps on the data, such as converting categorical variables to numerical values and standardizing the data, and loads the trained machine learning model from a pickle file. Finally, the application predicts the likelihood of heart disease and displays the result to the user.

The user can trigger the prediction by clicking on the "Predict" button, and the application will display a message indicating the likelihood of heart disease in the next 10 years. If the likelihood is high, the application will display a warning message. The application also includes some visualization elements such as balloons popping up in the case of a low likelihood of heart disease.

**5. Conclusion**

1. The dataset has a high class imbalance, with significantly fewer cases of CHD than non-CHD cases.
2. The dataset contains null values, which were imputed using techniques like KNN imputation.
3. Men are more likely to suffer from CHD than women, and risk increases with age. Both men and women aged 50-52 have a higher risk, and men aged 40-42 and over 65 are also at risk.
4. Smoking does not appear to significantly affect CHD risk, but having diabetes or hypertension does increase the risk.
5. The percentage of people with CHD is almost equal between smokers and non-smokers.
6. The logistic regression model had the best performance in terms of minimizing false negatives (Type 2 errors), precision, and recall, with an AUROC of 0.708.
7. Based on these metrics, the logistic regression model would be the best choice for predicting CHD in this dataset.
8. Overall, our project has succeeded in developing and refining machine learning models that can make accurate predictions about cardiovascular heart disease risk prediction. These models can help doctors to get idea of heart disease for every individual patient.

**References**

1. Towards Data Science
2. Stack overflow
3. Medium