Indian Institute of Technology, Dharwad

|| सा विद्या या विमुक्तये ||
ಭಾ.ತಂ.ಸಂ. ಧಾರವಾಡ
भा. प्रौ. सं. धारवाड
I.I.T. DHARWAD

CS209 : Artificial Intelligence
And
CS214 : Artificial Intelligence Laboratory
Project Report :**Data Exploration and Preprocessing Report: Student Enrollment Data Analysis**

Course Instructor:
Dr. Dileep A.D.
Mentor Name:
ABCD
Submitted by:

1. abcd
2. efgh
3. pqrs

4. wxyz

Contents

List of Figures

List of Tables

1 Introduction

The "Data Exploration and Preprocessing Report: Student Enrollment Data Analysis" project focuses on analyzing a dataset (`2.csv`) containing 37,582 records from a Peruvian university for the year 2023. This dataset encompasses critical student-related information, including enrollment types, tuition payment statuses, demographic details (e.g., gender), academic metrics (e.g., number of enrolled courses and at-risk courses), and geographic data. The project is driven by the need to address challenges such as missing values, inconsistent data formats (e.g., gender represented as M, F, U, 1, 2), and the presence of irrelevant or incomplete attributes, which hinder effective analysis and modeling.

The primary goal of this project is to preprocess and explore the dataset to prepare it for predictive modeling, with a specific focus on predicting tuition payment behavior for March 2023. This involves cleaning the data by handling missing values and inconsistencies, encoding categorical variables, and performing exploratory data analysis (EDA) to uncover trends and patterns. The project aims to provide actionable insights that can support university administrators in improving student retention, optimizing financial planning, and identifying at-risk students. Additionally, it seeks to establish a robust foundation for applying machine learning techniques, such as classification models, to address real-world educational challenges.

This project is important because it uses AI to help improve decision-making in education. By organizing raw data into a format that's easy to analyze, it helps identify the main factors that affect student success and financial involvement. The results can benefit not just the Peruvian university, but also other universities around the world, by providing a useful way to manage student enrollment data.

2 Data Handling

This section covers the preprocessing and exploration of the dataset.

2.1 Dataset Overview

The dataset contains **37,582 records** with **21 attributes** related to student enrollment. Key attributes include:

- **Enrollment**: Type (New, Re-enrolled, Reinstated).
- **Tuition Payment March 2022/2023**: Binary (0 = No, 1 = Yes).
- **Gender**: Gender (M, F, U, 1, 2).
- **Number of Enrolled Courses**: Course count.
- **At-Risk Course**: Courses at risk of failure.

Loaded with `pandas.read_csv("./2.csv", delimiter=";")`, initial inspection via `df.head()` revealed inconsistencies like gender formatting.

Attribute	Description	Type	Example Values
Enrollment	Type of enrollment	Categorical	Nuevo, Reincorporado
Tuition Payment March 2023	Paid tuition (0 = No, 1 = Yes)	Binary	0, 1
Gender	Student gender	Categorical	M, F, 1, 2, U
Number of Enrolled Courses	Count of enrolled courses	Numerical	0 to 6

Table 2.1: Dataset Overview

Column	Missing Count	Percentage (%)
Gender	2	0.005
Type of Educational Institution	21,714	57.77
Educational Institution	19,370	51.54
Institution Status	21,714	57.77
Department	736	1.96
Province	736	1.96
District	736	1.96
Classification	1	0.003
Faculty	1	0.003
Program/Major	1	0.003
Shift/Schedule	58	0.15
Age Range of Enrolled Student	4	0.01

Table 2.2: Missing Value Counts (Initial Dataset)

2.2 Handling Missing Values

Missing values were identified using `checknulls()`.

Actions:

1. **Gender:** Replaced 'U' with `np.nan` and dropped rows with missing gender (203 total). Standardized: 'M' and '1' → 1, 'F' and '2' → 2.

2. **Critical Columns:** Dropped rows with missing values in `CLASSIFICATION`, `FACULTY`, `PROGRAM/MAJOR`, `GENDER`, `AGE RANGE`, `DEPARTMENT`, `PROVINCE`, `DISTRICT`, and `SHIFT/SCHEDULE`. Reduced dataset to **36,584 records** (2.7% loss).

3. **Irrelevant Columns:** Excluded `TYPE OF EDUCATIONAL INSTITUTION`, `EDUCATIONAL INSTITUTION`, and `INSTITUTION STATUS` due to high missing rates (> 50%) .

Reason for Dropping Rows Instead of Filling with Mode: Rows with missing gender were dropped rather than filled with the mode to avoid introducing bias. Filling with the mode could skew the gender distribution, misrepresenting true trends and potentially affecting model accuracy.

Findings:

- Row deletion had minimal impact (2.7% loss).
- Irrelevant columns were removed, focusing on key enrollment data.

2.3 Feature Encoding

Categorical variables were encoded for modeling.

Actions:

- Applied LabelEncoder to categorical columns (excluding numerical and GENDER).
- Manually encoded GENDER: {'M':1, 'F':2, '2':2, '1':1}.

