

Phase-3: Development Part 1

Title: AI Based Diabetes Prediction System

Dataset link: <https://www.kaggle.com/datasets/mathchi/diabetes-data-set>

Software used: google colab

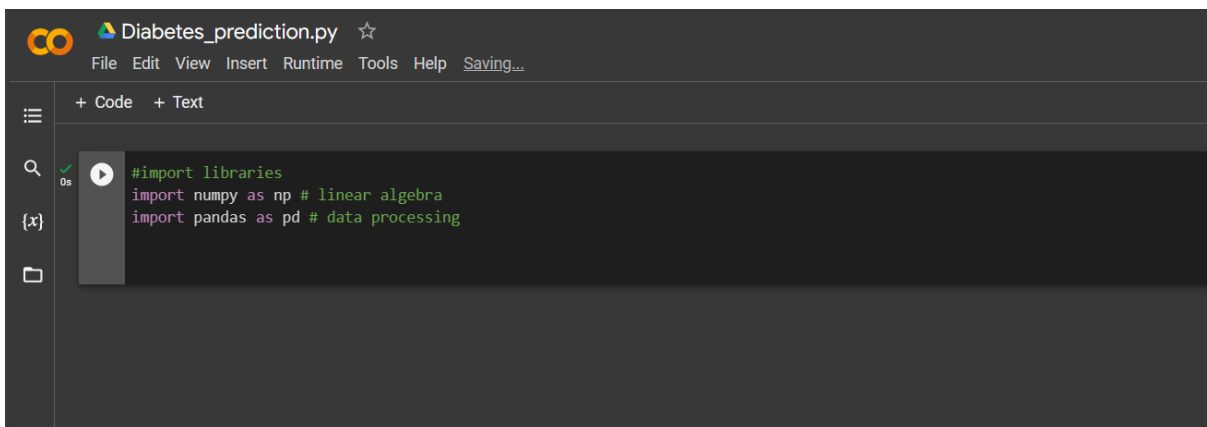
Introduction:

Diabetes is a health condition that affects how your body turns food into energy. Most of the food you eat is broken down into sugar (also called glucose) and released into your bloodstream. When your blood sugar goes up, it signals your pancreas to release insulin.

Process:

1.Installing Libraries:

In this first step I have imported most common libraries used in python for machine learning such as Pandas, Numpy etc.

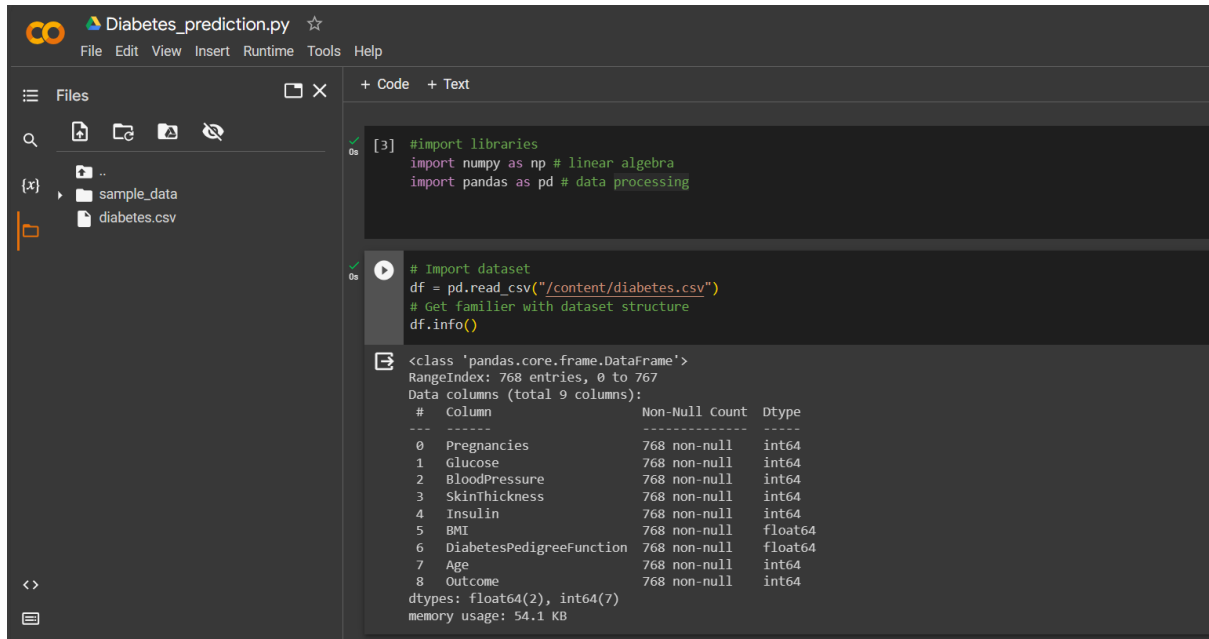


```
Diabetes_prediction.py ☆
File Edit View Insert Runtime Tools Help Saving...
+ Code + Text
#import libraries
import numpy as np # linear algebra
import pandas as pd # data processing
```

