Data Preprocessing Report – Student Dropout Prediction

Name: Andy Ndubuisi Joseph

Company: 3Signetltd

Data Science Intern

1. Description of Data Cleaning Steps

1.1. Handling Missing Values

- Identification: I checked for missing values using `isnull().sum()` to locate columns with incomplete data but no missing value was recorded

1.2. Removing Duplicates

- Identification: I checked duplicate rows using the `duplicated().sum()` function. No missing value

1.3. Correcting Data Types

- Conversion: checked for columns with inconsistent data types to the appropriate formats:

- Ensured numerical columns were of numeric type (e.g., `int`, `float`).

1.4. Standardizing Column Names

- Cleanup: I standardized column names by removing extra spaces and special characters for consistency and readability (e.g., 'Nacionality’ was cleaned to 'Nationality').

1.5. Normalizing Data

- Scaling: I normalized or scaled numerical columns using MinMax Scaler to ensure all features were on a similar scale, which is crucial for algorithms sensitive to feature scales.

1.6. Encoding Categorical Variables

- Encoding: I encoded categorical variables using methods like label encoding to convert them into a numerical format suitable for analysis.

2. Summary of Data Quality Issues Encountered and How They Were Resolved

2.1. Column Name Inconsistencies

- Issue: Column names contained extra spaces and special characters.

- Resolution: I standardized and cleaned the column names to improve consistency.

3. Justification for Chosen Data Transformation Methods

3.1. Normalization

- Justification: Normalizing data allows features to contribute equally to the analysis, especially for algorithms sensitive to feature scales.

3.2. Encoding Categorical Variables

- Justification: Encoding categorical data into a numerical format is essential for machine learning models to process and learn from these features effectively.