# SAVis: a Learning Analytics Dashboard with Interactive Visualization and Machine Learning

Zeynab Mohseni[1], Rafael M. Martins[1] and Italo Masiello[1]

[1]*Department of Computer Science and Media Technology, Linnaeus University, Sweden*

**Abstract**

A dashboard that provides a central location to monitor and analyze data is an efficient way to track multiple data sources. In the educational community, for example, using dashboards can be a straightforward introduction into the concepts of visual learning analytics. In this paper, the design and implementation of Student Activity Visualization (SAVis), a new Learning Analytics Dashboard (LAD) using interactive visualization and Machine Learning (ML) is presented and discussed. The design of the dashboard was directed towards answering a set of 22 pedagogical questions that teachers might want to investigate in an educational dataset. We evaluate SAVis with an educational dataset containing more than two million samples, including the learning behaviors of 6,423 students who used a web-based learning platform for one year. We show how SAVis can deliver relevant information to teachers and support them to interact with and analyze the students' data to gain a better overview of students' activities in terms of, for example, their performance in number of correct/incorrect answers per each topic.

**Keywords**

Learning Analytics Dashboard, Visual Learning Analytics, Educational Dataset, Machine Learning, Visualization, SAVis

## 1. Introduction

The increased use of technology in education has enabled educational institutions to collect a large variety of data about their students. Educational data, e.g., text answers, tests, numbers, timestamps, users' info and usage of the digital learning material or platform, are frequently large in amount, complex, and heterogeneous, and therefore difficult to be meaningfully interpreted by teachers [1, 2]. To aid in making sense of educational data, a relatively new field, commonly referred to as Learning Analytics (LA), has matured [3, 4]. Siemens et al. [5] define LA as: "The measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs". To further aid teachers in their interpretation of LA, data can be visualized in dashboards. LADs, defined by Schwendimann et al. [6] as "a single display that aggregates different indicators about learner(s), learning process(es) and/or learning context(s) into one or multiple visualizations", are developed with the intention to increase motivation, self-direction, learning effectiveness,

*Nordic Learning Analytics (Summer) Institute 2021, KTH Royal Institute of Technology, Stockholm, 23 August 2021*
zeynab.mohseni@lnu.se (Z. Mohseni); rafael.martins@lnu.se (R. M. Martins); italo.masiello@lnu.se (I. Masiello)
https://lnu.se/personal/zeynab.mohseni/ (Z. Mohseni); https://lnu.se/personal/rafael.martins/ (R. M. Martins); https://lnu.se/personal/italo.masiello/ (I. Masiello)
0000-0002-3297-0189 (Z. Mohseni); 0000-0002-2901-935X (R. M. Martins); 0000-0002-3738-7945 (I. Masiello)

performance of students, and teachers' engagement [7]. In this paper, we explore the design of a LAD with the use of interactive visualization and ML [8] with the intention to answer real pedagogical questions that teachers may have in their practice. Our SAVis is meant to be connected and used with other existing learning management systems or digital learning material to give teachers a broader overview of their students. SAVis provides an interactive platform that enables the comparison and exploration of students' activities by looking at, and interacting with, various visualizations presented in the dashboard.

The rest of the paper is organized as follows. Section 2 describes briefly the related work in this field. In section 3 we have presented the design of SAVis. In section 4 of this paper which is about a real-world problem, we have provided specific pedagogical elaborations on how an educator can quickly gain information about how the students perform and approach the learning activities. Section 5 discusses the conclusion and presents possible lines of future work.

## 2. Related Work

During the last decade, some LADs and visualization tools have been developed to support decision-making by considering various scenarios and datasets. For instance, Govaerts et al. [9] developed the Student Activity Meter (SAM) to visualize students' time spent and resource use to support awareness for teachers and students. In [10] Ez-zaouia et al. presented the EMODA dashboard which allows the teacher to monitor learners' emotions and better understand their evolution during an online learning session. Also, He et al. [11] developed LearnerVis to visualize the temporal features of the learning process and help users analyze how students schedule their multi-course learning. In this paper, we showcase and motivate the initial design and development of a LAD where interactive visualization is used in order to interpret the results of general ML algorithms, providing an interactive platform for teachers to explore and analyze students' data from many different perspectives at the same time.

