MAJOR PROJECT

BY: SWASTIK MOHAPATRA

Aim:

we have been given a “diabetes.csv” data file .This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

And hence we need to apply at minimum 2 classification algorithms and compare their respective accuracies.

Preview: this project classifying whether a person has diabetes or not considering the different data columns given in the csv file that where:

“pregnancies” , “glucose”, “blood pressure” , “skin thickness”, “ insulin” , “BMI”, ”diabetes pedigree function” and “age” .

We tend to find out the relation and apply 4 machine learning algorithms to the following dataset in order to predict their respective accuracies and hence point out the best prediction algorithm among them.

Tools used:

There are a number of tools used for completing this major project some of them being as follows :

Pandas – for implementing fataframe structures and using it to store the values of the given csv files (eg: pd.read_csv(“file_path”))

Matplotlib – for implementing graphical representation of the results as well as the analysis as it helps us to get a clear overview and errors in the data.

Sklearn – this library seems to help us a lot from data pre processing for generating usable data to dividing test and train data to implementing the 4 classification algorithms it helps us access pre defined classes for classifying data.

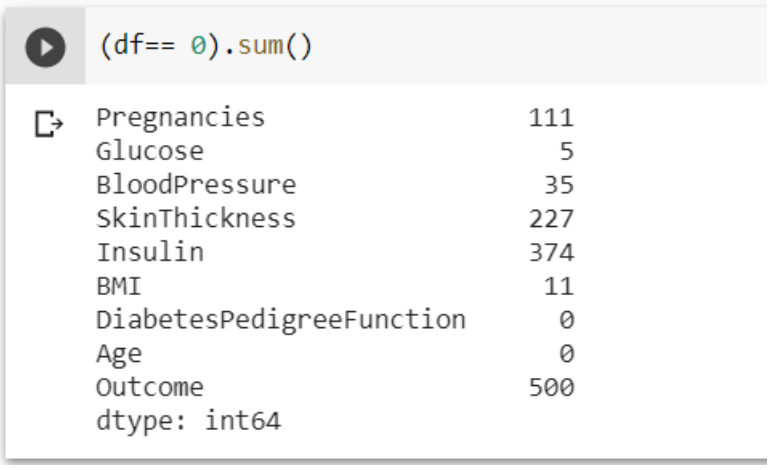
Python – is the programming language we have used to write the code implementing the machine learning concepts.

```
[ ]  import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     from sklearn.model_selection import train_test_split
     from sklearn.preprocessing import StandardScaler
     from sklearn.linear_model import LogisticRegression
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.ensemble import  RandomForestClassifier
     from sklearn.svm import SVC
     from sklearn.metrics import accuracy_score
```

Pre-processing and understanding:

We analyse the data before applying the algorithms to the dataset for machine learning. While analysing I plotted histograms to possibly guess important factors.

We also got abnormal values where BMI , glucose and skin thickness were 0 and insulin being 0 wasn’t an uncommon phenomenon .



```
[6] df_refined=df[(df.Glucose!=0) & (df.BloodPressure!=0) & (df.BMI!=0)]
    #data cleaning
    df_refined.shape

    (724, 9)
```

so we cleaned the data accordingly and then divided the it into train and test parts further scaling them relatively using StandardScaler.

```
[62] x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=0,test_size=0.33)
     sc=StandardScaler()

[63] x_train=sc.fit_transform(x_train)
     x_test=sc.fit_transform(x_test)
```

Algorithms :

In this project I have implemented 4 classification algorithm:

1. Logistic regression : the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. Basically it points the data points in a 2 dimensional graph and divides it into 2 classes like here being diabetic or not.
2. Decision tree : A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.it creates conditional clauses in a tree like format with a specific priority order helping us to classify the given dataset.
3. Random forest : Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

4. Support vector machines : In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyse data for classification and regression analysis. Here we have classified the data into 2 parts.

```
[61] lr=LogisticRegression(random_state=0)
     dct=DecisionTreeClassifier(random_state=0)
     rfc=RandomForestClassifier(random_state=0)
     sv=SVC(random_state=0,kernel='rbf')
     #TRIED CHANGING KERNEL VALUES BUT GOT EVEN LOWER ACCURACY
     x=df_refined.iloc[:,0:8].values
     y=df_refined.iloc[:,8].values
     print(x)
     print(y)
```

CONCLUSION :

We applied the 4 classification algorithms for the given diabetes dataset and found their respective accuracies which are :

Logistic regression : 78.24 %

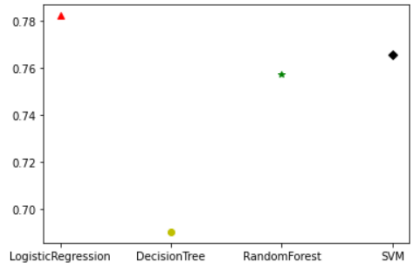
Decision tree : 69.04 %

Random forests : 75.73 %

Support vector machines(classifier) : 76.57 %

```
print(accuracy)
print(types)
i=0
colours=["r^","yo","g*","kD"]
for i in range(4) :
   plt.plot(types[i],accuracy[i],colours[i])
```

```
[0.7824267782426778, 0.6903765690376569, 0.7573221757322176, 0.7656903765690377]
['LogisticRegression', 'DecisionTree', 'RandomForest', 'SVM']
```



I HAVE USED 4 CLASSIFICATION ALGORITHMS TO POINT OUT THE ACCURACIES FOR THE TOP 2 FOR MY PROJECT AND THEN COMPARE THEM, WHICH ARE FOUND TO BE **LOGISTIC REGRESSION AND SUPPORT VECTOR MACHINE CLASSIFICATION ALGORITHMS WITH RESPECTIVELY 78.24 % and 76.57 % APPROX ACCURACIES IN PREDICTING THE TEST SET VALUES**

(THE .ipynb file has bin given in the zipped folder)

NAME- SWASTIK MOHAPATRA

STREAM- COMPUTER ENGINEERING

YEAR OF STUDY- 1st

COLLEGE- BHARATI VIDYAPEETH DEEMED TO UNIVERSITY BE COLLEGE OF ENGINEERING,PUNE.

College prn : 2014110710