Diabetes prediction

Under the Guidance of :
Srinivas

By:

Priyanka S K

Attributes in the dataset:

1. Pregnancies
2. Glucose
3. Blood Pressure
4. Skin Thickness
5. Insulin
6. BMI
7. Diabetes Pedigree Function
8. Age
9. Outcome

Explanation about the attribute of csv file:

1.Pregnancies:

- This column tells us the number of pregnancies that the patient has had.

2.Glucose:

- This column tells about the sugar level in the body of the patient.
- If the level of the sugar is more than 125 mg/dl then the patient is said to be diabetic.

3.Blood Pressure:

- Blood pressure column tells about the patient's diastolic blood pressure and is measured in millimetres of mercury

4.Skin Thickness:

- A measurement from the anterior abdomen's subcutaneous adipose tissue which has a average thickness of mercury.
- Since this dataset focuses on diabetes, it's likely that some of the data points would have much higher value for skin thickness then the average. Since diabetes mellitus is associated with increase created and retention of fat from glucose.

5.Insulin:

- When your blood sugar goes up,it signals your pancreas to release insulin. Insulin act like a key to let the blood sugar into your body's cells for use as energy.
- Measured in MIU/L or PMOI/L.

6.BIM(Body mass Index):

- It is measure of body fat based on height and weight that applies to adult men or women.
- Normal range of BIM 20 to 25 kg/m^2, less then 20 underweight, more than 25 over weight, more than 30 its OBES.

7.Diabetes Pedigree Function:

- It provides some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient.

8.Age:

- Age is one attribute which tells about the age of the patient or diabetic person or normal person age.

9.Outcome:

- Outcome says that after working on the data the prediction result is said to be coutcome whether 0 or 1,
- 1 means yes, 0 means not a diabetic person.

Different Model's we have tried:

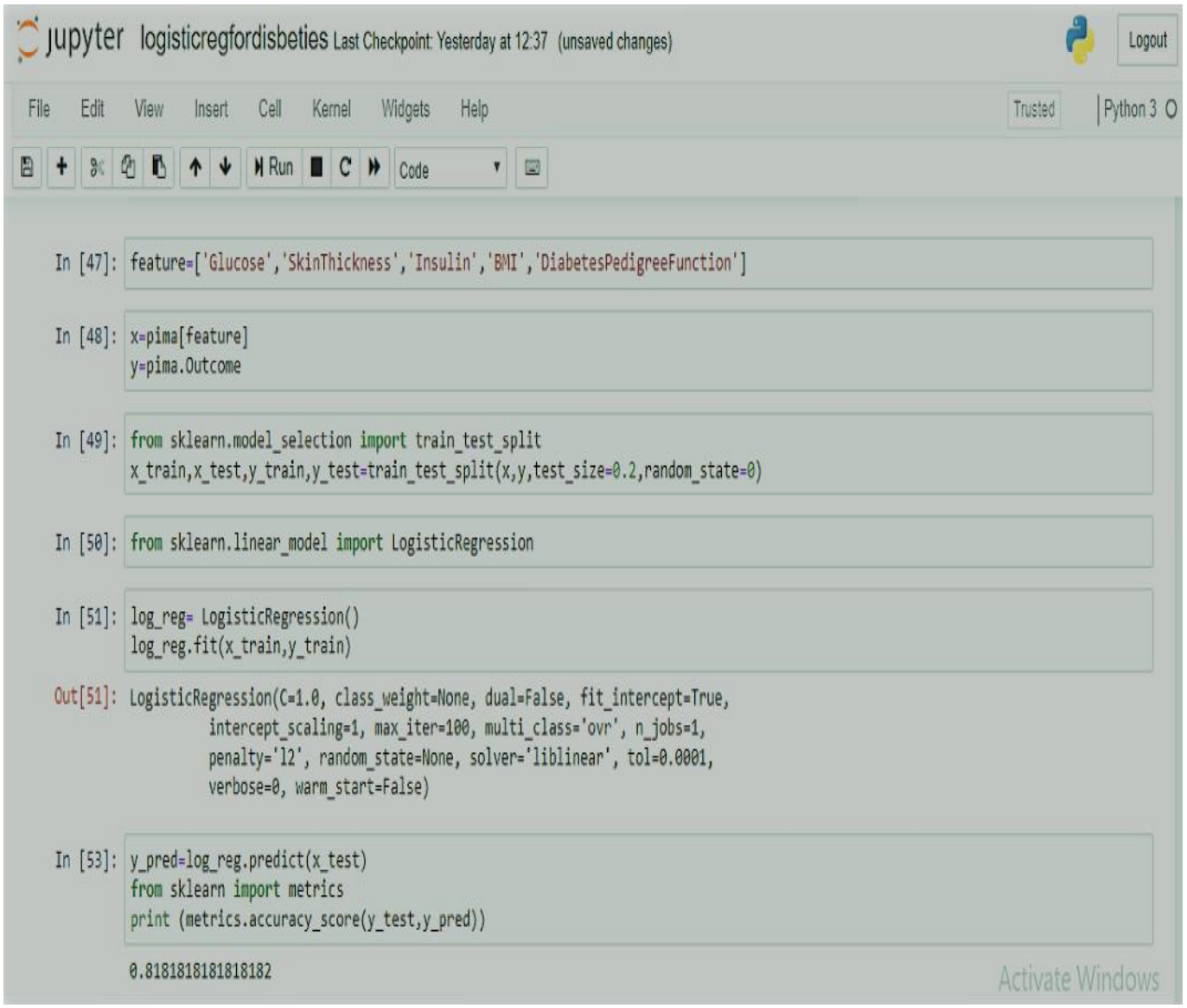
1.SVM

2.Decision tree

3. Logistic Regression.

Logistic Regression:

- While building this project I have selected some specific features which will give us more accuracy. Blood pressure and Age and pregnancy attributes are not much related to Diabetes. I have separated these attributes from the training data set before going for Prediction.
- Whereas blood pressure is no related to diabetes, pregnancies and age are independent and somewhat related to Diabetic person but we can exclude in feature selection.
- When I tested with testing data set I observed that accuracy is 81% when I worked with all the attributes the accuracy was lesser then what we have got now.
- By comparing complexity, accuracy and all I felt Logistic Regression is the best model.


jupyter logisticregfordisbeties Last Checkpoint: Yesterday at 12:37 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [47]: feature=['Glucose','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction']
In [48]: x=pima[feature]
y=pima.Outcome
In [49]: from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
In [50]: from sklearn.linear_model import LogisticRegression
In [51]: log_reg= LogisticRegression()
log_reg.fit(x_train,y_train)
Out[51]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
verbose=0, warm_start=False)
In [53]: y_pred=log_reg.predict(x_test)
from sklearn import metrics
print (metrics.accuracy_score(y_test,y_pred))
0.8181818181818182
Activate Windows

Conclusion:

- ✓ We stared the project with how Machine learning is used in the field of Medical science.
- ✓ Flowed by, we are doing the problem of Diabetic Prediction.
- ✓ Based on applying several models I decided Logistic Regression as the best model.