|  |  |
| --- | --- |
| **Roll Nos:** | 2845  2848  2852  2854 |
| **Student’s Name:** | Ojasvi Tummala  Aishwarya Pawar  Preeti Raut  Aparajita Sarkar |
| **Title:** | Open Ended Assignment - Disease Prediction |
| **Date:** | 30/05/22 |

**Open Ended Assignment - Disease Prediction**

**Assignment Statement:**

Build and evaluate the machine learning model for disease diagnosis from patients’ symptoms.

**Dataset used:**

The dataset is available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

1. Name: framingham.csv
2. Link: https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression

**Dataset description:**

This dataset contains records of individuals in the following format

* male: 0 = Female; 1 = Male
* age: Age of person
* education: 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = College
* currentSmoker: 0 = nonsmoker; 1 = smoker
* cigsPerDay: number of cigarettes smoked per day (estimated average)
* BPMeds: 0 = Not on Blood Pressure medications; 1 = Is on Blood Pressure medications
* prevalentStroke: 0 = Stroke not prevalent in family history; 1 = Stroke prevalent in family history
* prevalentHyp: 0 = Hypertension not prevalent in family history; 1 = Hypertension prevalent in family history
* diabetes: 0 = No; 1 = Yes
* totChol: total cholesterol (mg/dL)
* sysBP: systolic blood pressure (mmHg)
* diaBP: diastolic blood pressure (mmHg)
* BMI: BodyMass Index
* heartRate Beats/Min
* glucose: total glucose mg/dL
* TenYearCHD: 0 = Patient doesn’t have 10-year risk of future coronary heart disease; 1 = Patient has 10-year risk of future coronary heart disease

|  |  |
| --- | --- |
| **No. of Samples** | 4238 samples |
| **No. of Features (15)** | male, age, education, currentSmoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totchol, sysBP, diaBP, BMI, heartRate, glucose |
| **Target** | TenYearCHD |

**Model building:**

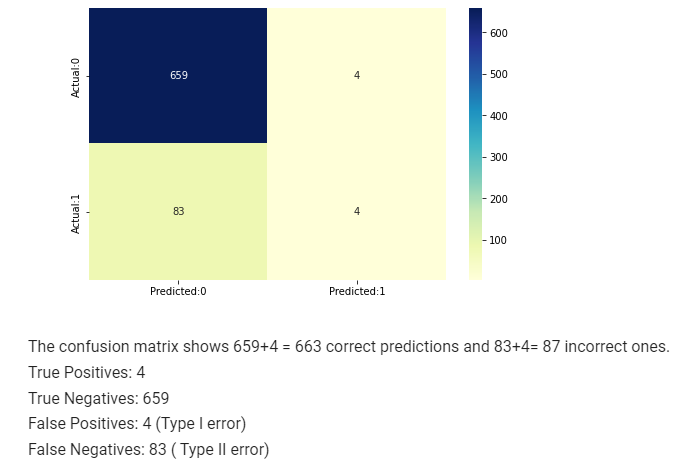
The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients’ information. Hence, we evaluated the dataset using different algorithms to find which one provides the best accuracy and suits our model.

**Algorithm:**

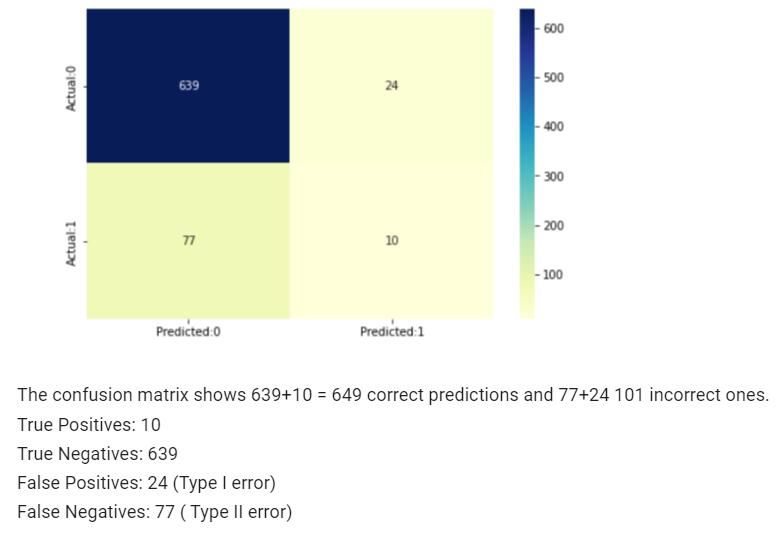
1. Import the necessary libraries into the workspace.
2. Add the csv file of the dataset and print the dataset.
3. From the dataset drop the education column and using .head() method print top 5 samples.
4. Using .isNull().sum() method , find the columns with null values. Since it is 12% of the entire dataset, the rows with missing values are excluded using drop method.
5. Visualize the data by plotting histogram for every feature.
6. Perform feature selection by plotting heatmap.
7. x stores the feature matrix which comprises of the columns: age, male, cigsPerDay, totChol, sysBP and glucose. Similarly, y is reponse vector which stores the target column: TenYearCHD.
8. Import LogisticRegression from sklearn.linear\_model and train the model using the training dataset using .fit(x\_train , y\_train) command . Find y1\_pred by applying it on test data (x\_test).
9. Model accuracy can be calculated by comparing values of y1\_pred and y\_test.
10. Plot the confusion matrix by importing it from library sklearn.metrics . The confusion matrix shows the number of correct and incorrect predictions done by the given model.
11. To evaluate the given model, calculate different evaluation parameters like accuracy, misclassification, precision, recall, sensitivity and specificity.
12. To choose the most accurate model, repeat steps 8-11 for other models like Naïve bayes and Decision Tree. Import these models from sklearn.

