Assignment: Customer Churn Prediction Using PySpark MLlib

Objective

Build a machine learning pipeline using **PySpark MLlib** to predict customer churn.

Part 1 – Data Preparation

1. **Dataset:** Use the provided `churn.csv`.
    - Columns: `customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn`.
2. **Tasks:**
    - Load the CSV into a Spark DataFrame.
    - Display schema using `printSchema()`.
    - Show the first 10 rows using `show()`.
    - Count the number of churned and non-churned customers using `groupBy().count()`.

Part 2 – Feature Engineering

1. Convert categorical columns to numeric using `StringIndexer`.
2. Assemble all features into a single vector using `VectorAssembler`.
3. Split the dataset into **training (70%)** and **test (30%)** sets.

Part 3 – Model Training

1. Train a **Logistic Regression** model to predict churn.
2. Train a **Decision Tree Classifier** and compare with Logistic Regression.
3. (Optional) Train a **Random Forest Classifier** for improved accuracy.

Part 4 – Model Evaluation

1. Use **BinaryClassificationEvaluator** to calculate **AUC (Area Under ROC)**.
2. Print **precision, recall, and accuracy**.
3. Display the **confusion matrix**.

Part 5 – Bonus Tasks

1. Tune hyperparameters using **CrossValidator** or **TrainValidationSplit**.
2. Try **feature importance** extraction using Decision Tree or Random Forest.
3. Export the final model and demonstrate how to load it back using `PipelineModel.load()`.