Column	Unique Values	Example (Before → After)
Enrollment	3	Nuevo → 0, Reingresado → 1
Department	25	LIMA → 14
Province	165	LIMA → 98
District	694	BRENA → 124
Classification	5	[Varies] → 1
Campus	14	UTP Lima Centro → 6
Faculty	8	Fac. Ing. Sist. Y Elect. → 5
Program/Major	83	ING. DE SISTEMAS → 40
Shift/Schedule	4	NOCHE → 1, MAÑANA → 0
Benefit Discounts	7	SIN BENEFICIO → 4
Study Mode	3	Presencial → 0, Online → 1
Age Range of Enrolled Student	5	4. 24-29 → 2
Disability	2	No → 0, Yes → 1
Gender	2	M → 1, F → 2

Table 2.3: Categorical Columns and Unique Values (Post-Dropping Rows)

Findings:

- Encoding enabled numerical analysis.
- High unique values in PROGRAM/MAJOR and DISTRICT suggest diversity.

2.4 Exploratory Data Analysis (EDA)

EDA explored feature distributions and relationships.

2.4.1 Box Plot

- Finding: The plot shows medians of 1 for tuition payments, 2 for enrolled courses, and 0 for at-risk courses; outliers up to 5 for at-risk courses are more frequent among females.

2.4.2 Violin Plot

- Finding: The plot reveals both genders peak at paid (1), with females showing a wider spread toward unpaid (0), suggesting greater payment variability among females.

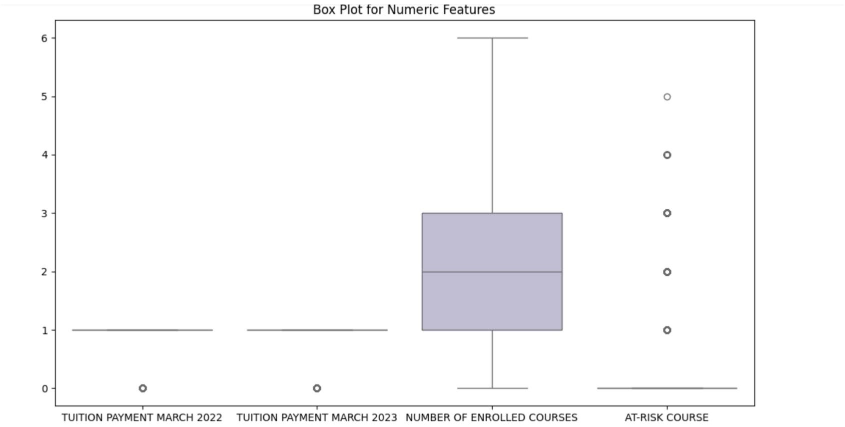

Figure 2.1: Box Plot of Numeric Features by Gender

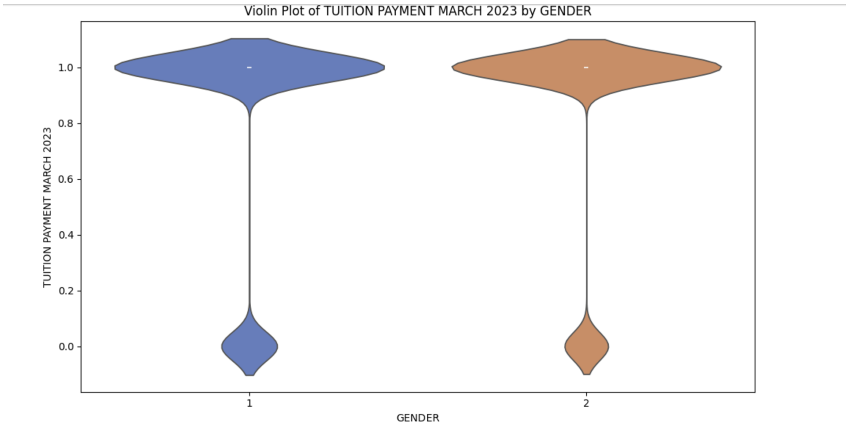

Figure 2.2: Violin Plot of Tuition Payment March 2023 by Gender

2.4.3 Count Plot

- Finding: The plot shows 20,607 males and 15,977 females, indicating a slight male majority (56.3% vs. 43.7%).

