Feature Engineering Report

Feature engineering is the process of using domain knowledge to select, modify, or create new features (variables) from raw data to improve the performance of machine learning models. It's a crucial step in the data preprocessing phase, as the right features can significantly enhance model accuracy and interpretability. Below shows the analysis of how the features were engineered:

1. **Features created**

Total Curricular units (grade): This shows the total grades gained for the session (first and second semester.

Total Curricular units (credited): This shows the total credits gained for the session (first and second semester.

Total Curricular units (enrolled): This shows the total units enrolled for the session (first and second semester.

Total Curricular units (evaluations): This shows the total units evaluated for the session (first and second semester.

Total Curricular units (approved): This shows the total units approved for the session (first and second semester.

Total Curricular units (without evaluations): This shows the total units not evaluated for the session (first and second semester.

2. **Justification for Feature Transformation**

All numerical columns were transformed using log transformation in order to transform them to a normal distribution because machine learning models assume that the data is normally distributed, binned and scaled to avoid overfitting and to avoid one variable dominating the other.

3. **Analysis of feature importance and selection results**

The feature importance were analyzed with chisquare, Recursive Feature Elimination, Ridge to find the 10 highest performed features. Below are the analysis of the selection:

a. **Chisquare:** Used chisquare to check the scores of the top 10 features

```
[8.081e+01 9.510e+02 3.933e+01 3.360e+00 3.138e+00 3.971e+02       nan
 1.223e+01 1.610e+02 5.725e+01 6.774e+01 1.379e+02       nan 2.608e+01
 6.346e-01 2.298e+02 9.829e+01 1.512e+02 3.081e+02 6.284e+01 1.248e+00
 2.285e+00 9.939e-01 5.325e+00 1.535e+02 4.745e+00 2.098e+00 1.617e+01
 2.705e+02 2.325e+01]
[[ 7.     0.    12.     9.     0.     1.     1.     0.     0.     0.   ]
 [ 5.     0.     0.     3.     0.     0.     1.     0.     0.925  0.691]
 [ 0.     0.    21.     9.     0.     0.     1.     0.     0.     0.   ]
 [ 7.     0.    22.     3.     0.     1.     0.     0.     0.907  0.646]
 [11.     0.    21.     9.     0.     1.     0.     0.     0.902  0.691]]
```

b. **Recursive Feature Elimination:** Shows the top 10 features

	Feature	Important
16	Tuition fees up to date	True
21	Unemployment rate	True
28	Total Curricular units (approved)	True
3	Course	True
27	Total Curricular units (evaluations)	True
26	Total Curricular units (enrolled)	True
24	Total Curricular units (grade)	True
19	Age at enrollment	True
10	Mother's occupation	True
11	Father's occupation	True
18	Scholarship holder	False
20	International	False

c. **Ridge:** Shows the coefficients of each features

```
Ridge model: 0.016 * X0 + -0.009 * X1 + -0.001 * X2 + -0.018 * X3 + -0.065 * X4 + 0.006 * X5 + 0.0 * X6 + -0.022 * X7 + -0.002 * X8 + 0.001 * X9 + 0.008 * X10
+ -0.001 * X11 + 0.0 * X12 + -0.027 * X13 + -0.11 * X14 + -0.178 * X15 + 0.427 * X16 + -0.09 * X17 + 0.216 * X18 + -0.243 * X19 + 0.35 * X20 + -0.015 * X21 + -0.0 * X22
 + 0.017 * X23 + -0.508 * X24 + -0.385 * X25 + -0.704 * X26 + -0.217 * X27 + 2.619 * X28 + -0.065 * X29
```

d. **Principal Component Analysis:** Shows the ratio of each 10 features

```
array([0.514, 0.148, 0.107, 0.102, 0.058, 0.034, 0.015, 0.01 , 0.005,
       0.001])
```

Fig 1: Visualization of high dimensionality Results

Shows clusters of groups that can be created from the features