Python Programming

Machine Learning Assignment

Consider below Dataset as

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0

🎯 Objective:

Build a Machine Learning model to **predict whether a patient is diabetic (1) or not (0)** based on medical attributes.

Task Instructions:

You must complete the following steps

1. Exploratory Data Analysis (EDA):

- Load the dataset using `pandas`.
- Display the first 5 rows.
- Show column info and check for null values.
- Display basic statistics using `.describe()`.
- Plot the **distribution of the target variable** (`Outcome`).
- Use graphs like `hist`, `boxplot`, or `pairplot` to identify patterns or outliers.

2. Data Preprocessing:

- Check and handle **missing or zero values** in columns like `Glucose`, `BloodPressure`, etc.
- Apply feature scaling using **StandardScaler or MinMaxScaler**.
- Split the dataset into **features (X)** and **target (y)**.

3. Model Building:

Train **at least 2 different algorithms** on the dataset:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Use `train_test_split` to divide the data.

4. Model Evaluation:

- Print **accuracy score**, **confusion matrix**, **precision**, **recall**, and **F1 score**.
- Use **matplotlib or seaborn** to visualize confusion matrix.

5. Final Output:

- Predict whether a patient is diabetic based on test data.
- Display predictions on screen and save them in a CSV file.