CDS6314 DATA MINING

Trimester 1, 2024/2025

ASSIGNMENT (20%)

INSTRUCTIONS:

1. This assignment carries 20% of the coursework assessment.
2. This is a group project, with a maximum of 4 members.
3. Deliverables for this assignment include Python code (*.ipynb*) and a report (*.pdf*).
4. Submission deadline: **5th January 2025 (Sunday), 11.59pm, Week 9.**
5. Late-Day policy applies (20% deduction per day late from deadline).
6. If ***plagiarism*** is detected, the assignment will be granted 0% with no negotiation.

OBJECTIVE:

The aim of this assignment is to guide you through the process of understanding a dataset, performing preprocessing, building classification models, including ensemble learning techniques, and evaluation using cross-validation and various metrics.

Dataset: *Student_dataset*

PYTHON TASK:

This part involves analyzing a provided dataset to perform exploratory analysis, preprocessing, and classification tasks using various machine learning techniques. The objective is to evaluate and compare the performance of individual classifiers and ensemble methods. Steps should include, but not limited to the following:

A. Data exploration and preprocessing

1. Data Exploration (statistics and visualization).
2. Data Preprocessing (cleaning, normalization, missing data, duplication, etc.).

B. Classification Tasks

1. Implement the following classification algorithms:
    - Logistic Regression
    - Decision Trees
    - Support Vector Machines (SVM)
    - K-Nearest Neighbors (KNN)
    - Random Forest
2. Cross-Validation and evaluation:
    - Perform k-fold cross-validation (with k=5 or k=10) for all classifiers to ensure robust performance evaluation.
    - Report the following metrics based on cross-validation: Accuracy, Precision, Recall, F1-Score

C. Ensemble Methods

1. Implement a Bagging classifier (e.g., Bagged Decision Trees) and Boosting techniques such as AdaBoost, Gradient Boosting, and XGBoost.
2. Use the same cross-validation strategy as in the classification tasks and report evaluation metrics, including Accuracy, Precision, Recall, and F1-Score.
3. Perform a simple comparison with individual classification algorithms.

TECHNICAL REPORT:

Write a technical report to introduce the domain including related research, and compile the details and results of the python task. The report should include the following items:

1. **Cover page:**
   a. Title
   b. Group number and members
   c. Contribution of each member
2. **Introduction**
   a. Introduce the background and motivations
   b. Review at least three related research papers discussing student performance prediction, e.g. "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques"
   c.
3. **Formulating Exploratory Questions**
   a. Create questions to explore and find out potentially frequent patterns in the dataset
   b. Explain and justify the potential subjective interestingness of the outcomes from answering the formulated questions
4. **Data Exploration**
   a. Describe the dataset
   b. Explain the initial exploration done the data by showing statistics and visualizations
5. **Data Preprocessing**
   a. Describe the data cleaning steps (if any) and justify the approach
   b. Describe the steps taken to transform the raw data into form suitable for data mining including justification
6. **Classification**
   Describe the classification algorithms applied and ensemble methods.
7. **Results Discussion**
   a. Discuss the results generated from association rule mining.
   b. How do the results answer the formulated exploratory questions?
   c. Are there certain factors in the preprocessing that influence the rule generation?
   d. What are the attributes affecting the classification result, listed in the descending order? Justify your findings.
8. **Conclusion**
   a. Summarize the overall findings of the work
   b. Discuss potential use case or importance of the findings
   c. Suggest potential future directions of the work (e.g. how to overcome limitations, other dimensions of exploration, etc.)
9. **References**

SUBMISSION:

Submit the following files to Ebwise:

- Python notebook codes (.ipynb)
  - Please ensure the codes can reproduce the results given the raw data
  - Include a Readme.txt of instructions to navigate the notebooks (if necessary, especially if multiple notebooks)
- Technical report (.pdf)

Note:

- *Name the submission files with your Group number (e.g. Group01.ipynb, Group01.pdf).*
- *Only group leaders need to submit. Submission by group members will not be marked. (The first name in the group list is the group leader).*
- *Do NOT submit zip files.*

MARKS DISTRIBUTION:

Code (12%)	Data Exploration	2
	Data Preprocessing	2
	Classification	4
	Ensemble Methods	3
	Visualizations	1
Report (8%)	Introduction and Literature Review	2
	Questions Formulated	2
	Analysis of Findings	3
	Conclusion	1
Total		**20%**