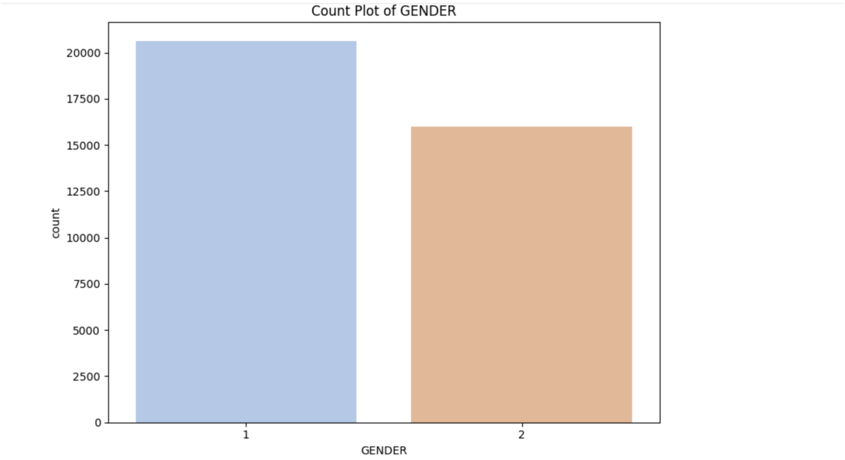

Figure 2.3: Count Plot of Gender

2.4.4 Correlation Heatmap

- Finding: The heatmap indicates a strong positive correlation (0.923) between tuition payments across years, weak links (0.037, 0.034) with enrolled courses, and negative correlations (-0.188, -0.200) with at-risk courses.

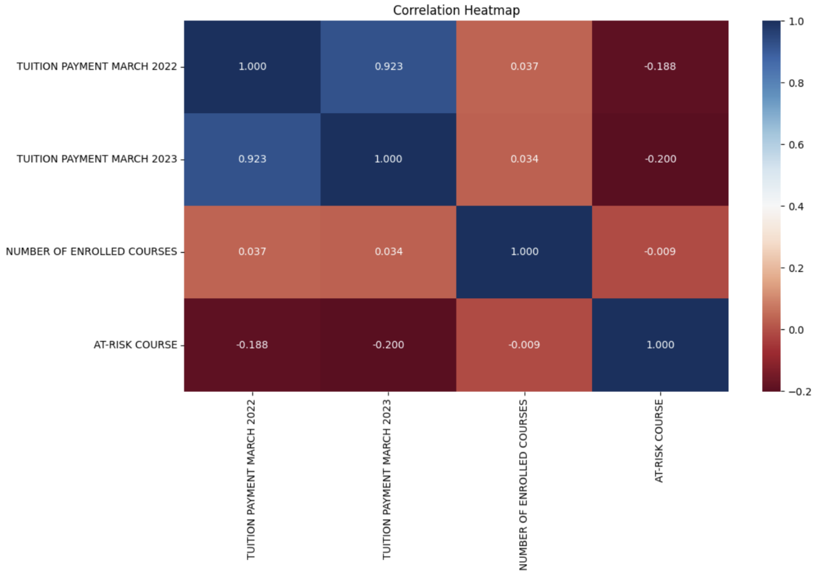

Figure 2.4: Correlation Heatmap

2.4.5 Distribution Plot

- Finding: The plot displays 84.1% paid (1) and 15.9% unpaid (0), confirming a skewed distribution favoring payment.

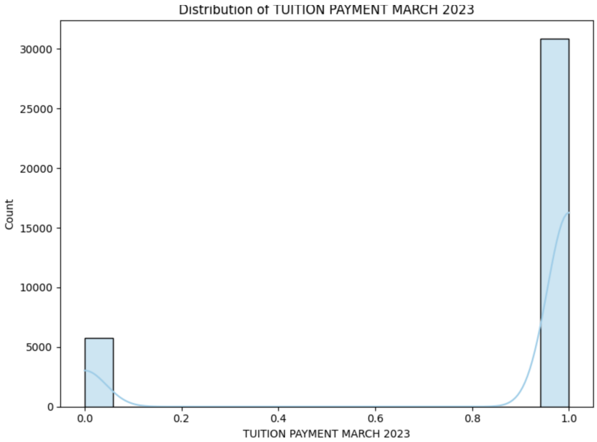

Figure 2.5: Distribution of Tuition Payment March 2023

Findings:

- Gender distribution favors males (56.32% vs. 43.68%).

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Tuition Payment March 2022	36,584	0.863	0.344	0	1	1	1	1
Tuition Payment March 2023	36,584	0.843	0.364	0	1	1	1	1
Number of Enrolled Courses	36,584	1.947	0.985	0	1	2	3	6
At-Risk Course	36,584	0.149	0.513	0	0	0	0	5

Table 2.4: Summary Statistics (Post-Preprocessing, n=36,584)

- Tuition payment is high (84.3% in 2023), with females showing slightly more unpaid cases.
- Enrolled courses average 1.95, with at-risk courses rare but more variable among females.
- Strong tuition payment correlation (0.923) suggests consistent behavior.

2.5 Data Splitting

The dataset was split into 80% training (29,267 records) and 20% test (7,317 records) sets, excluding `TUITION PAYMENT MARCH 2023` from features and using it as the target.

3 Conclusions

The preprocessing phase successfully transformed the student enrollment dataset into a robust, analysis-ready format. Key achievements include:

- **Data Quality**: Effectively removed 2.7% of rows with missing critical data and excluded irrelevant columns, while standardizing gender for consistency.
- **Insights**: Revealed high tuition payment compliance (84.3%), a slight male majority (56.3%), and a low incidence of at-risk courses, providing a foundation for understanding student behavior.
- **Encoding**: Converted all features to numerical formats, ensuring compatibility with modeling techniques.
- **Readiness**: The dataset is optimally split for predictive modeling and deeper analysis.