Course Project for Advance Techniques in Data Science

This project encourages creativity and practical application of data science skills, providing a hands-on experience of the entire lifecycle of a data science project.

1. **Problem Definition**
    - Identify a real-world problem to address using data. This could span domains such as healthcare, finance, marketing, or environmental science.
    - Examples:
        - Predicting house prices based on property features.
        - Analyzing customer churn for a subscription-based service.
        - Forecasting stock prices or energy consumption.
    - Clearly state the problem and outline its significance, objectives, and potential impact.
2. **Asking the Right Questions**
    - Formulate specific, actionable questions related to the chosen problem.
    - Example Questions:
        - What are the key factors influencing the target variable?
        - How accurate can predictions be for the target variable?
        - Are there any patterns or trends that offer insights into the problem?
3. **Data Collection**
    - Obtain a dataset relevant to the problem. Students may use sources such as:
        - Kaggle
        - UCI Machine Learning Repository
        - Public APIs or open data platforms (e.g., government datasets)
    - Ensure the data contains sufficient features and observations to perform meaningful analysis.
4. **Data Wrangling (Preprocessing)**
    - Clean the dataset to ensure it is usable:
        - Handle missing values.
        - Remove or address outliers.
        - Convert data types where necessary (e.g., dates to datetime format).
    - Feature engineering:
        - Create new variables if necessary.
        - Standardize or normalize numerical features.
5. **Exploratory Data Analysis (EDA)**
    - Conduct a thorough EDA to understand the dataset.
    - Use visualizations and summary statistics to:
        - Identify patterns and trends.
        - Examine the relationships between features and the target variable.
        - Highlight anomalies or inconsistencies.
    - Tools: Python libraries such as pandas, seaborn, matplotlib, and plotly.
6. **Predictive Analysis**
    - Develop a machine learning model to solve the problem:

        - Split the data into training and testing sets.
        - Choose appropriate algorithms (e.g., regression, classification, clustering).
    - Evaluate the model using relevant metrics:
        - Regression: R^2, RMSE, MAE.
        - Classification: Accuracy, precision, recall, F1-score.
    - Optimize the model through techniques like hyperparameter tuning or feature selection.

7. **Deliverables**
    - A detailed project report including:
        - Problem definition and objectives.
        - Questions posed and their relevance.
        - Data source and description.
        - Data wrangling and cleaning steps.
        - Insights from EDA.
        - Model choice, performance metrics, and interpretation.
    - Code submission (Jupyter Notebook).
    - Presentation of findings (Not more than 10 slides 15 minutes will be given for the presentation).

Evaluation Criteria

- **Clarity**: Well-defined problem and objectives.
- **Depth of Analysis**: Thorough data exploration and interpretation.
- **Modeling**: Appropriateness of the chosen model and performance evaluation.
- **Creativity**: Innovative approaches in feature engineering or problem-solving.
- **Presentation**: Clear and concise communication of results and insights.