## 3. Proposed LAD

We designed SAVis with the intent to answer 22 pedagogical questions (PQ)(real-world problem) that an educator might have when exploring students' learning activities:

*PQ1. How many correct and incorrect answers are there in general? PQ2. What are the maximum and minimum time students spend to answer a question? PQ3. What is the accuracy of the ML algorithm to classify the students to the right university? PQ4. How many correct and incorrect answers are there for each student? PQ5. How many correct and incorrect answers are there for every topic? PQ6. How many correct and incorrect answers are there per month? PQ7. How many correct and incorrect answers are there per day? PQ8. How many correct and incorrect answers are there per hour? PQ9. What are the percentages of correct and incorrect answers in general? PQ10. How many correct and incorrect answers are there for different time categories? PQ11. How many correct and incorrect answers are there for every student category? PQ12. How many correct and incorrect answers are there for every topic category? PQ13. How many correct and incorrect answers are there for every question type? PQ14. How many correct and incorrect answers are*

*there for every month of year/day of month/hour of day? PQ15. What hour of the day are students more active? PQ16. What day of month are students more active? PQ17. What month of year are students more active? PQ18. Which topic has the highest number of correct answers? PQ19. What are the top 10 topics in which students are more active? PQ20. What is the number of correct answers for top 10 topics? PQ21. Which question type is the most common? PQ22. What is the percentage of students' activities in different months of the year?*

Each sample of the dataset contained various features such as *student ID*, *topic ID*, *question type*, *resource name*, *resource type*, *student answer*, *answer duration*, and the *month*, the *day* and the *hour of student's activity*. Moreover, for the student's monthly, daily, and hourly activity, we referred to the number of correct/incorrect answers per month, day, and hour, respectively. For the analysis performed in this paper, and in order to improve performance and avoid overload of users in the dashboard, 10,000 random samples were selected from the dataset.

## 3.1. Overall Design of SAVis

In this subsection, we provide an overview of the dashboard by describing its different sections. As can be seen in figure 1, we used levels of increasing detail from "Key metric" to "Context" to "Detail". These three levels follow Shneiderman's mantra [12], "overview first, zoom and filter, then details on demand", to guide visual information-seeking behavior and the design of the interface. The ***Key metrics*** section shows the most important information on a general level about the dataset, including the total number of random complete samples and their related number of students, educational topics, correct and incorrect answers, and the minimum and maximum time in minutes for the answer duration. The ***Context*** section of the dashboard consists of a scatter plot which presents the students' activity using t-Distributed Stochastic Neighbor Embedding (t-SNE) [13], a slider menu to change the t-SNE properties, and a heatmap (explained below) that shows the performance of a Random Forest classifier according to its attempt to predict the university to which the student belongs.

### 3.1.1. t-SNE View

t-SNE is an unsupervised non-linear dimensionality reduction technique used for the visualization of high-dimensional datasets. It enables us to create compelling two-dimensional "maps" from data with hundreds of dimensions in such a way that similar objects are modeled by nearby points, and dissimilar objects are modeled by distant points [13]. In other words, t-SNE is a data mining algorithm that shows patterns in the multidimensional space of the data. The example shown in figure 1 includes 2,560 students and 835 educational topics. Every color in this view represents a different university ID. In this paper, we are simply using the university IDs as a layer of exploration dictated by the dataset, to check if the students that are coming from the same place happen to have similar learning patterns. In other scenarios, a teacher would probably use a class ID or similar. As can be seen in figure 1, there are nine university categories: the students belong to eight different medical universities, therefore eight registered university IDs (categories from "1" to "8"), and one for the unregistered university IDs (category "9"). Using the t-SNE plot allows us to find students with similar activity patterns in terms of the number of correct/incorrect answers for educational topics in different universities. By

selecting a small group of data points (or a cluster) and looking at *Detail* section, we get more detailed information on those points. By selecting a small/big cluster in the t-SNE plot, all Key metrics, heatmap in the *Context* section, and the visualizations in the *Detail* section of the LAD are updated accordingly. Parameters of the algorithm, such as the "Number of Iterations", "Perplexity", and "Learning Rate" can be manipulated with the controls on the left side of the screen (cf. figure 1).


