Capstone Project-Healthcare

DESCRIPTION

NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.

- The dataset used in this project is originally from NIDDK. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
- Build a model to accurately predict whether the patients in the dataset have diabetes or not.
-

Dataset Description

The datasets consists of several medical predictor variables and one target variable (Outcome). Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and more.

Variables	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skinfold thickness (mm)
Insulin	Two hour serum insulin
BMI	Body Mass Index
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age in years
Outcome	Class variable (either 0 or 1). 268 of 768 values are 1, and the others are 0

Project Task: Week 1

Data Exploration:

1. Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make sense and thus indicates missing value:

- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI

2. Visually explore these variables using histograms. Treat the missing values accordingly.

3. There are integer and float data type variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

Data Exploration:

4. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.

5. Create scatter charts between the pair of variables to understand the relationships. Describe your findings.

6. Perform correlation analysis. Visually explore it using a heat map.

Project Task: Week 2

Data Modeling:

1. Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.
2. Apply an appropriate classification algorithm to build a model.
3. Compare various models with the results from KNN algorithm.
4. Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc.

Please be descriptive to explain what values of these parameter you have used.

Data Reporting:

5. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

- Pie chart to describe the diabetic or non-diabetic population
- Scatter charts between relevant variables to analyze the relationships
- Histogram or frequency charts to analyze the distribution of the data
- Heatmap of correlation analysis among the relevant variables
- Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.

Steps followed to carry out this task

Imported the Dataset

Perform Basic data cleaning

Perform EDA

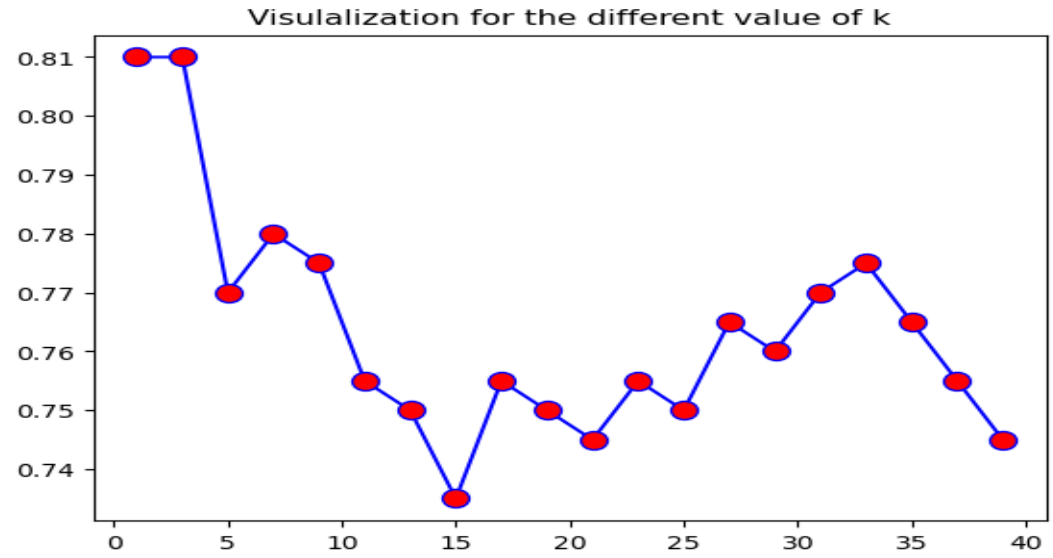
Perform Statistical Analysis

Model Building

Model Evaluation

Statistical Model Result:

KNN



K value=3
Accuracy score for KNN is : 0.81
Classification report for KNN is :

	precision	recall	f1-score	support
0	0.74	0.83	0.78	81
1	0.87	0.80	0.83	119
accuracy			0.81	200
macro avg	0.80	0.81	0.81	200
weighted avg	0.82	0.81	0.81	200

