Machine Learning for Behavioral Data: Milestone 4

Manon Landrieux, Ilias Merigh, Malen Raychev

Date: 30/04/2025

Abstract

In Switzerland today, technology has diversified learning methods, leading to the development of several educational platforms in recent years. One such platform is GoGymi, designed to help students prepare for the Gymi entrance exam. It offers exercises with solutions, tracks student activity, offers an AI-based tutor and essay correction, and delivers personalized support based on individual learning analytics. This platform provides rich data on users' learning behaviors, raising several questions: Can distinct student profiles be identified based on their learning behaviors? How do these profiles relate to academic performance? Do they evolve over time, and how is their development influenced by the use of the AI tutor?

I Research Question

In Milestone 4, we focus on the following research question based on behavioral clustering: Can we identify different student profiles based on their behaviors and understand how these profiles relate to academic performance?

II Data Preprocessing

The input features for the model were computed from the `activity` dataset after preprocessing.

- `user_id`: Unique identifier of the student.
- `active_days_count`: Number of days the student has been active since registration.
- `activities_count`: Total number of activities completed by the student since registration.
- `activity_types_count`: Number of distinct activity types the student engaged in.
- `median_gap_days`: Median number of days between consecutive login sessions.
- `median_activity_duration_minutes`: Median duration, in minutes, of the student's activities.

The output variable, `median_score`, was computed from the `all_scores` dataset, after the preprocessing, by aggregating each student's scores using the median.

Our preprocessing involved removing missing and inconsistent values, as well as filtering out outliers using the IQR method across multiple fields.

Finally, both preprocessed datasets were merged, combining the input features and output scores into a single dataset, where each entry corresponds to a user, resulting in a final dataset of 472 users.

III Model Building

To identify potential distinct learner profiles, an unsupervised learning approach with K-Means clustering was used. The features selected for clustering and the number of clusters were treated as key hyperparameters.

1. Hyperparameter Tuning of the Clustering Features

The hyperparameter tuning of the clustering features was conducted using the following two methods:

- **Spearman Correlation Analysis**: This was used to evaluate the relationship between the input features (student activity) and the target variable (student score). Spearman correlation was preferred over Pearson because, although the scores follow a Gaussian distribution, the input features are highly right-skewed, likely following an exponential distribution.
- **Linear Regression Coefficient Interpretation**: This was used to interpret the coefficients and identify which input features (student activity) are most strongly associated with the target variable (student score). Since the objective was interpretation rather than prediction, the model was trained on the full dataset without splitting it into training and evaluation sets. However, to ensure robustness, 5-fold cross-validation was performed, and the coefficient weights were averaged across the folds.

Due to the presence of relationships that appeared counterintuitive based on common sense, all features were retained, and their impact on clustering performance has been assessed in a subsequent analysis.

2. Hyperparameter Tuning of the Number of Clusters

The hyperparameter tuning of the number of clusters was performed using the Elbow method. Euclidean distance was used as the distance metric, as it is well-suited for our data where all features are numerical. The optimal number of clusters was found to be 4.

IV Model Evaluation

Our model is evaluated based on two key aspects: the reliability of feature weights and the performance of the clustering.

1. Reliability of Feature Weights

Reliability of feature weights is evaluated by Spearman correlation and the linear regression model. Spearman analysis reveals that `active_days_count` , `activities_count`, and `median_activity_duration_minutes` have weak but statistically significant monotonic relationships with the target.

Although the linear regression model yields a negative R² value (-0.007), indicating poor predictive power, the model's coefficients still reflect a consistent pattern with the Spearman correlations, further supporting the findings from the correlation analysis.

2. Performance of the Clustering

As depicted in Figure 1, the performance of the clustering produced by K-Means is evaluated using the silhouette scores.

