TASK 3

Title: Decision Tree Classifier on Bank Marketing Dataset

Objective

To build a **decision tree classifier** that predicts whether a customer will purchase a product or service based on their **demographic** and **behavioral** data, using the Bank Marketing dataset from the UCI Machine Learning Repository.

Dataset Information

- **Dataset Name:** Bank Marketing Dataset
- **Source:** UCI Machine Learning Repository
- **URL:** https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

- **Records:** 45,211
- **Features:**
- **Input Variables:** Age, Job, Marital Status, Education, Default, Balance, Housing Loan, etc.
- **Target Variable:** y (Has the client subscribed to a term deposit? Yes/No)

Steps Performed

1. Data Preprocessing

- Handled missing values (if any)
- Encoded categorical variables using Label Encoding or One-Hot Encoding
- Normalized numerical features like balance, duration, campaign

- Split dataset into **training** and **testing** sets (e.g., 80-20 split)

2. Model Building

- Used **Decision Tree Classifier** from sklearn.tree
- Trained the model using fit() on training data
- Visualized the tree using plot_tree() or export_graphviz()

3. Model Evaluation

- Calculated metrics:
- **Accuracy**
- **Precision**
- **Recall**

- **F1-score**
- **Confusion Matrix**
- Performed **cross-validation** to ensure robustness
- Evaluated feature importance

Key Insights

- **Duration** of the call was the most significant factor in determining subscription.
- Clients with housing or personal loans were less likely to subscribe.
- Job type and education level also influenced the prediction.
- Target marketing is more effective when calls are short and focused.

Tools & Technologies Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- **IDE:** Jupyter Notebook / Google Colab

Conclusion

The decision tree classifier efficiently predicted whether a customer would subscribe to a term deposit. Key features like call duration, job type, and loan status significantly impacted the model. This model can be further improved with ensemble methods like **Random Forest** or **XGBoost**.

Link to Dataset

🔗 Bank Marketing Dataset – UCI