Team 2: Student Enrollment Data Analysis

1 Introduction

The dataset `2.csv` contains student enrollment data from a Peruvian university for the year 2023. It presents challenges such as missing values and inconsistent formatting, particularly in gender representation (1, 2, M, F, U), which requires data cleaning and transformation. The dataset allows for in-depth analysis of enrollment trends, tuition payment patterns, study preferences, and academic risk factors.

2 Dataset Description

The dataset includes the following attributes:

- **Enrollment** – Type of student enrollment:
  - **New:** Student enrolling for the first time.
  - **Re-enrolled:** Student continuing studies without interruption.
  - **Reinstated:** Student returning after a period of inactivity.
- **Tuition Payment March 2022** – Indicates whether the student paid tuition in March 2022 (0 = No, 1 = Yes).
- **Tuition Payment March 2023** – Indicates whether the student paid tuition in March 2023 (0 = No, 1 = Yes).
- **Gender** – Student's gender (M, F, U, 1 $\rightarrow$ (M), 2 $\rightarrow$ (F)).
- **Program/Major** – Academic program or major the student is enrolled in.
- **Shift/Schedule** – Study schedule (Morning, Afternoon, Night, Mixed).
- **Study Mode** – Study modality:
  - **On-site:** Classes held at a physical campus.
  - **Online:** Fully online classes.
  - **Remote:** Online classes with some in-person activities.
  - **To be determined:** Study mode not yet selected.
- **Age Range of Enrolled Student** – Age range of enrolled students.
- **Department** – Department where the student resides or studies.
- **Province** – Province where the student resides or studies.
- **District** – District where the student resides or studies.
- **Type of Educational Institution** – Type of institution the student comes from (School, Institute, etc.).
- **Institution Status** – Status of the institution (Public or Private).
- **Benefit Discounts** – Indicates whether the student receives any financial benefits or discounts.
- **Number of Enrolled Courses** – Number of courses the student is enrolled in.
- **At-Risk Course** – Indicates whether the student has courses at risk of failure.

3 Tasks and Requirements

To analyze and extract meaningful insights from the dataset, the following tasks are required:

3.1 Data Exploration and Preprocessing

- Load and review the dataset.
- Handle missing values and clean inconsistent data
- Perform exploratory data analysis (EDA) to identify trends and anomalies.
- Normalize numerical features and encode categorical variables.
- Split the dataset into training and test sets for further analysis.

3.2 Data Analysis and Modeling

- Analyze enrollment trends based on different factors such as gender, major, and study mode.
- Compare tuition payment behaviors between 2022 and 2023.
- Identify the most popular academic programs among different demographics.
- Investigate the relationship between the number of enrolled courses and academic risk.

3.3 Visualization and Reporting

- Use appropriate visualizations (bar charts, histograms, heatmaps) to present insights.

4 Submission Requirements

- A well-structured report detailing the methodology, results, and analysis in a given report format.
- Python code is used for implementation.
- A presentation summarizing key findings and recommendations in a given presentation format.