CAPSTONE PROJECT
(Machine Learning)
Cardiovascular-Risk-Prediction

Auti Rupali
Data science trainee,
AlmaBetter, Bangalore

Abstract:

We always deserve the best health service from our society. However, most of the people from different parts of the world are deprived of the same due to a lack of doctors and other medical staff in quality and quantity and in many cases medical affairs are costly. Artificial intelligence has a great opportunity in the health sector to create ml models which can provide the perfect solution at a low cost to everyone.

Our model for Cardiovascular Risk Prediction was a small attempt from a toy dataset towards reaching the goal.

1. Problem Statement

We were provided with a labeled dataset on the details of patients with or without cardiovascular disease.

Our task was to explore and analyze the data and to build a classification model for 10 Years Coronary Heart Disease prediction.

2. Introduction

We are at a stage where a big revolution in the health sector is expected. Heart disease is one of the most important cases for human health.

In this project, we tried to create the best classification model for Coronary Heart Disease prediction from the limited available experience.

3. Steps Involved

Following steps were involved during this Supervised ML (Classification) Project

3.1. **Connection with the Data**

The dataset was actually a collection of 3,390 experiences about patients with or without cardiovascular disease with 17 features or dimensions.

We needed to decode the set of experiences to build a model for Cardiovascular Risk Prediction.

At first, we imported the libraries or functions for making our journey easy and then got connected to the set of experiences.

3.2. **First Feelings of the Data**

When we saw the head of the data, we could understand what the set of experiences was all about. Then we tried to understand the features of the experiences.

3.3. **Deeper Understanding of the Data**

As there was a huge no. of experiences, we took the help of statistics to measure each and every feature in different dimensions, and thus step by step, we found the most important features or the exact way to decode the experiences.

“what gets measured gets done“.

3.4. **Cleaning the Data**

We handled all null values in 'cigsPerDay', 'totChol', 'BMI', 'heartRate', 'glucose', 'education', and 'BPMeds' columns with imputation. Thus there was no loss of data.

We encoded ‘sex’ column with two categories: F: 0, M: 1

We encoded ‘is_smoking’ column with two categories: NO:0, YES:1

We also checked the statistics several times on clean data to confirm the completion of processing.

3.5. **Treating Anomalies in the Data**

While we were finding out the general formula from the experiences, we were supposed to identify the true outliers or exceptional or abnormal experiences and keep them aside.

For most of the features, class 1 targets were outliers. Thus we needed more experience with class 1 targets to bring a balance in prediction.

In simple words, with the available experience set, our model was going to become an expert in predicting the features for which there will be no heart disease.

3.6. **Final Feature Selection from the Data**

We needed to understand the distribution of the features and the relationship among the features for the decision of transformation, scaling, and final selection of features.

Here, the distribution of 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate' and 'glucose' were positively skewed. Thus we did log transformation on these features to normalize their distribution.

We removed ‘sysBP’, ‘diaBP’, ‘BMI’, ‘Age’, ‘heartRate’, ‘totChol’, ‘is_smoking’ and ‘Diabetes’ in sequence from our dataset to bring all the VIF values below 10. Thus all our input variables became truly independent.

3.7 **Preparation of Input and Output Data**

Finally, we prepared the inputs (X) and output (y) for our model in three steps:

i. Normalization (Log transformation),

ii. Train-Test Splitting and

iii. Scaling

3.8 **Building and Evaluation of Model-1**

In our final Random Forest Model, after cross-validation and hyperparameter tuning, the best parameters were 'max_depth': 25,

'max_features': 'auto', 'min_samples_leaf': 5, 'min_samples_split': 15, 'n_estimators': 100

3.9 **Building and Evaluation of Model-2**

In our final KNN Model, after cross-validation and hyperparameter tuning, the best parameters were 'leaf_size': 30, 'n_neighbors': 19

3.10 **Building and Evaluation of Model-3**

In our final SVC Model, after cross-validation and hyperparameter tuning, the best parameters were 'C': 6, 'gamma': 0.1

4. Challenges Faced

4.1 The first challenge was to find the relevant features as most of the features were looking important as a first feeling.

4.2 It was very challenging to find maximum truly independent features as there were many features with multicollinearity and VIF values above 10.

4.6 The most challenging work was to create a dashboard for comparing model performance.

5. Approach Used

The performance of a machine learning model depends on three factors:

5.1 Quality of Data- cleaner experiences for better learning: We gave the highest importance to the exploration and pre-processing of data to produce quality data.

5.2 Quantity of Data-more experiences for better learning: We tried to minimize the loss of data to the extent it is possible.

5.3 Quality of Model-right model and right hyperparameters for better learning: We selected the model considering the volume of clean data and type of expected output. We also tuned the hyperparameters to produce the optimum model.

6. Conclusion

Conclusions drawn were as follows:

6.1 In our final Random Forest model, test accuracy is 84% and variation in performance is 2%

6.2 In our final KNN model, test accuracy is 84% and variation in performance is 1%

6.3 In our final SVC model, test accuracy is 84% and variation in performance is 2%

6.4 On the basis of the performance study of our three models, we selected KNN classifier for predicting 10 Years Coronary heart disease, as it had a low variation in performance, good f1_score, and good ROC_AUC score among all three models