AI BASED DIABETES PREDICTION SYSTEM

USING MACHINE LEARNING ALGORITHM

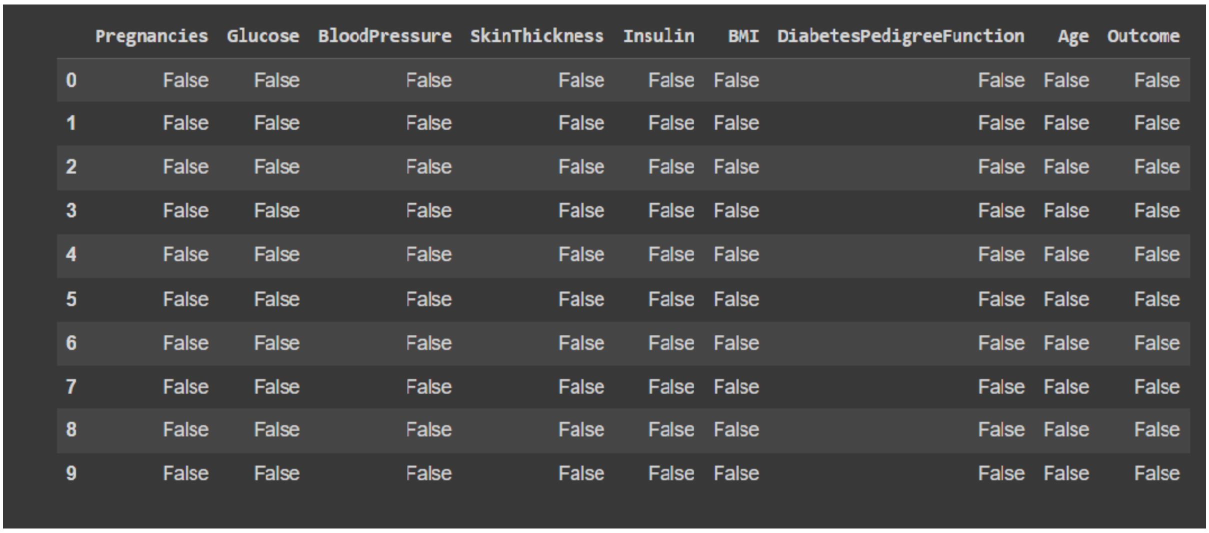
Phase 3: Here, in this phase developing the diabetes prediction system by preparing the data and selecting relevant features.

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

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

from mlxtend.plotting import plot_decision_regions

import missingno as msno

from pandas.plotting import scatter_matrix

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import confusion_matrix

from sklearn import metrics

from sklearn.metrics import classification_report

import warnings

warnings.filterwarnings('ignore')

%matplotlib inline

diabetes_df = pd.read_csv('./sample_data/diabetes.csv')
```

# Now let's check that if our dataset have null values or not

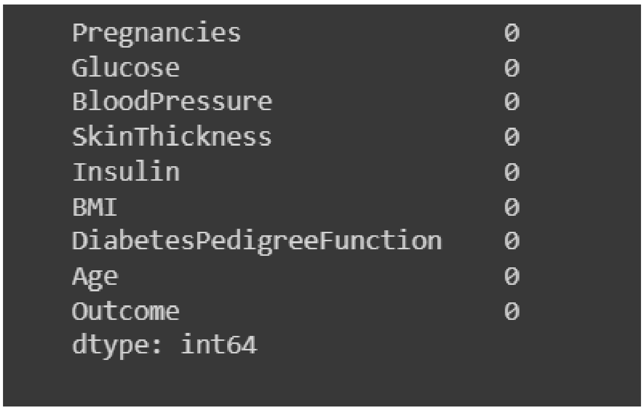
diabetes_df.isnull().head(10)



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False	False	False
8	False	False	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False	False	False

# Now let's check that if our dataset have null values or not

diabetes_df.isnull().sum()



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

Here from above code we first checked that is there any null values from isnull() function then we are going to take the sum of all those missing values from sum() function and the inference we now get is that there are no missing values but that is actually not a true story as in this particular dataset all the missing values were given the 0 as value which is not good for the authenticity of the dataset. Hence we will first replace the 0 value to NAN value then start the imputation process.

4 Machine learning algorithm to proceed the prediction analysis :

- Random Forest
- Decision Tree
- XgBoost classifier
- Support Vector Machine (SVM)

FEATURES SELECTED:

- Pregnancies
- Glucose
- Blood pressure
- Skin thickness
- Insulin
- BMI
- Age
- Diabetes pedigree Function