**Figure 1:** The main screen of SAVis with all the three sections containing several visualizations.

### 3.1.2. Heatmap

A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors [14]. The heatmap, at the right of the *Context* section, shows the accuracy of a Random Forest classifier for detecting the students' university ID. To improve the performance of the Random Forest classifier, Synthetic Minority Over-sampling Technique (SMOTE) is used to oversample the minority university categories [15, 16]. The result of the high-performance Random Forest classifier shown in the heatmap gives confidence to an educator using the LAD that the patterns in the visualizations of the data belonging to a specific students' university ID are accurate and not random.

### 3.1.3. Detail Section

The ***Detail*** section of the dashboard contains five tabs with specific visualizations in each of them, to aid the discovery of insights about different features' categories. SAVis enables the user to interact with every visualization separately and drill-down at further levels of details. Most of the visualizations support interactive techniques such as *brushing*, *zoom*, and *filter*. The main

purpose of brushing is to highlight brushed data items in different views of the dashboard. By selecting a part of each visualization and zooming the view, the user can reach further details. Additionally, by clicking on the hued small objects (square or circle) in the right part of each visualization, the user can filter the view according to the correct/incorrect answers, the month, or the question type. Figure 2 displays the "Student Answer" tab. This tab contains a scatter plot, a pie chart, and a dropdown component. The number of correct and incorrect answers for the selected feature from the dropdown on the right side of the tab is shown in the scatter plot. By moving the mouse in the scatter plot, the user can get extra information about students. The Scatter plot showed in figure 2 allows identifying the number of correct and incorrect answers for every student, every topic, each month, each day, and every hour. Also, the pie chart on the right shows the overall percentages of correct and incorrect answers.

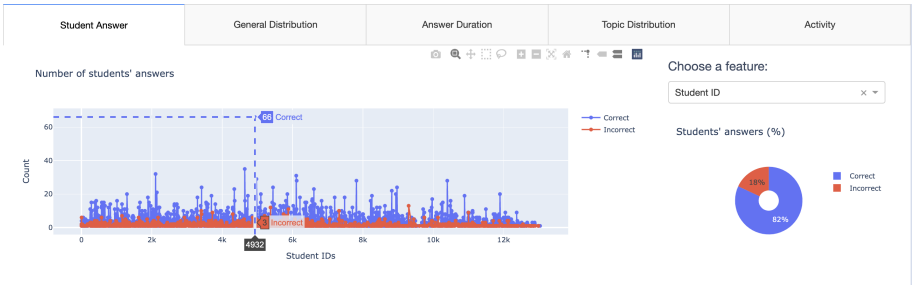


**Figure 2:** Student Answer tab: the first tab of the Detail section showing two visualizations.

The "General Distribution" tab showed in figure 3 illustrates the number of correct and incorrect answers for different features' categories shown in the radio-button list on the right side. By moving the mouse over the histogram and the strip plot on top of that, users can obtain more information about the number of correct and incorrect answers in relation to the different categories in the radio-button list, i.e., students, topics, questions, answer durations, month, day, and hour of students' activities.


**Figure 3:** General distribution tab: the second tab of the Detail section showing one visualization.

Figure 4 represents the "Answer Duration" tab. This tab illustrates the amount of time in minutes students spent answering questions and the distribution of answer duration. The Scatter plot on the left shows the students correct/incorrect answers for different educational topics over several months. The size of the circles in the view is dependent on the answer

duration. This view can help users to identify the minimum and maximum answer duration for correct and incorrect answers per month. The histogram on the right shows the mean and standard deviation of the 10,000 random samples. These values represent the average time in minutes for answering different topics' questions.

