UNIVERSITI MALAYA
Faculty of Computer Science and Information Technology
WIE 3007 – Data Mining (2025/2026 – Semester 1)
Group Project Specification

Group Size: Maximum 3 Members

Due Week 13

Objective

Apply Data Mining and AI-enhanced analytics in the financial or business domain using Generative AI (GenAI), Large Language Models (LLMs), and Small Language Models (SLMs). Students must perform dataset simulation, feature engineering, predictive modelling, and model interpretation with AI support.

GitHub Usage Requirement (Compulsory)

To ensure fairness and transparency, each group member must actively use GitHub for version control. The instructor will verify individual contributions through commit logs.

Each member must:
- Make at least 6 commits between Weeks 7–13
- Create and contribute through their own branch
- Write meaningful commit messages
- Contribute to at least one notebook section and one modelling task
- Review or comment on at least one pull request

Groups failing to demonstrate individual contributions may receive mark deductions.

1️⃣ Dataset Simulation & Feature Engineering (4 marks)

Simulate or generate at least 1000 financial or business-related records. Use GenAI to create realistic numerical and textual patterns. Use LLMs/SLMs for feature extraction such as sentiment, risk categorization, or customer segmentation.

2️⃣ Predictive Model Development (5 marks)

Develop predictive models (classification or regression) using algorithms such as Random Forest, Logistic Regression, XGBoost, or Neural Networks. Use AI tools for text-based feature engineering. Compare results across at least two models.

3️⃣ Model Evaluation & Interpretation (4 marks)

Evaluate models using Accuracy, F1-score, ROC-AUC, RMSE, or other suitable metrics. Use LLMs to summarise findings, provide insights, and interpret feature importance.

4️⃣ Final AI-Assisted Report (2 marks)

Prepare a 5–7 page report including objectives, dataset details, EDA, feature engineering, modelling, results, business insights, AI usage disclosure, and a GitHub contribution summary.

Deliverables

1. GitHub Repository (Compulsory)
2. Jupyter Notebook
3. Final Report (PDF/DOCX)
4. Recorded Project Presentation

Mark Distribution (15 marks)

Component	Marks	Description
Dataset Simulation & Feature Engineering	4	Financial/business dataset + AI-enhanced feature extraction
Predictive Model Development	5	Model building, comparison, GitHub contributions
Model Evaluation & Interpretation	4	Correct metrics + LLM-supported insights
Final Report	2	Clarity, professionalism, AI disclosure

Learning Outcomes

- Apply data-mining workflows in real financial/business contexts.
- Integrate GenAI/LLMs/SLMs into data analysis.
- Build and evaluate predictive models.
- Work collaboratively using professional GitHub practices.