WQD 7005 - 2024/2025 S2

Assignment (Due: Week 7)

Objective:

Perform exploratory data analysis (EDA) and advanced data preprocessing on simulated patient data leveraging Generative AI (GenAI), Large Language Models (LLMs), and Small Language Models (SLMs). The dataset will cover six vital signs (oxygen saturation, heart rate, temperature, blood pressure, weight, and blood glucose), questionnaire responses, and timestamps.

Tasks:

1. **Dataset Simulation using GenAI (3 marks)**
   o Simulate a dataset representing 500 patients monitored over 1 month. Utilize GenAI to produce realistic numerical variations in vital signs and generate plausible textual questionnaire responses or clinical notes, incorporating scenarios with missing data.
2. **Exploratory Data Analysis (EDA) enhanced by LLMs (4 marks)**
   o Conduct comprehensive exploratory data analysis using visualizations and statistical summaries.
   o Utilize Large Language Models (e.g., GPT-4) to interpret complex patterns, automatically summarize findings, identify trends, anomalies, and provide clinically relevant insights.
3. **Advanced Data Preprocessing utilizing SLMs/LLMs (4 marks)**
   o Implement preprocessing techniques, including intelligent missing value handling, normalization, and categorical encoding.
   o Apply Small Language Models or fine-tuned LLMs to handle textual data preprocessing tasks, such as classifying questionnaire responses, sentiment analysis, or textual data imputation.
4. **AI-Assisted Summary Report and Visualization (4 marks)**
   o Prepare a short, insightful report (2-3 pages) summarizing findings, preprocessing techniques, and key insights from the analysis.
   o Leverage LLMs to draft clear, coherent explanations for visualizations and data-driven insights.

Deliverables:

- Jupyter notebook with clearly documented steps, code explanations, and AI-generated insights.
- AI-assisted short summary report, including key visualizations and findings.

Mark Schema (Total: 15 marks):

- Dataset Simulation with GenAI (3 marks)
- EDA with LLM-generated insights (4 marks)
- Advanced Data Preprocessing using SLMs/LLMs (4 marks)
- AI-assisted Summary Report and Visualization (4 marks)