DATA PREPROCESSING REPORT

By: Blessing Ilesanmi

3Signet Data Science Internship Week 1

13th September, 2024

INTRODUCTION

Data preprocessing is a crucial step in the data analysis pipeline, serving as the foundation for robust, insightful analysis and reliable machine learning models. Before diving into analytics, it is essential to prepare the raw data to ensure it is free from errors, inconsistencies, and irregularities that could skew the results or impede the model's performance. This report presents a comprehensive overview of the preprocessing steps performed on a dataset related to academic performance.

The dataset in question encompasses various fields that capture student demographics, previous academic qualifications, admission grades, curricular units (credits and grades), and other key attributes reflective of students' educational backgrounds and progress. Given the wide range of variables and their diverse data types—ranging from numerical values to categorical data—multiple preprocessing techniques were employed. These techniques include data cleaning, type conversions, handling missing data, dealing with outliers, and applying normalization and feature transformations.

The goal of this report is to detail each step of the preprocessing journey, explain the rationale behind the methods employed, and justify the decisions made to ensure the dataset was transformed into a clean, consistent, and analytically useful state. This preprocessing effort aims to facilitate subsequent analysis and model training, particularly in exploring factors that influence academic success or failure.

DESCRIPTION OF DATA CLEANING STEPS

1. **Data Loading:** The dataset was initially provided in an Excel file, with the raw data presented in a semicolon-delimited format. The data was imported into a pandas dataframe without specifying any headers. Upon inspection, it was found that the first row contained key column names that should serve as headers. To correct this, the first row was reassigned as the header, ensuring proper organization of the dataframe.

2. **Renaming Columns:** One key observation during the data inspection phase was the presence of a misspelled column: “Nacionality” instead of “Nationality.” To ensure consistency and uniformity, this column was renamed. Standardizing column names is essential, as misspellings or inconsistencies in naming conventions could lead to confusion during analysis or hinder the merging of multiple datasets.

3. **Data Type Conversion:** Upon loading, several columns were identified as “object” data types, even though they contained numerical data. This posed a challenge for analysis, as certain operations like aggregation or statistical analysis cannot be performed on non-numeric data types. The following data type conversions were carried out:

- Integer-based columns (e.g., "Previous qualification (grade)" and "Admission grade") were converted using the *.astype(int)* function.
- Floating-point numbers, which represent grades and other continuous variables, were converted using *.astype(float)*. Moreover, some numerical columns were rounded to

one decimal place to maintain consistency across the dataset, ensuring that grades and other metrics were accurately represented.

4. **Handling Missing Values:** One of the initial steps in preprocessing is to identify and handle missing values. Upon using the *.isnull().sum()* method to assess missing data across all columns, it was found that the dataset contained no missing values. This was a fortunate discovery, as missing data can complicate analysis by introducing biases or reducing the available dataset for modeling. As no imputation or removal of records was necessary, this step confirmed that the dataset was complete and ready for further transformations.

5. **Duplicate Data:** Another critical aspect of data cleaning is ensuring the uniqueness of records. Duplicate entries can skew the results of an analysis, artificially inflating certain metrics or leading to biased models. Using the *.duplicated()* method, the dataset was scanned for any duplicate rows, no duplicates were found, ensuring that the dataset was free from redundancy and each record represented a unique student's data.

6. **Handling Outliers:** Outliers in datasets, particularly those involving academic grades can significantly distort results, as extreme values may not accurately reflect typical performance. To address this, outliers were identified and capped using a method based on the mean and standard deviation. Specifically, any values beyond three standard deviations from the mean were capped at the upper or lower limits. This allowed for the reduction of the outlier effect while preserving the overall dataset.

   Boxplots were generated for numerical columns to visually inspect the presence of outliers. Outliers were primarily found in the columns:
   - Previous qualification (grade)
   - Admission grade
   - Curricular units 1st semester (grade)
   - Curricular units 2nd semester (grade)

   By capping the extreme values, I ensured that the analysis would not be unduly influenced by students with unusually high or low performance, while still retaining all records.

SUMMARY OF DATA QUALITY ISSUES ENCOUNTERED AND RESOLUTIONS

- **Unstructured Data:** The raw data was initially provided in a semicolon-delimited format, requiring preprocessing to restructure into proper columns for analysis.

- **Incorrect Headers:** The first row of the dataset was initially used as a regular data row. Upon inspection, this row was identified as containing header information, so it was reassigned as the header row.

- **Special Characters and Quotes:** One of the columns header contained values enclosed in double quotation marks (""). The dataset was reviewed and cleaned of extra quotes to ensure uniformity.

- **Inconsistent Column Names:** The column 'Nacionality' was identified with a spelling error and was renamed to 'Nationality.'
- **Incorrect Data Types:** Several columns were initially read as object types but were meant to be numerical. These were converted to their appropriate data types (int or float) to allow for further analysis and transformations.
- **Outliers:** The outliers in the grade columns, such as "Previous qualification (grade)," were addressed using the standard deviation capping method. This method capped extreme values beyond three standard deviations from the mean to prevent skewing the analysis.
- **No Missing Data:** Upon checking for missing values, the dataset had no null values, indicating that it was complete and no further imputation was necessary.

JUSTIFICATION FOR CHOSEN DATA TRANSFORMATION METHOS

- **Normalization of Numerical Features:** Numerical features such as “Application order,” “Admission grade,” and “Age at enrolment” were normalized using Min-Max scaling to bring all features into a range of [0, 1]. This transformation was chosen to ensure that all numerical features were on a consistent scale, preventing any one feature from dominating the model during the analysis phase.
- **Categorical Encoding:** Categorical variables, including "Marital status," "Course," and "Mother's qualification," were label-encoded using LabelEncoder. This method was selected because it assigns a unique integer value to each category, making the categorical data ready for use in machine learning models that require numerical input.
- **Derived Features:** Several new features were created to capture important relationships in the data. For instance:
    - **Total Curricular Units per Semester:** The sum of credited and enrolled curricular units was calculated to assess each student’s total workload.
    - **Weighted Grades and GPA:** Weighted grade averages were computed to derive each student’s GPA, an important metric for academic performance. This was done by calculating the weighted sum of grades based on credits and dividing by the total credits.

These transformations aimed to enhance the dataset's representational power and ensure that key aspects such as academic performance and course loads were properly captured for further analysis.

CONCLUSION

The preprocessing phase of this academic performance dataset ensured that the data was clean, consistent, and ready for robust analysis. Through data cleaning methods such as correcting headers, renaming columns, converting data types, and capping outliers, the dataset became a reliable foundation for subsequent analysis. Additionally, the normalization, encoding, and feature engineering steps applied helped in making the data more interpretable and suitable for machine learning algorithms.

By addressing data quality issues and applying these necessary transformations, I have significantly enhanced the dataset and ensured that future analyses or models will be based on a sound and accurate dataset. This preprocessing process lays the groundwork for meaningful exploration of the factors that contribute to student performance and academic success.