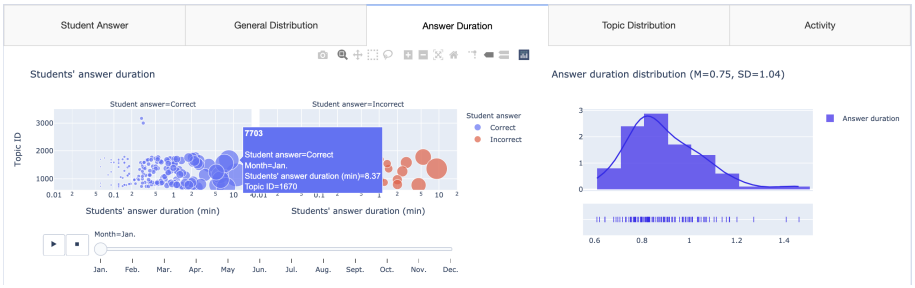


**Figure 4:** Answer duration tab: the third tab of the Detail section showing two visualizations.

Figure 5 shows the "Topic Distribution" tab which includes the distribution of topics for students' answers in general, students' correct answers more specifically, the top 10 topics that the students were more active in, and the number of correct and incorrect answers for the top 10 topics. In the two histograms on the left side of the tab, users can detect the number of students' answers and students' correct answers for different topic categories.

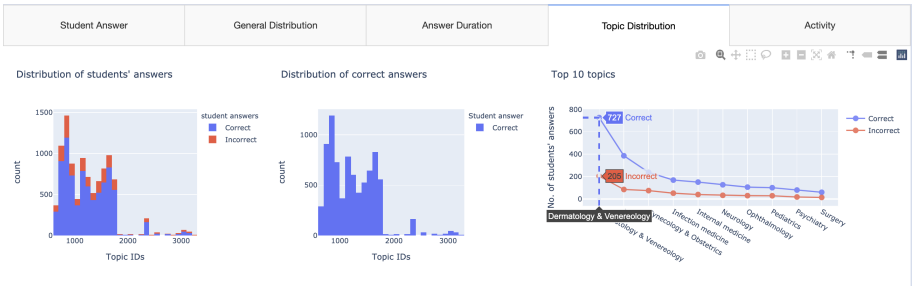


**Figure 5:** Topic distribution tab: the fourth tab of the Detail section showing three visualizations.

The "Activity" tab in figure 6 illustrates the percentages of students' activities during twelve months, the percentages of the students' activities in terms of the number of correct/incorrect answers in different question types, and the distribution of students' monthly activities in terms of the number of correct/incorrect answers for all educational topics. Viewing this tab enables users to find the question type that students have used more often, and the percentage of students' activities per month.

