Machine Learning

Course-End Project- Health Care

Problem statement:

Cardiovascular diseases are the leading cause of death globally. It is therefore necessary to identify the causes and develop a system to predict heart attacks in an effective manner. The data below has the information about the factors that might have an impact on cardiovascular health.

Task to be performed:

1. Preliminary analysis:
    a. Perform preliminary data inspection and report the findings on the structure of the data, missing values, duplicates, etc.
    b. Based on these findings, remove duplicates (if any) and treat missing values using an appropriate strategy

2. Prepare a report about the data explaining the distribution of the disease and the related factors using the steps listed below:
    a. Get a preliminary statistical summary of the data and explore the measures of central tendencies and spread of the data
    b. Identify the data variables which are categorical and describe and explore these variables using the appropriate tools, such as count plot
    c. Study the occurrence of CVD across the Age category
    d. Study the composition of all patients with respect to the Sex category
    e. Study if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient
    f. Describe the relationship between cholesterol levels and a target variable
    g. State what relationship exists between peak exercising and the occurrence of a heart attack
    h. Check if thalassemia is a major cause of CVD
    i. List how the other factors determine the occurrence of CVD
    j. Use a pair plot to understand the relationship between all the given variables
3. Build a baseline model to predict the risk of a heart attack using a logistic regression and random forest and explore the results while using correlation analysis and logistic regression (leveraging standard error and p-values from statsmodels) for feature selection

Steps followed to carry out this task

Imported the Dataset

Perform Basic data cleaning

Perform EDA

Perform Statistical Analysis

Model Building

Model Evaluation

Statistical Model Result:

Logistic Regression Model

```
Accuracy score of LR model is:  0.8524590163934426
Classification report of LR model is:
               precision    recall  f1-score   support

           0       0.96      0.76      0.85        33
           1       0.77      0.96      0.86        28

    accuracy                           0.85        61
   macro avg       0.87      0.86      0.85        61
weighted avg       0.87      0.85      0.85        61
```

SVM Model

```
Accuracy score of SV model is:  0.8524590163934426
Classification report of SV model is:
               precision    recall  f1-score   support

           0       1.00      0.73      0.84        33
           1       0.76      1.00      0.86        28

    accuracy                           0.85        61
   macro avg       0.88      0.86      0.85        61
weighted avg       0.89      0.85      0.85        61
```

Decision Tree Model

```
Accuracy score of DT model is:  0.819672131147541
Classification report of DT model is:
               precision    recall  f1-score   support

           0       0.89      0.76      0.82        33
           1       0.76      0.89      0.82        28

    accuracy                           0.82        61
   macro avg       0.83      0.83      0.82        61
weighted avg       0.83      0.82      0.82        61
```

Random Forest Model

```
Accuracy score of RF model is:  0.8688524590163934
Classification report of RF model is:
               precision    recall  f1-score   support

           0       1.00      0.76      0.86        33
           1       0.78      1.00      0.88        28

    accuracy                           0.87        61
   macro avg       0.89      0.88      0.87        61
weighted avg       0.90      0.87      0.87        61
```

Gradient Boosting Model

```
Accuracy score of GB model is:  0.8524590163934426
Classification report of GB model is:
               precision    recall  f1-score   support

           0       0.90      0.82      0.86        33
           1       0.81      0.89      0.85        28

    accuracy                           0.85        61
   macro avg       0.85      0.86      0.85        61
weighted avg       0.86      0.85      0.85        61
```

XGBOOST Model

```
Accuracy score of XG model is:  0.8360655737704918
Classification report of XG model is:
               precision    recall  f1-score   support

           0       0.93      0.76      0.83        33
           1       0.76      0.93      0.84        28

    accuracy                           0.84        61
   macro avg       0.85      0.84      0.84        61
weighted avg       0.85      0.84      0.84        61
```

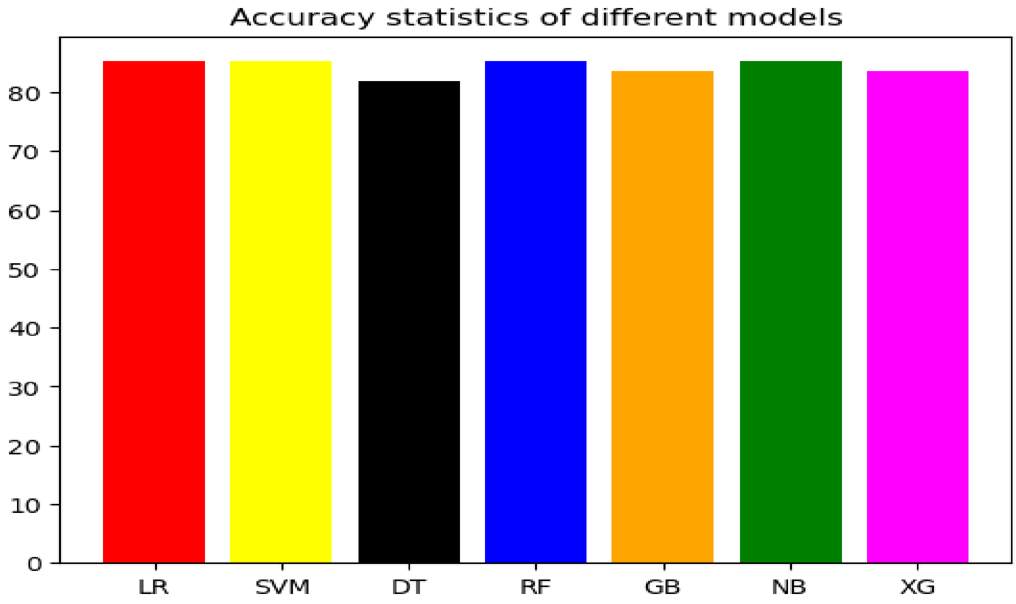
Naïve Bayes Model

```
Accuracy score of NB model is:  0.8524590163934426
Classification report of NB model is:
               precision    recall  f1-score   support

           0       0.90      0.82      0.86        33
           1       0.81      0.89      0.85        28

    accuracy                           0.85        61
   macro avg       0.85      0.86      0.85        61
weighted avg       0.86      0.85      0.85        61
```

CONCLUSION:

```
Accuracy Result of the models is as follows:
LR: 85.24590163934425
SVM: 85.24590163934425
DT: 81.9672131147541
RF: 85.24590163934425
GB: 83.60655737704919
NB: 85.24590163934425
XG: 83.60655737704919
```

The accuracy of the model from LR, SVM, RF and NB is maximum upto 85.25%



Thank You