AISE3010: Final Project

<u>Due date</u>: April 4th, 2025, by 11:55 pm Eastern Time.

<u>Total marks</u>: 15

<u>Late penalty</u>: Late assignments will <u>not</u> be accepted after 11:55 pm on April. 11th, 2025, and a mark of zero will be given.

This final project will be done **individually or in a group (maximal three students in a group),** and the final mark for the assignment will be assigned to each of you by our Teaching Assistant. Note that I allow you to keep resubmitting until the deadline. Please get in touch with TA if you are unable to resubmit your files.

If you are in a group, please submit only ONE assignment document in OWL but includes ALL team members' name. Otherwise, one of your team members will not receive marks.

OBJECTIVES

The overall objective of your project is to demonstrate your ability to leverage Google Cloud Platform (GCP) tools and machine learning (ML) techniques to solve a real-world relational dataset problem.

Marking regulation

1) Dataset Selection (2 mark):
   a) Relational Dataset: Students must use a dataset that is stored in a relational database format (e.g., multiple tables with relationships like one-to-many or many-to-many). Example: A dataset with separate tables for customers, orders, and products.
   b) Minimum Size: The dataset should have at least 10,000 rows and 3+ tables to ensure complexity.
   c) Data Relationships: The dataset must require **JOIN operations** to combine tables for analysis.
   d) Include features suitable for a classification problem. Examples of complex datasets include e-commerce (predict churn/returns), healthcare (predict diagnosis/readmission), financial (predict loan default/fraud), and social media (predict engagement/popularity), each with multiple relational tables requiring SQL handling.
2) GCP Data Processing (5 mark):
   a) Upload the dataset to Google Cloud Storage and import it into BigQuery
   b) Provide the SQL scripts used for preprocessing and analysis
   c) Clean and Preprocess Data with SQL. Such as, handling missing values, removing duplicates, data transformation.
   d) Feature Engineering: Create features relevant to your classification problem using SQL. Remove the features that are not relevant to the task.
   e) Export your cleaned and transformed data from BigQuery to a format suitable for ML training (split training and testing data).
3) Training Workflow (5 mark):
   a) Custom Model with TensorFlow or PyTorch then use Vertex AI custom training job or Colab (I would suggest not use a super large model and use transfer learning).
   b) Use Vizier for hyperparameter tuning.
   c) Train an AutoML model and compare the performance compared to your custom model hyperparameter finetuned result. Check metrics like precision, recall, and feature importance and training efficiency.

4) Final Report (3 mark):

   a) Explain your workflow: Data → SQL processing → Model → Deployment.

   b) Include SQL snippets, model metrics, and screenshots of the data, training visualization.

Submission

- Submit all **the code**.
- Submit a **PDF** report that shows your workflow, SQL snippets, model metrics, and screenshots of the data, training visualization.

Challenges & Tips

- **Cost Management:** Use smaller datasets for training to avoid high GCP costs. You can train the custom model on Colab if the credit is not enough.
- **Model Drift:** Mention how you'd monitor model performance over time.
- Stop or delete your instance if you are not using it.

All code and PDF report requested below must be submitted using OWL. To provide answers via OWL:

1. **You should log into OWL and access the course web site accordingly.**
2. **Select the "Assignments" tool.**
3. **From the page that comes up, select "Final Project".**
4. **You will now reach the submission page for Final Project. Attach the files you did.**