Machine Learning for Education Laboratory
ML4ED

Do As I Do

Optimal Behavior Identification Through Success Prediction on Time Series Data

Kai Cooper, María Isabel Ruiz Martínez, Nicolas Thierry d'Argenlieu

Machine Learning for Behavioral Data (MLBD) Course, EPFL, 2022

1 INTRODUCTION + RESEARCH QUESTION

Improved web technologies have made more accessible online learning format such as MOOCs and dedicated learning platforms such as Moodle. **Lernnavi** is one of these **online education platforms**, focusing on preparing Swiss students for their Matura exam. They offer Math and German lessons and exercises. Students track their progress through a mastery score updated by level checks, and proper to each topic taught. Using the extracted **data from this platform**, we aim to answer the following research question:

Is it possible to identify one or more studying behaviors leading to a significant improvement in level check results?

2 METHODOLOGY

To tackle our question, we followed **2 approaches**. Each time, we try to **cluster** our dataset to see if behavioral profiles can be detected in the raw data and improve the prediction. This is followed by training a model to **predict** the students' level check results. **Approach 1 [A1]**: we cluster the dataset based on *regularity metrics*[1], then we train a classifier on the *time series* from the students' weekly activity. **Approach 2 [A2]**: we cluster the *time series* from the students' weekly activity, then we train a classifier on the students' *engagement*[2] *and regularity*[1] *scores*.

Dataset

Raw data is a set of dataframes containing information on student profiles (demographics), activity tracking and platform content. We selected data concerning navigation and participation activities. Students we cannot **track over time** are removed from the data, criteria for this included too few actions of any given type, for example. From their level-check evaluation data, we create the classification **labels** (*0* or *1* if the student's mastery score improved or decreased following the previous level check, respectively.)

Model Architecture

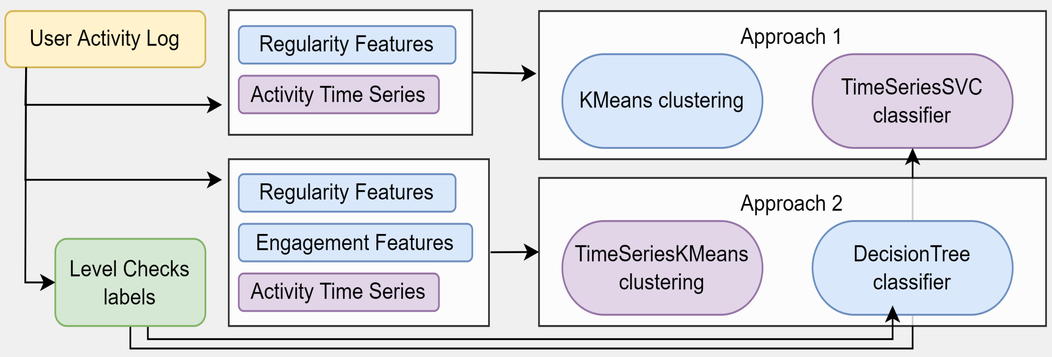

Figure 1. Overview of our experiments' architecture. [A1] Clustering is done using the **KMeans algorithm**. The optimal *k* is chosen according to the **silhouette score** of the clustered samples. The time series classification is done using a **TimeSeriesSVC** [3] model (SVM-based time series binary classifier). [A2] Clustering is done using a **TimeSeriesKMeans** [3] model, with the optimal *k* chosen using the **silhouette score** technique. The binary classification is done on the computed regularity [1] and engagement [2] features using a **decision tree classifier**.

Choice	Justification
Time series classification/clustering	Student success is not a memoryless process
Regularity/engagement based features/clustering	Identification of 'best student' patterns could inform 'best usage' of platform
Tree-based classification	Highly interpretable approach to categorise and rank importance of student trend metrics.

3 RESULTS

[A1] Clustering on regularity metrics yielded ***k=2***. Classification was subsequently done with 2 and no clusters. Training was performed over the number of training samples due to few hyperparameters. The best performing model achieves **79.9% accuracy**, **86.6% F1-score** for 828 samples and no clustering.

