Assignment 2: Predict Students Dropout and Academic Success

Assigned: 28/11/2024 **Due: 26/12/2024**

Overview

Welcome to the Student Performance and Retention Prediction Competition! In this challenge, you will work with a real-world dataset to predict student outcomes, such as whether a student will drop out, remain enrolled, or graduate. The competition is designed to allow you to practice Exploratory Data Analysis (EDA), visualization, and machine learning to gain insights and build predictive models.

This dataset comes from a higher education institution and contains multiple features related to:

- Demographic Information: Age, gender, socio-economic background.
- Academic Performance: Grades and results from the first two semesters.
- Behavioral and Social Data: Attendance records, participation in extracurricular activities.

You will use this dataset to predict one of the following outcomes for each student:

- Dropout: The student has withdrawn from the program.
- Enrolled: The student remains enrolled in the program but has not graduated.
- Graduate: The student has successfully completed the program.

Competition Entrance

https://www.kaggle.com/t/728f638f02854aab8f16be47cbeaef6b

Tasks 1 (60 Marks)

1. Create an account at https://www.kaggle.com/. You MUST set a **TEAM NAME** with the format in (Student ID). Example: 2136673.

2. Create a notebook on Kaggle. Conduct Exploratory Data Analysis and data processing, train and validate your models, and generate your 'submission.csv' on test data using the notebook on Kaggle. Add necessary comments to your notebook. Download the notebook from Kaggle and **submit it to Learning Mall,** with the name in (Name_Student ID). (*30 Marks*)

3. Submit your predictions ('submission.csv') for the test solution to Kaggle. Also, you are required to include your Kaggle score in your report (see below in Task 2). (*30 Marks*)

Tasks 2 (40 Marks)

Write a 1-page report, which **must contain** 2 or 3 tables or figures.

- Name your report with Name_Student ID.
- **Submit your report to Learning Mall**.

The report must cover:

- **Introduction:** (4 *Marks*)
  What is the background of this project? How is it related to Big Data?
- **Methodology:** (*8 Marks*)
  A. Data Preprocessing
  What are the steps of data pre-preprocessing explored before training? Data visualization, data cleaning and reduction, normalization and discretization, feature selection, imbalanced data, etc. No need to cover all of them.
  B. Classification Algorithm
  How does it work? Explain the algorithm or framework.
- **Results:** (*10 Marks*)
  Are there other competitor models for this project? How does it compare to your technology?
- **Discussion:** (*4 Marks*)
  What are the good aspects, and what are the bad aspects? Be sure to add a

sentence on "**contributor thoughts:**" What are your own unique thoughts on the pros and cons of the technology? Do you envision an extension that might be helpful?

- **Conclusion:** (*4 Marks*)
  Summarize the 2 to 4 points you think are most important.

Concise, information-rich content. For each of the sections above, you will not simply be graded on having content but on the quality of the content and how well it answers the questions in concise, clear, and engaging terms.

Style. (*10 Marks*)
In order to make your report consistent and visually appealing, as well as to make the evaluation of your work fairly, each page should be conformed to the following specifications:

- Margins: approx. 0.5" on all 4 sides.
- Columns: 2 with approx. 0.3in margin; justified text
- Fonts:
  - Body text: Times New Roman, 11pt.
  - Section headings: Calibri 13pt bold-Italic
  - Within captions, tables, figures, or images: Calibri 9-11pt.
- Line Spacing:
  - Body text: Single (1.0)
  - Section headings: 6pt spacing above heading

Academic Honesty. Copying chunks of code or problem-solving answers from other students, online or other resources is prohibited. You are responsible for both (1) not copying others' work, and (2) making sure your work is not accessible to others. Assignments will be extensively checked for copying of others' work. Problem-solving solutions are expected to be original, using concepts discussed in the book, class, or supplemental materials but not using any direct code or answers. Please see the syllabus for additional policies.