13-09-2024

Deliverables of Week One Project

Data Science Internship: **Group 4**

Precious Benjamin

Description of Data Cleaning Steps

- Stripping columns of whitespace using .strip()
- Renamed Necessary columns using .rename
- Checked for missing values
- Got more information on my dataset using .info()
- Changed features to their appropriate datatypes
- Extracted numerical features from the dataframe
- Extracted categorical features
- Created a subplot of boxplots to clearly display outliers
- Identified Outliers using Interquartile Range (IQR)
- Plotted a subplot of histograms to understand the skewness of the data
- Took care of outliers using median
- Capp the remaining outliers to reduce their impact using winsorizing

- Encoded the Target feature
- Concatenated the numerical and categorical columns
- Normalized the numerical features using MinMaxScaler
- Data validation
- Visualized the value count for the targets

Summary of Data Quality

- The quality of the data was not bad. The dataset was a csv file, however, it was semi colon separated instead of comma separated. Most of the features were already encoded and given identification tags. The dataset had no missing values, or duplicated values. They were negative values in the GDP, and inflation columns, however, after thorough research, I noticed they weren't error but represents low economic conditions.

Issues Encountered | How They Were resolved

- The csv file was column separated, instead of comma separated. I had to specify the delimiter before I could load the dataset properly.

- Most of the columns were numerical, I had to I had to analyze the data to understand which columns should be categorical and converted them appropriately.
- Outliers was the major issue I encountered. I handled the outliers using various methods, but they were not totally eliminated. I couldn't determine if they are indeed outliers or errors, I had to normalize them.

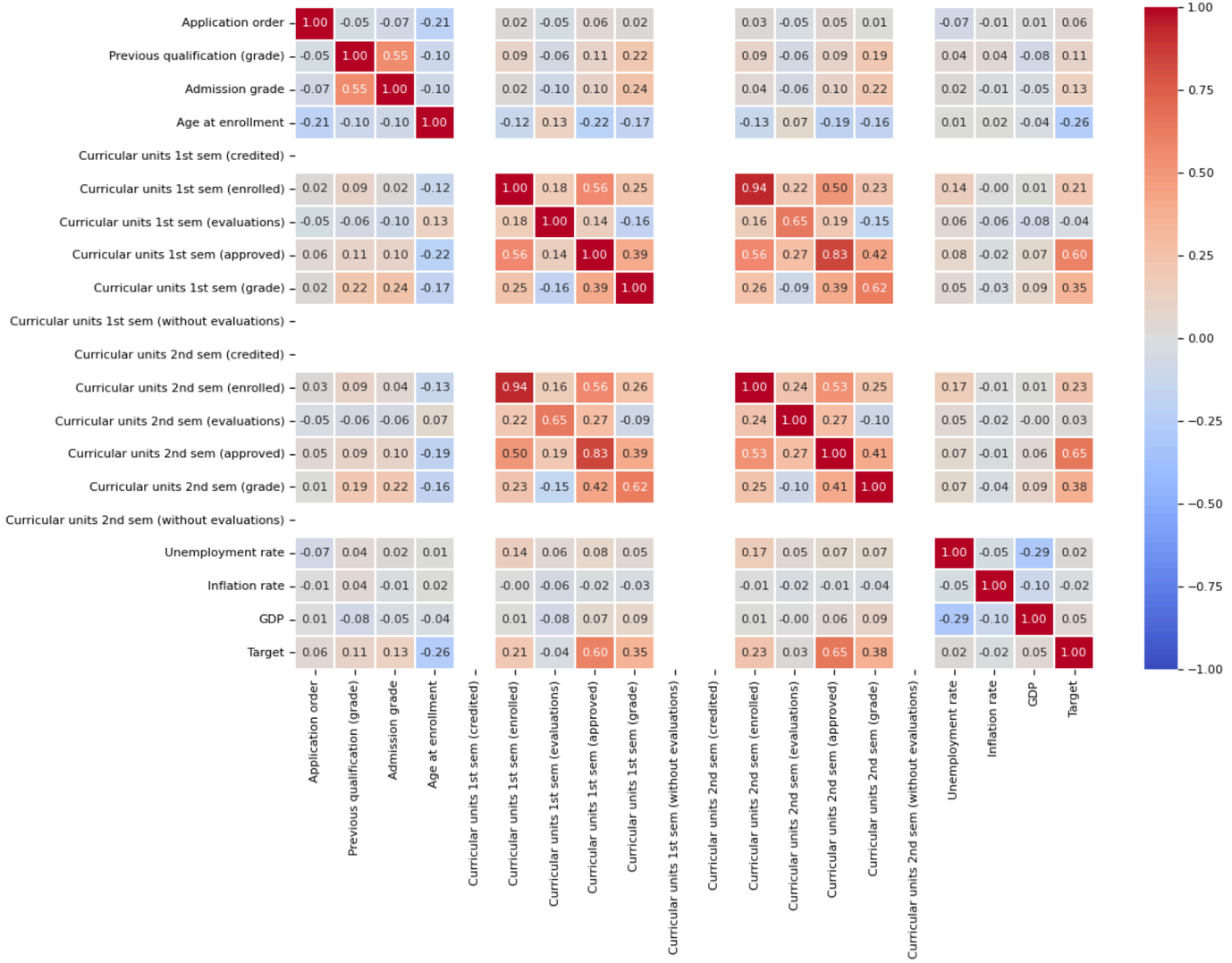
Justification for chosen data transformation methods

