CSCI 32092 | Data Mining and Data Warehousing

Assignment 02 - Project

Weightage: 20% of the total course grade
Submission Deadline: [15th May 2025]

Objective

Design and implement a data warehouse schema (Star or Snowflake as suitable) for a retail dataset, perform necessary preprocessing and ETL, and apply data mining techniques to generate business insights. The data should be stored in Snowflake (or another warehouse platform) and used directly for data mining tasks.

Instructions

Use the given "sales-data.csv" file to complete your project.

Step 1: Understand and review the dataset.
Deliverables:

- Data understanding summary
- Initial data quality report

Step 2: Data cleaning, preprocessing, and feature engineering.
Deliverables:

- Cleaned dataset
- Python script or notebook

Step 3: Shema Design.
Deliverables:

- Justification for schema choice
- Schema Diagram
- ER Diagram showing table relationships

Step 4: Implementing the in Snowflake (or Similar Warehouse).
Create tables in Snowflake according to the schema design and load cleaned data into the appropriate tables.
Deliverables:

- SQL scripts for schema creation
- Screenshots or logs showing successful data load

Step 5: Data Mining Tasks (Using Data from Warehouse).
Use necessary datamining techniques to complete below tasks.

- Predict Shipment Status

- Predict Sales Forecast
- Predict Order Profitability
- Cluster Orders by Buying Behavior - Group orders/customers by similar buying/shipping patterns
- Visualize shipment performance across geographies
- Analyze delivery delays or profit trends by latitude/longitude

Deliverables:

- Python script or notebook with:
  - ✓ Data pulled from warehouse
  - ✓ Feature engineering
  - ✓ Data mining techniques, evaluation
  - ✓ Evaluation metrics: Accuracy, RMSE, etc.
  - ✓ Graphs and visualizations
  - ✓ Brief explanation of findings

Bonus steps:

- Integrate a BI tool (example: Power BI) to analyze and visualize data to gain insights
- Use Snowflake's Snowpark for featuring one of the datamining tasks in step 5 (if you are using any other warehousing software utilize the features available for ML tasks)
- Try adding a real-time element using Streams and Tasks to ingest a stream of data to your warehouse and prepare for analytics.