Exposys Data Labs

Data Science Internship Program

Souvik Chakraborty

- ## *Problem Statement*

Diabetes is a type of chronic disease which is more common among the people of all age groups. Predicting this disease at an early stage can help a person to take the necessary precautions and change his/her lifestyle accordingly to either prevent the occurrence of this disease or control the disease.

- ## *Task*

1. Prepare the data-set using several methods to train the model.
2. Build a model which can give high accuracy of predicting the disease.

- ## *Approach to Address*

1. Selecting correct parameters for creating the diabetes dataset.
2. Evaluating the parameters and dropping the unnecessary one.
3. Normalizing all the input parameters at a standard level.
4. Creating the train and test datasets for the ML algorithms.
5. Predicting through various algorithm, to find the most accurate one.

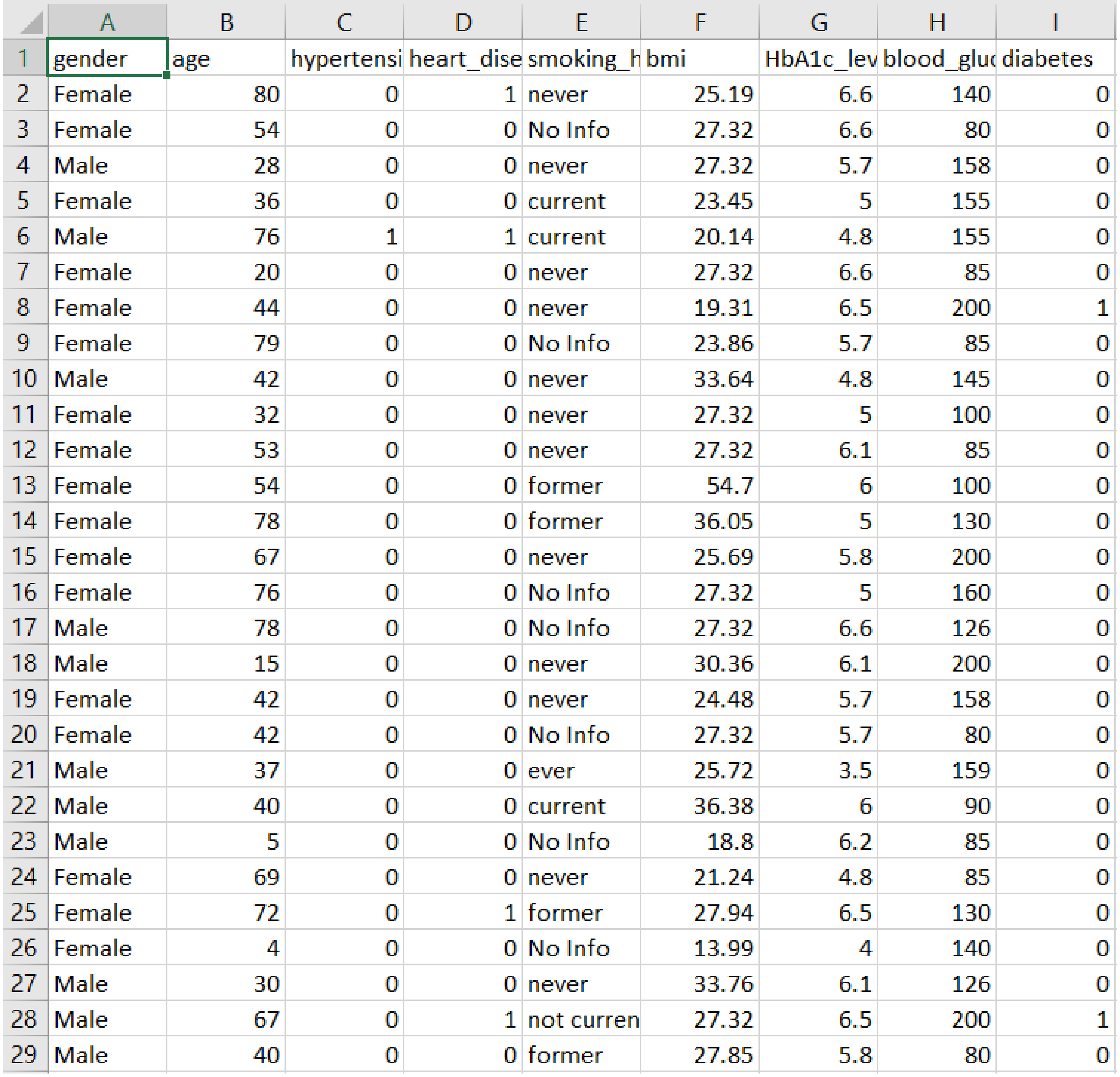
- ## *Selected Parameters*

1. **Gender**: - It has been observed that diabetes mellitus can have some gender peculiarities and some data show that women have more years of disease on average than their male counterparts and have a higher body mass index (BMI).
2. **Age:** - Middle age is when diabetes diagnoses really start to spike. An estimated 14% of Americans ages 45 to 64, or 11 million people, are diagnosed with type 2. That is almost five times the rate for those 18 to 44. Diabetes rates jump even higher at the onset of your senior years. Almost 25% of Americans 65 and older have been diagnosed with type 2. Undiagnosed cases may account for another 4.7%. That means more than 1 of every 4 oldest Americans lives with type 2 diabetes.
3. **Hypertension: -** Hypertension and diabetes share a number of common causes and risk factors. A person who has one condition is at an increased risk for developing the other. Likewise, a person who has both conditions may find that each condition worsens the other.