- **Using IQR to Handle Outliers:** IQR is a robust measure of dispersion, meaning it's less sensitive to outliers compared to the standard deviation. This makes it a suitable choice for identifying outliers in datasets with extreme values. IQR can be used in conjunction with various outlier handling techniques, such as winsorization or trimming.

- **Winsorizing Outliers:** Winsorization replaces outliers with the nearest non-outlier value, preserving the overall shape of the distribution while mitigating the impact of extreme values.

- **Normalizing Numerical Columns After IQR and Winsorization:** Many machine learning algorithms, such as linear regression, logistic regression, and support vector machines, perform better when features are on a similar scale. Normalization helps ensure that features contribute equally to the model's predictions. Normalized data can make it easier to interpret the results of your analysis. For example, if two features have similar magnitudes after normalization, it's easier to compare their importance in the model.

Correlation Matrix Heatmap



Results and Interpretation of Hypothesis Test

According to the hypotheses, the features, Nationality, International, and Education special needs are identified as null hypotheses, which means they are assumed to have no significant effect on the target variables (Graduate, Dropout, and Enrolled). This implies that these features do not contribute meaningful information in predicting the outcomes of the target variables.

Given this, it is appropriate to drop these features from the dataframe when building the predictive model. By excluding these variables, we streamline the model, reducing complexity and potentially improving its performance by focusing only on features that have a demonstrated impact on the target variables. This approach helps to prevent overfitting and ensures that the model is based on relevant, impactful features.

In summary, removing Nationality, International, and Education special needs from the dataset aligns with the hypothesis that these features do not significantly influence the target outcomes, thus enhancing the model's efficiency and accuracy.

Descriptive Analysis

```
l]:
# Descriptive statistics for all variables
descriptive_stats = num_cols.describe(include='all')

# Display descriptive statistics in a tabular form
descriptive_stats.head()
```

ıt[31]:

	Application order	Previous qualification (grade)	Admission grade	Age at enrollment	Curricular units 1st sem (credited)	Curricular units 1st sem (enrolled)	Curricular units 1st sem (evaluations)	Curricular units 1st sem (approved)	Curricular units 1st sem (grade)	Curricular units 1st sem (without evaluations)	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)
4424.000000	4424.000000	4424.000000	4424.000000	4424.0	4424.000000	4424.000000	4424.000000	4424.000000	4424.0	4424.0	4424.000000
1.263336	132.597536	126.171022	21.062613	0.0	6.116184	7.850814	4.354430	12.621496	0.0	0.0	6.200949
0.577718	10.970809	13.132899	3.824539	0.0	0.949348	3.346976	2.430971	1.212549	0.0	0.0	1.070260
1.000000	110.000000	99.000000	18.000000	0.0	5.000000	0.000000	0.000000	10.000000	0.0	0.0	5.000000
1.000000	125.000000	117.900000	19.000000	0.0	6.000000	6.000000	3.000000	12.000000	0.0	0.0	6.000000