Capstone Project Submission

Team Member's Name, Email and Contribution:

Name: Sethupathy M
Email: sethuqr@gmail.com
Contribution:

- Data cleaning
- Treating NULL values
- Box plot to detect outliers
- Bar graph for avg of independent variables with and without risk of CHD
- Multicollinearity
- Finding class Imbalance in given dataset
- Handling class imbalance by various sampling techniques (Feature engineering)
- Feature selection
- Feature transformation (Standardization)
- Logistic Regressor (hyperparameter tunning)
- Choosing appropriate evaluation metrics
- Plotting Precision Recall AUC curve
- Visual representation of confusion matrix
- Random Forest Classifier (with cv and hyperparameter tunning)

Name: Sri harish A
Email: sriharishanand@gmail.com
Contribution:

- Z score technique to remove outliers (outlier treatment)
- Distribution plot for numeric variables
- Count plot for categorical variables
- Pie chart
- Line plot for cigarettes consumed
- Variance inflation factor
- Feature transformation (Standardization)
- Logistic Regressor (hyperparameter tunning)
- Gradient Boost Classifier (with cv and hyperparameter tunning)
- XGB Classifier (with hyperparameter tunning)
- KNNeighbors Classifier (with hyperparameter tunning)
- Gaussian Naïve Bayes
- Model performance analysis
- Model performance visual representation

GitHub Link: -
https://github.com/SethupathyM/Supervised_ML_Classification_Cardiovascular_Risk_Prediction

Supervised Machine Learning – – Cardiovascular Risk Prediction

Summary

This project contains the data record of Patients who could possibly may or not develop a risk of Cardiovascular disease in upcoming 10 years. The Dataset contains the details like id, sex, age, medical records etc. There are a total of 17 variables describing 3390 observations. Each observation represents whether the patient will or will not develop a risk of Cardiovascular disease in upcoming 10 years.

The purpose of this project is to try a machine learning approach for Cardiovascular Risk Prediction by given the id, sex, and information about the health condition of the person. This project contains: Exploratory data analysis, feature engineering, choosing appropriate features, cross algorithms, cross validation, tuning the algorithms, analysis of feature importance, analysis of model performance. The predictions of future could help for saving a person's life and could help them to get rid of unhealthy habits. Another point of view is to test the machine learning algorithms how good are at solving this problem

We began by loading the data set from the google drive into our colab notebook, which was in.csv format, and performing basic operations such as shape, describe and info, is null, and so on to gain a basic understanding of the dataset's contents.

Data cleaning was the next process to remove unwanted variables and values from our dataset and get rid of any irregularities in it. Since these values disproportionately skew the data and hence adversely affect the results.

Our dataset had many outliers which could affect the effectiveness of the model. So, by using bar plot outliers had been detected and by using Z- score technique all the outliers had been imputed with median.

Exploratory Data Analysis was performed on the dataset to understand the relationship between the target variable and independent variables.
From the EDA the hypothesis generation was done which was the Target variable is influenced by sex, age, education, Is_smoking, CigsPerDay, BPMeds,
prevalentStroke, prevalentHyp, Diabetes, Tot_Chol, Sys_BP, Dia_BP, BMI, Heart rate, Glucose.

As the given dataset is highly imbalanced dataset various sampling techniques were tried. Out of those Random over sampling is found to be more effective.

Standardization transformation was done to centers the values around mean with unit standard deviation.

Train_test_split was used to Split arrays or matrices into random train and test subsets.

The transformed dataset was fitted in Logistic Regressor model then appropriate metrics like ROC AUC scores, Precision Recall scores, Confusion matrix were chosen but the desired scores were not achieved.

Decision Tree Classifier is robust to outliers and multicollinearity but it is prone to overfitting.
We got 100% Precision Recall score for train dataset and 75% Precision Recall score for test dataset.

So, Ensemble of Decision Tree was chosen. In RandomForestClassifier number of tress are generated and mean of those trees are taken into account so overfitting is reduced. Desired Precision Recall score of 97.8% is achieved by using RandomForestClassifier.

And further Gradient Boost Classifier, XGB Classifier, KNNeighbors Classifier were tried but none of the above-mentioned models gave the scores high as RandomForestClassifier.

Model Performance were analyzed in which the Precision Recall score, ROC AUC scores of the different models were compared with visualization