2.Importing Data:

The diabetes data set was originated from <https://www.kaggle.com/datasets/mathchi/diabetes-data-set>.

Diabetes dataset containing 768 instances with 9 features. The objective is to predict if the patient is diabetic or not.



The screenshot shows a Jupyter Notebook interface with a file explorer on the left and a code editor on the right. The file explorer shows a folder named 'sample_data' containing a file named 'diabetes.csv'. The code editor shows the following code:

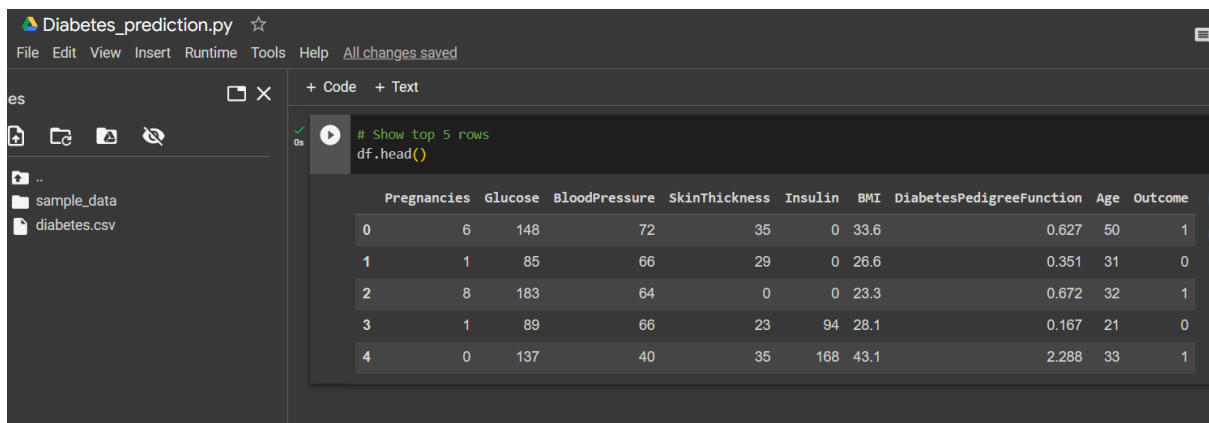
```
[3]: #import libraries
import numpy as np # linear algebra
import pandas as pd # data processing

# Import dataset
df = pd.read_csv("/content/diabetes.csv")
# Get familiar with dataset structure
df.info()
```

The output of the code is displayed below the code cell:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
#   Column                      Non-Null Count  Dtype
---  ---                      ---
0   Pregnancies                 768 non-null   int64
1   Glucose                    768 non-null   int64
2   BloodPressure              768 non-null   int64
3   SkinThickness              768 non-null   int64
4   Insulin                    768 non-null   int64
5   BMI                        768 non-null   float64
6   DiabetesPedigreeFunction    768 non-null   float64
7   Age                        768 non-null   int64
8   Outcome                    768 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Displaying Data:



The screenshot shows a Jupyter Notebook interface with a file explorer on the left and a code editor on the right. The file explorer shows a folder named 'sample_data' containing a file named 'diabetes.csv'. The code editor shows the following code:

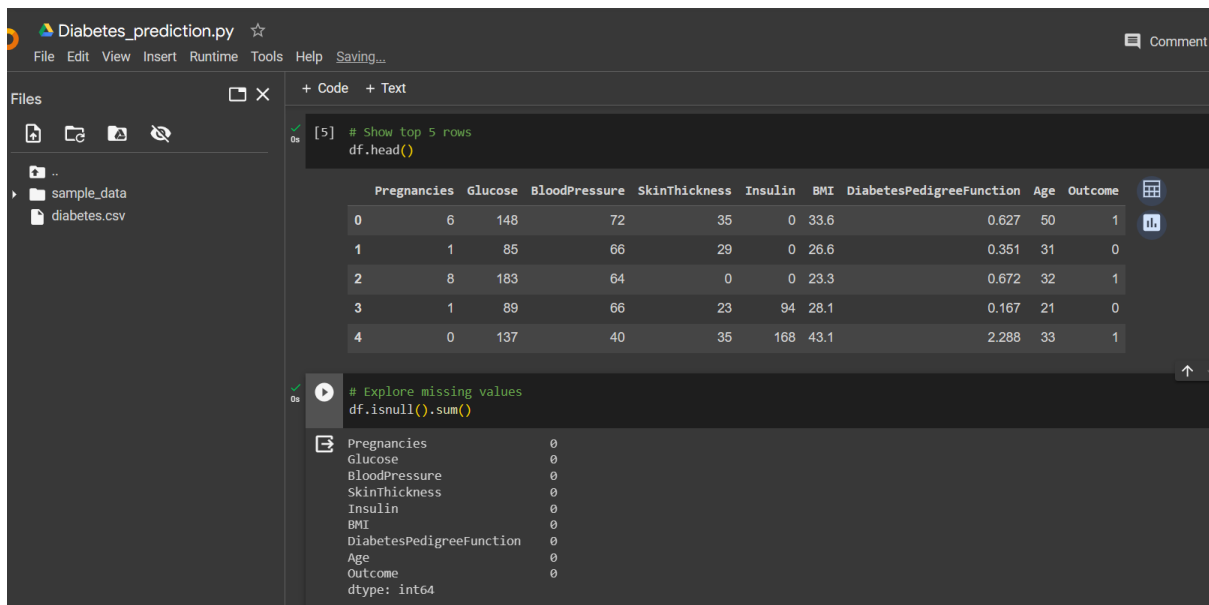
```
# Show top 5 rows
df.head()
```

The output of the code is displayed below the code cell, showing the first 5 rows of the dataset:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Data Preprocessing:

Next, i will cleanup the dataset which is the important part of data science. Missing data can lead to wrong statistics during modeling and predictions.



```
Diabetes_prediction.py ☆
File Edit View Insert Runtime Tools Help Saving...

Files
  sample_data
  diabetes.csv

+ Code + Text

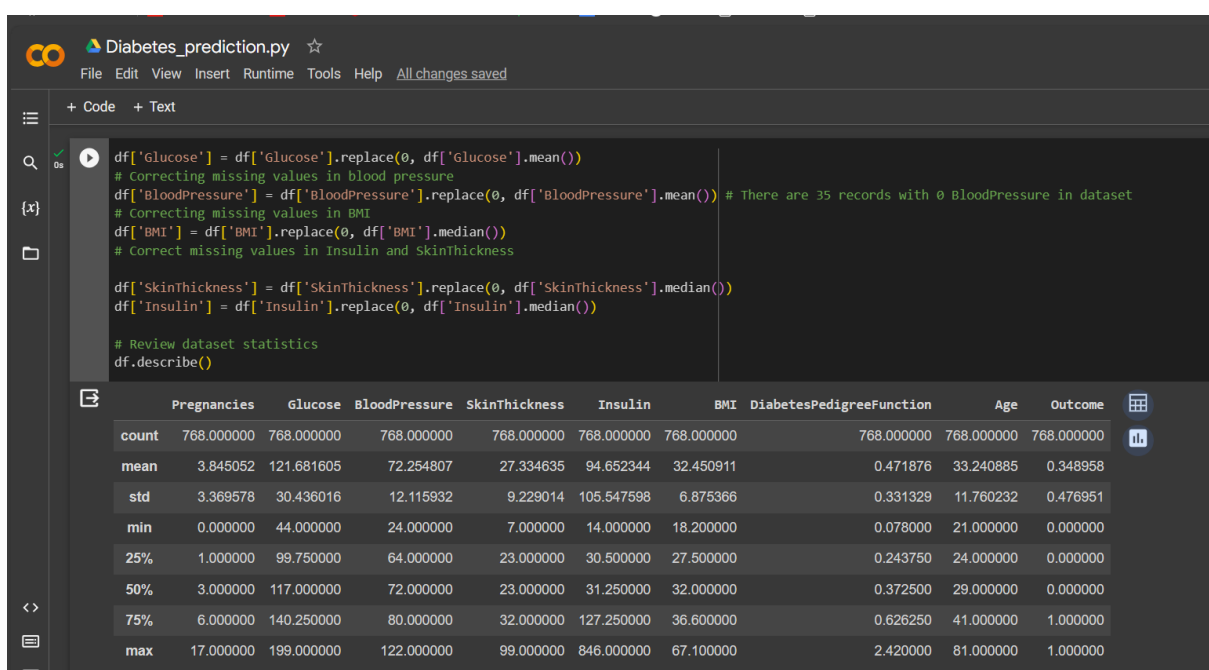
[5] # Show top 5 rows
df.head()

Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin  BMI  DiabetesPedigreeFunction  Age  Outcome
0           6       148             72             35         0   33.6                0.627    50         1
1           1        85             66             29         0   26.6                0.351    31         0
2           8       183             64              0         0   23.3                0.672    32         1
3           1        89             66             23        94   28.1                0.167    21         0
4           0       137             40             35       168   43.1                2.288    33         1

# Explore missing values
df.isnull().sum()

Pregnancies      0
Glucose           0
BloodPressure     0
SkinThickness     0
Insulin           0
BMI              0
DiabetesPedigreeFunction  0
Age              0
Outcome          0
dtype: int64
```

Missing value analysis:



```
Diabetes_prediction.py ☆
File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

df['Glucose'] = df['Glucose'].replace(0, df['Glucose'].mean())
# Correcting missing values in blood pressure
df['BloodPressure'] = df['BloodPressure'].replace(0, df['BloodPressure'].mean()) # There are 35 records with 0 BloodPressure in dataset
# Correcting missing values in BMI
df['BMI'] = df['BMI'].replace(0, df['BMI'].median())
# Correct missing values in Insulin and SkinThickness

df['SkinThickness'] = df['SkinThickness'].replace(0, df['SkinThickness'].median())
df['Insulin'] = df['Insulin'].replace(0, df['Insulin'].median())

# Review dataset statistics
df.describe()

Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin  BMI  DiabetesPedigreeFunction  Age  Outcome
count  768.000000  768.000000  768.000000  768.000000  768.000000  768.000000  768.000000  768.000000  768.000000
mean    3.845052  121.681605   72.254807   27.334635   94.652344   32.450911    0.471876   33.240885   0.348958
std     3.369578   30.436016   12.115932    9.229014  105.547598    6.875366    0.331329   11.760232   0.476951
min      0.000000    44.000000   24.000000    7.000000   14.000000   18.200000    0.078000   21.000000   0.000000
25%      1.000000   99.750000   64.000000   23.000000   30.500000   27.500000    0.243750   24.000000   0.000000
50%      3.000000  117.000000   72.000000   23.000000   31.250000   32.000000    0.372500   29.000000   0.000000
75%      6.000000  140.250000   80.000000   32.000000  127.250000   36.600000    0.626250   41.000000   1.000000
max     17.000000  199.000000  122.000000   99.000000  846.000000   67.100000    2.420000   81.000000   1.000000
```

Result:

Thus the Dataset can be loaded and preprocessed.