Data Science & AI Project Plan (October – April)

Project Topic: Finance

This project will focus on finance-related data, guiding us through data engineering, data analytics, and machine learning phases. The goal is to learn by doing and produce tangible results by the end of the academic year.

Overall Timeline:

- **Data Engineering & Data Fundamentals**: Mid-October to November
- **Data Analytics (Non-ML)**: December to January
- **Machine Learning & AI**: February to Early April

Phase 1: Data Engineering & Data Fundamentals (Mid-October to End-November)

Goal: Build a strong foundation for data collection, cleaning, and storage, ensuring our dataset is reliable for further analysis and modeling.

Key Topics:

- **Data Collection**: Gathering financial data from sources like Yahoo Finance, stock market APIs, and financial reports.
- **Data Cleaning**: Handling missing values, correcting inconsistencies, and organizing data for analysis using tools like Python's pandas library.
- **Data Storage**: Setting up efficient storage solutions using CSV files, databases (e.g., SQL), or cloud storage.
- **ETL (Extract, Transform, Load)**: Building simple data pipelines to automate data collection, cleaning, and storage.

Deliverables:

- Clean and well-structured finance-related datasets ready for analysis.
- ETL pipelines set up to automatically gather and process data.

Team Tasks:

- General members work on smaller projects, such as data cleaning tasks or building basic ETL pipelines for simpler financial datasets.
- Core team members focus on creating the main pipeline for collecting and preparing the large dataset that will be used throughout the project.

Phase 2: Data Analytics (Non-ML) (December to January)

Goal: Explore and analyze the data to extract meaningful insights without using machine learning.

Key Topics:

- **Exploratory Data Analysis (EDA)**: Identifying trends, patterns, and correlations in financial data.
- **Descriptive Statistics**: Calculating and understanding basic statistics (mean, median, standard deviation, etc.).
- **Data Visualization**: Visualizing financial trends using Matplotlib, Seaborn, or similar tools.
- **Feature Engineering**: Preparing data for machine learning by creating and selecting important features.

Deliverables:

- Financial data reports, including statistical summaries and visualizations.
- Well-prepared datasets with engineered features for use in the upcoming ML phase.

Team Tasks:

- General members work on small-scale EDA projects or create visualizations from their datasets.
- Core team members perform an in-depth analysis of the main financial dataset, producing reports and preparing data for machine learning.

Phase 3: Machine Learning & AI (February to Early April)

Goal: Apply machine learning algorithms to solve finance-related problems such as predicting stock prices or classifying financial trends.

Key Topics:

- **Supervised Learning**: Algorithms like linear regression, decision trees, and random forests to predict stock prices or other financial metrics.
- **Unsupervised Learning**: Clustering and anomaly detection in financial data (e.g., grouping stocks by performance).
- **Deep Learning (Optional - if time permits)**: Time-series forecasting using neural networks for more advanced financial predictions.

Deliverables:
- Machine learning models that predict financial trends (e.g., stock prices) or identify patterns in the data.
- Performance reports on the models, including accuracy and error rates.

Team Tasks:

- General members work on simpler ML models or assist in building models for individual finance datasets.
- Core team members develop more complex machine learning models based on the main dataset, documenting results and refining models as needed.

General Logistics & Work Distribution:

- General Members: Will work on smaller, individual projects to develop skills in data manipulation, analysis, and machine learning throughout each phase.
- Core Team Members: Will work on the main finance project, contributing to every phase (Data Engineering, Analytics, and ML) with a focus on specific responsibilities like building pipelines, analyzing data, or developing models.

Work Review and Progress Monitoring:

Core team members will regularly check the progress of general members and provide feedback. Core members will review deliverables such as cleaned datasets, analysis reports, and ML models to ensure quality and consistency with project goals.

Final Presentation (April):

The final presentation will showcase the results of our project, including:
- The data pipeline we built.
- Insights from data analysis.
- Results from our machine learning models.