Assignment 4 – Random Forest

Problem Statement:

You work in XYZ Company as a Python Data Scientist. The company officials have collected some data on diabetes based on years of experience and wish for you to create a model from it. Dataset: diabetes.csv

Tasks To Be Performed:

1. Load the dataset using pandas
2. Extract data from outcome column is a variable named Y
3. Extract data from every column except outcome column in a variable named X
4. Divide the dataset into two parts for training and testing in 70% and 30% proportion
5. Create and train Random Forest Model on training set
6. Make predictions based on the testing set using the trained model
7. Check the performance by calculating the confusion matrix and accuracy score of the model

In [1]:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import *
from sklearn.ensemble import RandomForestClassifier
```

In [2]:

```
df = pd.read_csv(r"csv files/diabetes-2.csv")
df.head()
```

Out[2]:

	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

In [3]:

```
X = df.drop(columns=['Outcome'])
y = df['Outcome']
```

In [4]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.30, random_state=2)
```

In [5]:

```
rand_forest = RandomForestClassifier()
rand_forest.fit(X_train,y_train)
```

Out[5]:

```
▾ RandomForestClassifier
RandomForestClassifier()
```

In [6]:

```
y_pred = rand_forest.predict(X_test)
```

In [7]:

```
confusion_matrix(y_test, y_pred)
```

Out[7]:

```
array([[138,  17],
       [ 38,  38]], dtype=int64)
```

In [8]:

```
accuracy_score(y_test, y_pred)
```

```
Out[8]: 0.7619047619047619
```

```
In [ ]:
```