Data preprocessing report

Data Validation

The Datasets “data.csv” chosen for this assignment are regarding “dropout rates” downloaded from 3signet. The Datasets contains information regarding school students and aims to identify students at risk of dropping out. \

This data set has 4424 rows and 37 columns. All variables except Target variable contained numerical datatypes but were changed to their correct data types. I have validated all variables and I have made changes after validation. No missing values were found but there were inconsistent categories in most categorical columns but were encoded with Label Encoder and saved with encoded attached at the end of the variable names. New variables were created to store the information of total Curriculum units for each status(i.e. enrolled, credited, evaluation etc). The data types were converted to their correct data types. There were outliers in the numerical columns but were not treated since it has to do with student's performance. The numerical features were normalized with StandardScaler to make them be on the same scaled and make it normally distributed.\

Data cleaning steps

Step 1

Imported the necessary libraries for the project \

Step 2

Loaded the data.csv dataset to have a sense of what the datasets look like using the 'data.shape' attribute to see how many rows and columns it contains, summary statistics with the '.describe()', and to view the data types with the '.info() and also checked for missing and duplicated methods and viewed each column to check the values each contained in each columns.

Step 3

I converted the the necessary columns to the categorical columns by iterating over the necessary columns, feature engineered the curriculum units columns, adding to the dataframe and dropping the columns used for feature engineering

Step 4

Normalized numerical columns by checking the variance of each columns,and looking at the distributions and merging to the dataframe.

Step 5

I saved the cleaned and transformed dataset to 'cleaned_data.csv'

Issues Encountered

There were no description provided to enable us understand what the variables represents and how the data types should be but we were able to convert them to their correct data types. Also the categorical columns contains inconsistent values but we were able to convert them to catergoical values.

The following variables were converted to categorical columns because the contained unique values:
'Marital status', 'Displaced', 'Application mode', 'Application order','Previous qualification', 'Nacionality', "Mother's qualification", "Father's qualification", "Mother's occupation", "Father's occupation", 'International','Scholarship holder', 'Gender', 'Tuition fees up to date', 'Debtor','Educational special needs', "Daytime/evening attendance\t", 'Target', 'Course'.