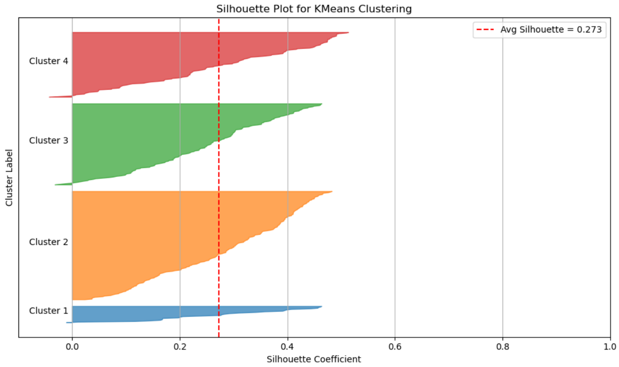


Figure 1: Silhouette Plot for KMeans clustering

The average silhouette score of 0.273 indicates modest clustering quality but suggests some underlying structure. The modest clustering performance may be attributed to the limited richness of the data or to unclear cluster boundaries. It also reflects the limitations of KMeans, which assumes spherical and equally sized clusters, and of the silhouette score, which may not fully capture more complex or overlapping structures in the data.

V Clusters Interpretation

Once the clusters were identified, additional processing was carried out to support their interpretation.

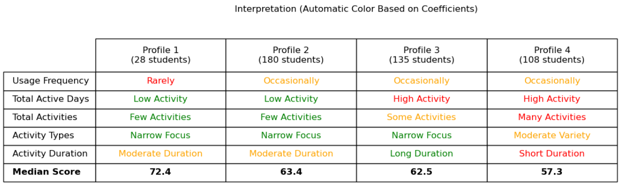
For each cluster, the median value of each feature was calculated across all students. Using the median as an aggregation method reduces the influence of outliers and provides a more reliable basis for comparison.

To enhance clarity, a color-coding system was applied to the features. Students were divided into three equally sized groups (low, medium, high) based on their relative feature values, with text colors indicating the interpretation (e.g., green for favorable values and red for unfavorable ones).

Two color-coding strategies were implemented: a "common-sense" approach, based on manual, intuitive interpretations, and a "coefficient-based" approach, driven by feature importance scores.

Comparing both methods revealed that the coefficient-based interpretation more accurately reflected the actual performance outcomes.

Figure 2, inspired by the approach of Mejía-Doméznain et al. [1], presents the different profiles using coefficient-based coloring and ranked in descending order according to their median performance scores.



Interpretation (Automatic Color Based on Coefficients)

	Profile 1 (28 students)	Profile 2 (180 students)	Profile 3 (135 students)	Profile 4 (108 students)
Usage Frequency	Rarely	Occasionally	Occasionally	Occasionally
Total Active Days	Low Activity	Low Activity	High Activity	High Activity
Total Activities	Few Activities	Few Activities	Some Activities	Many Activities
Activity Types	Narrow Focus	Narrow Focus	Narrow Focus	Moderate Variety
Activity Duration	Moderate Duration	Moderate Duration	Long Duration	Short Duration
Median Score	**72.4**	**63.4**	**62.5**	**57.3**

Figure 2: Different Student Profiles and their Median Score

Based on this analysis, four distinct behavioral profiles were identified:

- Profile 1 suggests that focused, efficient study sessions might be particularly effective.
- Profile 2 has a lower score of 63.4, indicating a possible difference in study effectiveness compared to Profile 1.
- Profile 3 might suggest that higher activity does not necessarily translate to higher performance, especially if it lacks focus.
- Profile 4 implies that short, scattered interactions may be less effective for learning outcomes.

Interestingly, from Figure 2, Profile 1 and 2 share similar characteristics and could potentially be merged based on our profile analysis.

In conclusion, the analysis successfully revealed distinct user profiles and allowed us to compare their performance in terms of median scores. This, in turn, provided valuable insights into the factors that may influence academic success, highlighting meaningful patterns and differences across student profiles.

VI Team Members' Contributions

For Milestone 4, we all contributed equally. At the beginning, we individually explored the data to better understand its structure and potential directions. We then came together to collaborate on the final version of the code and report.

References

[1] P. Mejia-Domenzain, E. Laini, S. P. Neshaei, T. Wambsganss, T. Käser, Visualizing Self-Regulated Learner Profiles in Dashboards: Design Insights from Teachers, **2023**.