Final Project Proposal

Project Title: Customer Insights & Prediction System

Project Description:

The project aims to build a complete Data Engineering and Machine Learning pipeline that collects, processes, and analyzes customer data to generate meaningful insights and predictions. The system will help businesses understand customer behavior, identify patterns, and predict future trends to improve decision-making and marketing strategies.

Group Members & Roles:

Team Leader	**Youssef Ali Hegazy – Machine Learning**
Member	Osama Essam Azab – Data Preprocessing
Member	Mohammed Hani Ebraheem – Data Collection
Member	Ahmed Tarek Elmenoufy – Database
Member	Tomas Amir Gerges – Airflow
Member	Mostafa Sobhy Mahmoud – Data Warehouse

Objectives:

- Collect and integrate customer-related data from multiple sources.
- Clean and preprocess the data for better quality and accuracy.
- Store and manage data efficiently using SQL Server and a Data Warehouse.
- Automate data workflows using Apache Airflow.
- Build predictive Machine Learning models to analyze customer behavior.
- Visualize KPIs and insights using interactive dashboards.

Tools & Technologies:

Python, SQL Server, Apache Airflow, Power BI, Data Warehouse, Machine Learning Models

Milestones & Deadlines:

Milestone	**Description**	**Deadline**
Data Collection	Gather customer datasets from various sources	15/10/2025
Data Preprocessing	Clean and transform raw data	20/10/2025
Database Integration	Design and implement SQL schema	25/10/2025
Data Warehouse Setup	Build and populate the warehouse	30/10/2025
Airflow Pipeline	Automate ETL workflow	05/11/2025
Machine Learning Model	Train and evaluate prediction models	10/11/2025
Dashboard Creation	Visualize KPIs and insights	15/11/2025

Final Presentation	Submit report and presentation	12/12/2025

KPIs (Key Performance Indicators):

1. Data Preprocessing (Python script, cleaned CSV)
• % of missing/duplicate data correctly handled (Target: 100%)
• Script efficiency (execution time within expected threshold)

2. SQL Integration (Schema, queries)
• Query accuracy (Target: ≥ 95%)
• Query performance (average execution time under X sec)

3. Visualization (Charts, dashboard)
• Dashboard load time (Target: < 3 sec)
• % of required KPIs/metrics visualized (Target: ≥ 90%)

4. Presentation (Report, slide deck)
• Report completeness (Target: 100%)
• Stakeholder clarity/feedback score (Target: ≥ 4/5)