Feature Engineering Report

1. Detailed Explanation of Each New Feature Created

a. Progress Rate:

Description: This feature calculates how much progress a student has made in their curriculum by measuring the proportion of approved curricular units relative to the total enrolled curricular units across both semesters. It gives insight into how much of their coursework the student has successfully completed.

Formula: (Curricular units 1st sem (approved) + Curricular units 2nd sem (approved)) / (Curricular units 1st sem (enrolled) + Curricular units 2nd sem (enrolled))

Justification: The progress rate allows us to quantify how much of the curriculum a student has successfully completed. A high progress rate indicates that the student is likely performing well, while a low progress rate might signal potential academic struggles.

b. Improvement Rate:

Description: The improvement rate measures the student's performance improvement between the 1st and 2nd semesters by evaluating the change in approved curricular units. This feature is useful for understanding how a student's academic performance has evolved over time.

Formula: (Curricular units 2nd sem (approved) - Curricular units 1st sem (approved)) / Curricular units 1st sem (approved)

Justification: The improvement rate is a good measure of a student's ability to adapt and improve over time. A positive improvement rate shows that the student is improving their academic performance, while a negative rate could indicate declining performance.

c. Completion Time:

Description: This feature reflects how much of a student's overall curriculum has been completed relative to the total units they were enrolled in across both semesters. It helps to assess whether a student is keeping up with their expected pace of completion.

Formula: (Curricular units 1st sem (approved) + Curricular units 2nd sem (approved)) / (Curricular units 1st sem (enrolled) + Curricular units 2nd sem (enrolled))

Justification: This feature is important to gauge how quickly a student is progressing through their courses. If the completion time is high, it suggests that the student is efficiently completing their enrolled units; if low, it could indicate potential delays or difficulties in finishing courses on time.

d. Study Time Per Credit:

Description: This feature estimates the average amount of study time allocated per enrolled credit by dividing the total number of enrolled curricular units by the total number of evaluations completed in both semesters. It helps to assess how much time a student might be spending on their courses.

Formula: (Curricular units 1st sem (enrolled) + Curricular units 2nd sem (enrolled)) / (Curricular units 1st sem (evaluations) + Curricular units 2nd sem (evaluations))

Justification: Understanding how much study time is spent per credit can offer insights into the student's workload and time management. A higher value might indicate a heavier workload, while a lower value could mean that the student is efficiently managing their study time.

e. Performance Ratio:

Description: This feature calculates the overall performance by measuring the ratio of approved units to enrolled units across both semesters. It provides a clear metric for evaluating the student's academic success.

Formula: (Curricular units 1st sem (approved) + Curricular units 2nd sem (approved)) / (Curricular units 1st sem (enrolled) + Curricular units 2nd sem (enrolled))

Justification: The performance ratio gives a direct measure of a student's academic success by evaluating how many of the enrolled units were approved. It is a critical feature for predicting academic performance and identifying students at risk of underperforming.

f. Polynomial Features (for Interaction Effects):

Description: Polynomial features are created to capture non-linear relationships between variables. By interacting different features such as academic performance metrics and financial status, we can better understand how these factors influence student outcomes.

Implementation: Polynomial transformations were applied to some of the base features, raising them to powers or creating interactions between features such as Curricular units and financial status.

Justification: Polynomial features allow the model to capture more complex, non-linear relationships between features, which may not be evident in the base features alone. This helps improve model accuracy by including interaction effects.

2. **Justification for Feature Transformations:** The newly created features offer deeper insight into the performance metrics across both semesters for each student. These features encapsulate key aspects like academic progress, improvement over time, and consistency. The transformations enable the model to capture longitudinal academic trends, providing a richer representation of a student's academic journey rather than focusing on isolated static snapshots.

3. **Analysis of Feature Importance and Selection Results**: The importance of these features was evaluated using Random Forest and Lasso regression methods. Both methods highlighted key predictors like 'progress_rate', 'performance_ratio', 'Curricular units 2nd sem (approved)', 'Curricular units 2nd sem (grade)', and 'Curricular units 1st sem (approved)' as significant contributors. Recursive Feature Elimination (RFE) helped in narrowing down the most significant features, ensuring that the dataset is more manageable while preserving predictive performance.

4. **Visualization of Dimensionality Reduction Results:** The dimensionality reduction was visualized using t-SNE and PCA techniques. These methods provided clear visual insights into the distribution of student performance across a lower-dimensional space. Both techniques helped in confirming the separability and clustering of student performance outcomes, supporting the validity of the newly created features.