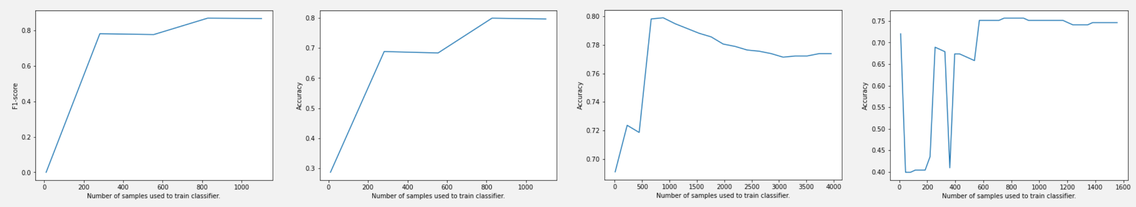

Figure 2. F1-score and accuracy with TimeSeriesSVC and no clustering

Figure 3. Accuracy with TimeSeriesSVC for clusters 0 & 1 (2 clusters).

Accuracy	Balanced Accuracy	Adjusted Balanced Accuracy	F1 Score	Weighted F1 Score
0.799	0.702	0.403	0.868	0.784

Table 1. Metrics details for best performing TimeSeriesSVC model (no clustering)

[A2] Classification was done with no clusters, since our clustering methods did not cluster significantly the dataset. Hyperparameters were tuned for each of the models. Two different experiments were carried out, the first one with more features than the second. The best performing model achieved **97.8% accuracy**, **98.5% F1-score** for 962 samples.

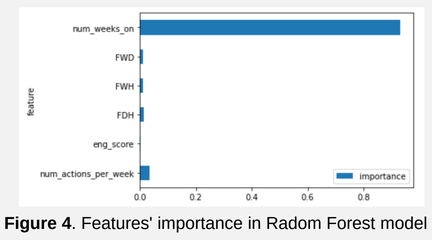

Figure 4. Features' importance in Radom Forest model 1.

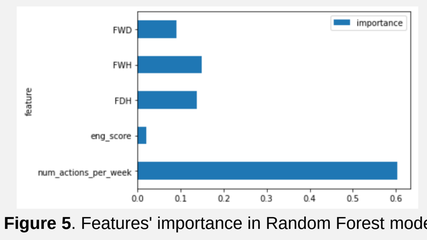

Figure 5. Features' importance in Random Forest model 2.

	Accuracy	Balanced Accuracy	Adjusted Balanced Accuracy	F1 Score	Weighted F1 Score
Experiment 1	0.978	0.983	0.966	0.985	0.978
Experiment 2	0.562	0.552	0.103	0.659	0.588

Table 2. Metrics details for Random Forest Classification models (no clustering)

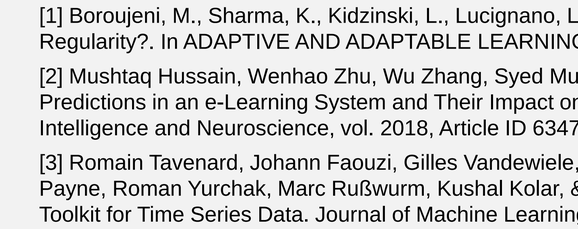

Figure 6. F1-score and accuracy with the Radom Forest Classifier obtained with the features shown in **Figure 4** and no clustering.

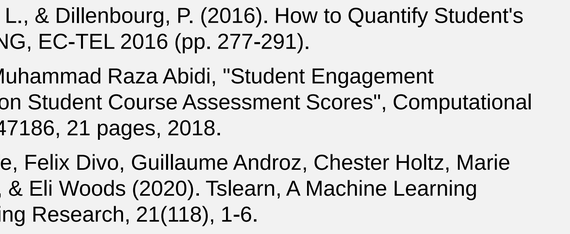

Figure 7. F1-score and accuracy with the Radom Forest Classifier obtained with the features shown in **Figure 5** and no clustering.

4 CONCLUSION

Our results demonstrate that **differing patterns** of **student behaviours** certainly **influence their performance** on the platform. Even as we could not identify a specific optimal behavior, both experiments clearly demonstrated its potential by giving good prediction accuracy. We believe that in future work, this platform could be used to carry out directed **causal** experiments where students demonstrate explicit patterns and use the technology to study what brings about student success.

REFERENCES

[1] Boroujeni, M., Sharma, K., Kidzinski, L., Lucignano, L., & Dillenbourg, P. (2016). How to Quantify Student's Regularity?. In ADAPTIVE AND ADAPTABLE LEARNING, EC-TEL 2016 (pp. 277-291).

[2] Mushtaq Hussain, Wenhao Zhu, Wu Zhang, Syed Muhammad Raza Abidi, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores", Computational Intelligence and Neuroscience, vol. 2018, Article ID 6347186, 21 pages, 2018.

[3] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, & Eli Woods (2020). Tslearn, A Machine Learning Toolkit for Time Series Data. Journal of Machine Learning Research, 21(118), 1-6.

École polytechnique fédérale de Lausanne