4. **Heart Disease: -** Diabetes and heart disease often go hand in hand. Learn how to protect your heart with simple lifestyle changes that can also help you manage diabetes.
5. **BMI: -** The risk of developing DM at $30 \leq \text{BMI} \leq 39.99$ relative to persons with normal BMI was nearly the same for both men and women (HR=1.98 for men vs. HR=1.96 for women). At $\text{BMI} \geq 40$, the risk of DM was higher among men (HR=2.85 for men vs. HR=2.51 for women).
6. **HbA1c level: -** A high HbA1c means you have too much sugar in your blood. This means you are more likely to develop diabetes complications, like serious problems with your eyes and feet. Knowing your HbA1c level and what you can do to lower it will help you reduce your risk of devastating complications

- ## *Software used*

1. MS-Excel (to build and analyze the dataset)
2. VS Code (used to write the python code for predictive analysis)

- ***Dataset***



	A	B	C	D	E	F	G	H	I
1	gender	age	hypertensi	heart_dise	smoking_h	bmi	HbA1c_lev	blood_gluc	diabetes
2	Female	80	0	1	never	25.19	6.6	140	0
3	Female	54	0	0	No Info	27.32	6.6	80	0
4	Male	28	0	0	never	27.32	5.7	158	0
5	Female	36	0	0	current	23.45	5	155	0
6	Male	76	1	1	current	20.14	4.8	155	0
7	Female	20	0	0	never	27.32	6.6	85	0
8	Female	44	0	0	never	19.31	6.5	200	1
9	Female	79	0	0	No Info	23.86	5.7	85	0
10	Male	42	0	0	never	33.64	4.8	145	0
11	Female	32	0	0	never	27.32	5	100	0
12	Female	53	0	0	never	27.32	6.1	85	0
13	Female	54	0	0	former	54.7	6	100	0
14	Female	78	0	0	former	36.05	5	130	0
15	Female	67	0	0	never	25.69	5.8	200	0
16	Female	76	0	0	No Info	27.32	5	160	0
17	Male	78	0	0	No Info	27.32	6.6	126	0
18	Male	15	0	0	never	30.36	6.1	200	0
19	Female	42	0	0	never	24.48	5.7	158	0
20	Female	42	0	0	No Info	27.32	5.7	80	0
21	Male	37	0	0	ever	25.72	3.5	159	0
22	Male	40	0	0	current	36.38	6	90	0
23	Male	5	0	0	No Info	18.8	6.2	85	0
24	Female	69	0	0	never	21.24	4.8	85	0
25	Female	72	0	1	former	27.94	6.5	130	0
26	Female	4	0	0	No Info	13.99	4	140	0
27	Male	30	0	0	never	33.76	6.1	126	0
28	Male	67	0	1	not curren	27.32	6.5	200	1
29	Male	40	0	0	former	27.85	5.8	80	0