Logistic regression

Accuracy score for LR is : 0.765
Classification report for LR is :

	precision	recall	f1-score	support
0	0.80	0.72	0.76	102
1	0.73	0.82	0.77	98
accuracy			0.77	200
macro avg	0.77	0.77	0.76	200
weighted avg	0.77	0.77	0.76	200

SVM

Accuracy score for SVM is : 0.76
Classification report for SVM is :

	precision	recall	f1-score	support
0	0.80	0.71	0.75	103
1	0.72	0.81	0.77	97
accuracy			0.76	200
macro avg	0.76	0.76	0.76	200
weighted avg	0.76	0.76	0.76	200

Decision Tree

Accuracy score for Decision Tree is : 0.775
Classification report for Decision Tree is :

	precision	recall	f1-score	support
0	0.79	0.73	0.76	98
1	0.76	0.81	0.79	102
accuracy			0.78	200
macro avg	0.78	0.77	0.77	200
weighted avg	0.78	0.78	0.77	200

Random Forest

```
Accuracy score for Random Forest is : 0.83
Classification report for Random Forest is :
              precision    recall  f1-score   support

           0       0.86      0.79      0.82        99
           1       0.81      0.87      0.84       101

    accuracy                           0.83       200
   macro avg       0.83      0.83      0.83       200
weighted avg       0.83      0.83      0.83       200
```

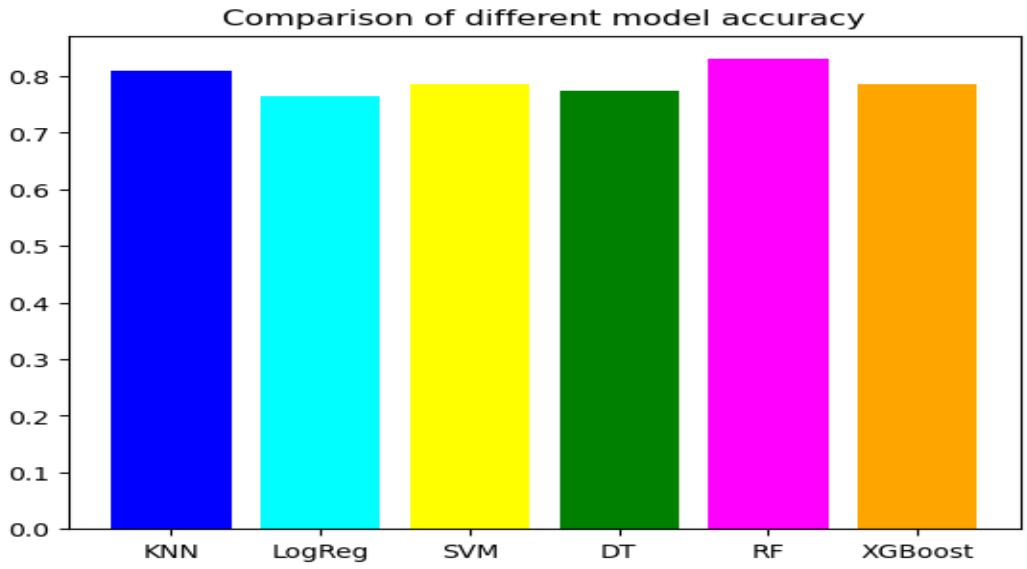
XGBOOST

```
Accuracy score for Xgboost is : 0.785
Classification report for Xgboost is :
              precision    recall  f1-score   support

           0       0.81      0.74      0.77       100
           1       0.76      0.83      0.79       100

    accuracy                           0.79       200
   macro avg       0.79      0.78      0.78       200
weighted avg       0.79      0.79      0.78       200
```

CONCLUSION:

Accuracy Result of the models is as follows:

'KNN': 0.81
'LogReg': 0.765
'SVM': 0.785
'DT': 0.775
'RF': 0.83
'XGBoost': 0.785



The accuracy of the model from Random Forest is maximum upto 83%