**Justifications:**

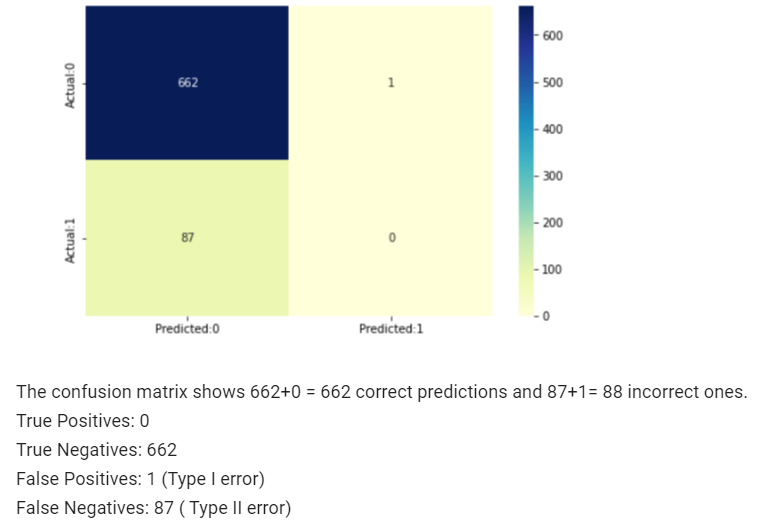
1. **Logistic Regression:**



1. **Naive Bayes Classifier:**

****

1. **Decision Tree:**



**Results obtained:**

**Evaluation parameters:**

1. **Logistic Regression:**
2. Accuracy = TP+TN/(TP+TN+FP+FN) = 0.884
3. Precision = TP/(TP+FP) = 0.5
4. Recall = TP/(TP+FN) = 0.045
5. F-Measure = (2\*Precision\*Recall)/(Precision+Recall) = 0.084
6. **Naive Bayes Classifier:**
7. Accuracy = TP+TN/(TP+TN+FP+FN) = 0.865
8. Precision = TP/(TP+FP) = 0.294
9. Recall = TP/(TP+FN) = 0.114
10. F-Measure = (2\*Precision\*Recall)/(Precision+Recall) = 0.165
11. **Decision Tree:**
12. Accuracy = TP+TN/(TP+TN+FP+FN) = 0.8826666666666667
13. Precision = TP/(TP+FP) = 0.0
14. Recall = TP/(TP+FN) = 0.0
15. F-Measure = (2\*Precision\*Recall)/(Precision+Recall) = nan

**Conclusion:**

* The LOGISTIC REGRESSION model predicted with 88.4% accuracy.

The model is more specific than sensitive, i.e., it predicts class 0 with much more accuracy than class 1.

* The NAIVE BAYES CLASSIFIER model predicted with 86.5% accuracy.

It has highest recall value among all the 3 models.

* The DECISION TREE model predicted with 88.2% accuracy.

The model is more specific than sensitive.

* Models 1 and 3 have poor performance for detecting positives, hence will be highly inefficient disease prediction models. In this application, it is better to have false positive (type I error) than false negative (type II error) as false negative could cost someone’s life. As model 2 is matching this pattern it is best fit for this application.
* Thus, overall best model for disease prediction is NAIVE BAYES CLASSIFIER.