**WQD7005 Data Mining 2024/2025 Semester 1**

**Group Project Overview:**

This group project involves applying data mining methodologies in real-world applications using the **SAS SEMMA** methodology (Sample, Explore, Modify, Model, Assess). The project emphasizes modern data mining tools, focusing on minimal coding requirements and utilizing **no-code** or **low-code** platforms. Each group should consist of **at most 4 members**.

**Assignment Instructions:**

**Step 1: Dataset Selection (Sample Stage)**

* **Simplified Approach:** Select an accessible dataset from platforms like **Kaggle Datasets** or **Google Cloud Public Datasets**. Choose a pre-structured dataset to avoid the complexity of data integration and API setup.
* **Automated Dataset Exploration:** Use tools such as **Pandas Profiling** or **Sweetviz** to automatically explore and summarize the dataset. These tools generate reports and insights without requiring code.

**Step 2: Exploratory Data Analysis (Explore Stage)**

* **Simplified Approach:** Instead of coding visualizations, use **Google Data Studio**, **Tableau Public**, or **Power BI** to create interactive visualizations using simple drag-and-drop interfaces.
* **Automated EDA:** Use tools like **AutoViz** or **D-Tale** to automatically generate exploratory data analysis reports. These tools require minimal coding and allow you to focus on interpreting the results.

**Step 3: Data Preprocessing (Modify Stage)**

* **Simplified Approach:** Use **Google Cloud AutoML**, **DataPrep**, or **H2O.ai** to automate data cleaning and feature engineering. These platforms handle missing values and outliers automatically.
* **Dimensionality Reduction:** If the dataset is large, use AutoML's built-in dimensionality reduction features (e.g., **PCA**) to simplify the data without manual coding.

**Step 4: Modeling (Model Stage)**

* **Simplified Approach:** Use **Google Cloud AutoML**, **H2O.ai**, or **DataRobot** for automated model training. Upload the cleaned dataset, and the platform will handle model selection and training.
* **Pre-trained Models:** If applicable (e.g., for text or image data), use pre-trained models like **BERT** or **ResNet** available through AutoML. These models reduce the need for extensive training.
* **Automated Hyperparameter Optimization:** Leverage AutoML’s built-in hyperparameter tuning features to optimize models without needing to write code.

**Step 5: Model Evaluation (Assess Stage)**

* **Simplified Approach:** Use built-in evaluation tools in **AutoML** or **H2O.ai** to automatically assess model performance (e.g., **accuracy**, **precision**, **F1-score**).
* **Model Interpretability:** Use AutoML’s built-in tools like **SHAP** to interpret and explain model predictions visually. These features simplify the interpretation process, especially for complex models.
* **Model Fairness:** Where applicable, check for bias using tools like **Fairlearn**, integrated into AutoML platforms.

**Step 6: Deployment and Monitoring (Bonus - Optional)**

* **Simplified Approach:** Use **Google AI Platform** or **AWS SageMaker** for one-click model deployment, which simplifies the process of creating an API from your model.
* **Optional Monitoring:** Use built-in monitoring tools in Google AI Platform or AWS SageMaker to track model performance over time without needing to manually configure monitoring systems.

**Step 7: Presentation and Visualization**

* **Simplified Approach:** Use **Google Data Studio**, **Tableau Public**, or **Power BI** to build interactive dashboards. These no-code tools allow for real-time visualizations, which can be easily shared with stakeholders.
* **Key Metrics to Visualize:** Focus on presenting key insights such as **sales trends**, **customer segmentation**, or **product performance** through simple visualizations like **bar charts**, **pie charts**, and **trend lines**.

**Step 8: Collaborative and Reproducible Workflows**

* **Simplified Approach:** Use **Google Colab** or **Jupyter Notebooks** for collaborative work. These platforms allow team members to work together in real-time without needing Git or complex version control systems.
* **Cloud Resources:** If the dataset or model is large, utilize free-tier resources from **Google Cloud** or **AWS** for computational tasks.

**Evaluation Criteria:**

1. **Data Understanding & Preparation (25%)**
   * **Dataset Selection (5%)**: Did the group select a relevant and well-structured dataset from an accessible source like Kaggle or Google Cloud Public Datasets?
   * **Data Preprocessing (10%)**: Did the group use automated tools (e.g., AutoML, DataPrep) to effectively clean the data, handle missing values, and perform feature engineering?
   * **Exploratory Data Analysis (10%)**: Did the group generate meaningful insights through automated EDA tools and simple visualizations?
2. **Modeling and Methodology (35%)**
   * **Model Selection (10%)**: Was an appropriate model selected and trained using automated tools like AutoML? Were the model choices suitable for the dataset and problem type?
   * **Model Training (10%)**: Was the model training process automated effectively? Was hyperparameter tuning handled by the platform?
   * **Feature Engineering (5%)**: Did the group use AutoML’s feature engineering tools to improve model performance?
   * **Advanced Techniques (10%)**: Did the group utilize pre-trained models or advanced techniques (like AutoML’s built-in algorithms) to enhance the modeling process?
3. **Model Evaluation & Interpretation (20%)**
   * **Evaluation Metrics (10%)**: Were the model’s performance metrics (e.g., accuracy, precision, F1-score) clearly presented using AutoML’s evaluation tools?
   * **Model Interpretability (5%)**: Did the group use built-in tools like SHAP to explain how the model made predictions?
   * **Limitations and Assumptions (5%)**: Were the model’s limitations and assumptions clearly discussed?
4. **Presentation & Reporting (20%)**
   * **Clarity of Report (10%)**: Was the report structured logically, with clear explanations, automated outputs, and insights presented in an easy-to-understand way?
   * **Visualizations (5%)**: Were no-code visualizations (Google Data Studio, Tableau) effectively used to present key insights?
   * **Conclusion and Recommendations (5%)**: Did the group provide clear conclusions with actionable recommendations based on their findings?
5. **Collaboration & Innovation (15%)**
   * **Team Collaboration (5%)**: Was there balanced participation from all 4 team members, with clear contributions from each?
   * **Innovation (10%)**: Did the group go beyond basic methods by incorporating advanced tools (AutoML, DataPrep, pre-trained models, cloud platforms)?

**Group Size:**

**Each group should have a maximum of 4 members.**

**Deadline:**

**Week 14**