This is a limited section of the dataset. The actual size of the dataset is (100001, 9)

- ***Python Libraries***

Some of the python libraries used during the implementation of the project are: -

1. Pandas
2. NumPy
3. Scikit-Learn
4. Matplotlib

- *ML algorithm*

   1. Logistic Regression
   2. Decision Tree
   3. Linear Regression

- *Code*

```
#___Importing all the Libraries needed for the project___#

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score
import sklearn.metrics as metrics

#___Importing the Raw data___#

data = pd.read_csv("F:\My_python_programs\Diabetes Dataset\diabetes_prediction_dataset.csv")
#print(data)

#___Preprocessing of the inserted raw data___#

#print(data.shape)
#iszero = data.isnull().values.any()
#print(iszero)
Parameter = data.iloc[:,2:7]
#print(Parameter)
resultant = data.iloc[:,data.columns == 'diabetes']
#print(resultant)
initial_dataset = pd.concat([Parameter,resultant],axis=1)
#print(initial_dataset)
str2int = preprocessing.LabelEncoder()
initial_dataset['smoking_history'] = str2int.fit_transform(initial_dataset['smoking_history'])
#print(initial_dataset)


#___visualising the data using matplotlib library___#

outcome_count = pd.value_counts(initial_dataset['diabetes'], sort = True).sort_index()
outcome_count.plot(kind = 'bar', color='yellow')
```

```
plt.title('diabetes histogram', fontweight='bold', fontsize='15', color='red')
plt.xlabel('diabetes', fontweight='bold', fontsize='15', color='gray')
plt.ylabel('Frequency', fontweight='bold', fontsize='15', color='gray')
#print(plt.show())

#___all columns have to standard with respect to other___#

stan = preprocessing.StandardScaler()
initial_dataset['new_bmi'] = stan.fit_transform(initial_dataset['bmi']. values.reshape(-1,1))
final_dataset = initial_dataset.drop(['bmi'], axis = 1)
#print(final_dataset)

#___Split data for training and testing___#

final_parameter = final_dataset.iloc[:,final_dataset.columns != 'diabetes']
final_resultant = final_dataset.iloc[:,final_dataset.columns == 'diabetes']
#print(final_parameter)
#print(final_resultant)
x_train, x_test, y_train, y_test = train_test_split(final_parameter, final_resultant, test_size = 0.2)

#___Applying Logistic Regression___#

model = LogisticRegression()
model.fit(x_train, y_train.values.ravel())
result = model.predict(x_test)
#print(result)
accuracy = model.score(x_test,y_test)
print("Accuracy using Logistic regression:", accuracy)

#___Applying Decission Tree Classifier___#

clf = DecisionTreeClassifier(criterion="entropy", max_depth=4)
clf = clf.fit(x_train,y_train)
yPred = clf.predict(x_test)
#print(yPred)
scores = cross_val_score(clf, y_test, yPred, cv=10)
print("Accuracy using Decision tree:",metrics.accuracy_score(y_test, yPred))
#print(scores)

#__Applying Linear Regression__#
reg = LinearRegression()
reg.fit(x_train,y_train)
y_pred = reg.predict(x_test)
print("Root mean squared error: ", np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

Output

```
Accuracy using Logistic regression: 0.94425
Accuracy using Decision tree: 0.95485
Root mean squared error:  0.2424602229405095
```

After observing the program outcome, we can say that Linear regression is the perfect algorithm to be used in order to get maximum accuracy for the diabetes predicting task.

Conclusion

It was a good opportunity with an exciting project to step into the field of Data Science. There are countless number of application of Data Science. In this project we are creating a diabetes dataset and with the help of libraries, we are performing predictive analysis. It is done to catch an early onset of diabetes in a healthy person, so that we can treat the disease in an early stage. By completing the whole project, I have gathered knowledge on various ML algorithms, Data pre-processing and creating a perfect dataset.