Data Science Assignment

Objective

This assignment is structured to assess your skills in three key areas:

1. **Traditional Machine Learning Algorithms**
2. **Generative AI (Gen AI)**
3. **Logical and Analytical Problem-Solving**

Each section evaluates your ability to analyze data, implement solutions, and generate insights. Follow the tasks, structure your code well, and ensure clarity in your explanations.

Section 1: Traditional ML Algorithms

Task: Predict Customer Churn

Problem Statement:

You are provided with a **telecom customer dataset** that contains historical customer usage patterns, demographics, and churn labels (binary target: `1` for churn, `0` for no churn). Your goal is to build a machine learning model to predict customer churn and generate actionable insights.

Dataset:

Download the Telecom Churn Dataset

Your Tasks:

1. **Perform Exploratory Data Analysis (EDA)**:
    - Visualize key patterns and relationships in the data.
    - Identify important features contributing to customer churn.
2. **Build and Evaluate ML Models**:
    - Train at least **3 traditional ML models** (e.g., Logistic Regression, Random Forest, XGBoost).
    - Use proper cross-validation and hyperparameter tuning techniques.
    - Evaluate model performance using metrics: **Precision, Recall, F1-score, and ROC-AUC**.
3. **Provide Insights**:
    - Identify the top predictors of churn and explain their impact.
    - Suggest strategies to reduce churn based on your findings.
4. **Deliverable**:
    - A well-documented Jupyter Notebook containing:
        - EDA, model implementation, evaluation metrics, and final insights.

Section 2: Generative AI (Gen AI)

Task: Summarize Product Reviews and Generate Synthetic Reviews

Problem Statement:

You are provided with a **product reviews dataset** containing raw customer reviews and sentiment labels (Positive, Neutral, Negative). Your goal is to implement a **text summarization pipeline** and generate synthetic reviews using a Generative AI model.

Dataset:

Download the Product Reviews Dataset

Your Tasks:

1. **Text Summarization**:
    - Summarize product reviews into concise 2-3 sentence summaries using a **pre-trained transformer model** (e.g., Hugging Face T5, GPT-based models).
2. **Sentiment Analysis**:
    - Classify reviews into Positive, Neutral, or Negative sentiment using a transformer-based sentiment analysis model.
3. **Synthetic Review Generation**:
    - Use a **fine-tuned GPT model** or any pre-trained generative model to generate synthetic product reviews for the following cases:
        - Positive review for a product rated 5 stars.
        - Negative review for a product rated 1 star.
4. **Sentiment Consistency**:
    - Compare the sentiment of the **synthetic reviews** with the original sentiment using a sentiment analysis classifier.
5. **Deliverable**:
    - A Python script or Jupyter Notebook demonstrating:
        - Text summarization results.
        - Sentiment classification accuracy.
        - Synthetic reviews for both positive and negative cases.
        - A comparison of sentiments between original and generated reviews.

Section 3: Logical and Analytical Problem Solving

Task: Resource Allocation Optimization – Production Scheduling

Problem Statement:

A mid-sized manufacturing company produces **3 types of products**:

- **Product A**, **Product B**, and **Product C**.

Each product requires different amounts of resources (labor hours, raw material) to produce. The company has a limited availability of **240 labor hours** and **180 kg of raw materials** per month. Your goal is to determine the optimal production quantities to **maximize profit** while ensuring resource constraints are met.

Details and Constraints:

Product	Profit per Unit ($)	Labor Hours per Unit	Raw Material per Unit (kg)
Product A	30	5	3
Product B	20	4	2
Product C	50	6	4

- **Labor Hours**: 240 hours per month
- **Raw Material**: 180 kg per month

Your Tasks:

1. **Define the Problem**:
   - Formulate the total profit function: Profit=(30×A)+(20×B)+(50×C)
   - Add constraints for labor hours and raw materials: 5A+4B+6C≤240(Labor Hours) 3A+2B+4C≤180(Raw Material)
2. **Optimize Using Linear Programming**:
   - Use `scipy.optimize.linprog` to determine the optimal production quantities for Products A, B, and C.
3. **Simulate Changes**:
   - Simulate the impact of **increasing raw material availability by 10%** and compare the new profit with the original scenario.
4. **Deliverable**:
   - A Python script that outputs:
     - Optimal production quantities for each product.
     - Total profit.
     - A comparison table showing the impact of increased raw material availability.
5. **Output Example**:

Product	Units Produced	Profit Contribution ($)

Product A	X	Y
Product B	X	Y
Product C	X	Y
Total	-	Z

Evaluation Criteria

1. **Section 1 (Traditional ML)**:
    - Depth of EDA, feature engineering, and model performance.
2. **Section 2 (Gen AI)**:
    - Quality of summarization, synthetic reviews, and sentiment consistency.
3. **Section 3 (Logical Problem Solving)**:
    - Correct formulation, optimal results using linear programming, and logical clarity.
4. **Code Quality**:
    - Readable, modular, and well-documented code.
5. **Insights and Communication**:
    - Clear presentation of results, conclusions, and insights.

Submission Guidelines:

- Submit the assignment as a **GitHub repository** or a zipped folder containing:
    - Jupyter Notebooks / Python Scripts
    - Final outputs and insights in Markdown or PDF format.

Duration: 3 days