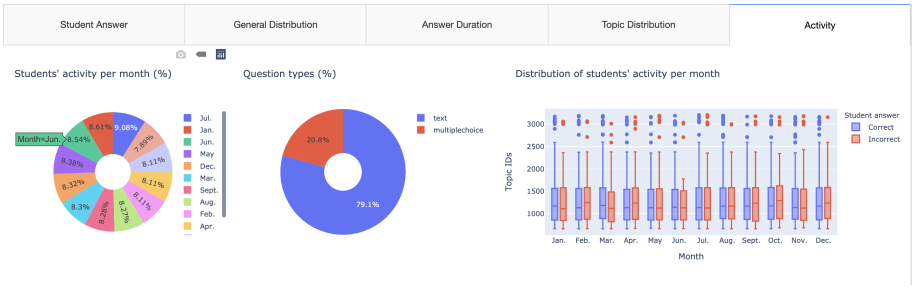


**Figure 6:** Activity tab: the fifth tab of the Detail section showing three visualizations.

## 4. Real-world Problem: How the LAD Can Answer 22 Pedagogical Questions

In this section, we provide a pedagogical explanation of the various visualizations of SAVis by attempting to answer the 22 pedagogical questions and explain how a user, in our case an educator/teacher, can make use of SAVis to get a broader picture of the students' learning activities. The ***Key metric*** section showed in figure 1 in the paper presents the number of correct and incorrect answers for 10,000 random samples out of the 1,322,097 available. These values are 8,196 and 1,804, respectively. Also, the minimum and the maximum answer durations are 0 and 19.73 minutes, respectively, for the selected samples. With this visualization, an educator could answer therefore PQ1 and PQ2 by simply looking at the key metrics of the dashboard. The heatmap showed in the ***Context*** section of the LAD allows teachers to answer PQ3, and according to figure 1, accuracy is 99.48%. This specific information helps to realize the accuracy of the provenience of the students. This high-accuracy classifier gives confidence in the data, meaning that a teacher can trust that the patterns in the visualizations are well-defined and match the students from the dataset. To answer the PQ4 – PQ22, the ***Detail*** section of SAVis should be explored (figure 2). The scatter plot showed in figure 2 helps answering PQ4 – PQ8 regarding the number of correct and incorrect answers for each individual student, every topic, each month, each day, and each hour. Table 1 shows different examples of how many correct and incorrect answers there are for PQ4 – PQ8, related to either the Student ID, the topic, the month, the day or the hour. By looking also at the pie chart presented in figure 2, teachers could answer PQ9 concerning the percentages of correct and incorrect answers in general. According to figure 2 the percentages of the correct and incorrect answers are 82% and 18%, respectively. The data in the "Student Answer" tab give an insight about individually student's progress, those students who work more than others, the popular educational topics, and when students' activity occurs. This sort of analysis can enrich the understanding of student activities so that an educator can better cater the pedagogical effort to most students. The histogram presented in figure 3 answers PQ10 – PQ17. Table 2 presents answers for PQ10 – PQ14, again showing unrelated examples of the number of correct and incorrect answers for each different category. In addition, by selecting "Monthly activity", "Daily activity" and "Hourly activity" from the radio-button list showed in figure 3, teachers are able to answer PQ15 – PQ17 zooming in the time, the day and the month students are more active, for example right before an examination

of a specific topic. The blue/red bar with the highest height in the histogram represents the month/day/hour with the highest activity. The data presented in this tab provide an educator with useful information on the learning progress and activity of a group of students taking a medical subject so that he/she can make an informed decision on possible course improvements for the next time. By looking at the line plot showed at the right side of figure 5 teachers are able to answer to PQ18 – PQ20. The data produced in this tab give information about the top 10 topics that students are answering more questions about, therefore being more active by taking more quizzes. By looking at two pie charts on the left side of figure 6 teachers could answer PQ21 and PQ22. As it can be seen figure 6, "text" with 79.1% is the most common question type for the 10,000 random samples. The percentage of students' activities for every month is comparable and ranges between 7.89%-9.08% for the 10,000 random sample. This outcome shows that the student's activity is equally spread every month, demonstrating that the student's workload is possibly evenly distributed and considered important by the student.

**Table 1**
Outcomes of PQ4 – PQ8

| Features | Selected item | No. correct Answ. | No. incorrect Answ. |
|---|---|---|---|
| Student ID | 4,932 | 66 | 3 |
| Topic ID | 1,708 | 112 | 24 |
| Month | Feb. | 655 | 156 |
| Day | 20 | 289 | 68 |
| Hour | 12 | 354 | 82 |

**Table 2**
Outcomes of PQ10 – PQ14

| Category name | Category | No. correct Answ. | No. incorrect Answ. |
|---|---|---|---|
| Student ID | 5,000 – 5,200 | 201 | 42 |
| Topic ID | 850-900 | 649 | 133 |
| Question type | Text | 6,628 | 1,287 |
| Answer duration | 0.35-0.44 | 746 | 170 |
| Monthly activity | Jul. | 733 | 175 |
| Daily activity | 24 | 250 | 72 |
| Hourly activity | 23 | 321 | 76 |

## 5. Conclusion and Future Work

In this paper, we described the design and development of SAVis, a new Learning Analytics Dashboard (LAD) by interpreting the visualization of ML algorithms to provide an interactive platform for teachers. The proposed LAD can be used to enable teachers to explore students' learning and activities by interacting with various visualizations of data. SAVis allows to compare groups of students as well as individuals on a different number of categories. An educator can choose which features to focus on while using SAVis and this allows for greater impact on educational issues rather than technical. An open challenge is to focus our future

research on a user study on the developed LAD to know more about teachers' needs with such a pedagogical instrument.

## References


[1] G. Akcapınar, M. N. Hasnine, R. Majumdar, B. Flanagan, H. Ogata, Developing an early-warning system for spotting at-risk students by using ebook interaction logs, Smart Learning Environments, Springer (2019).

[2] B. Daniel, Big data and analytics in higher education: Opportunities and challenges, British Journal of Educational Technology 46 (2015) 904–920. doi:10.1111/bjet.12230.

[3] A. L. Sonderlund, E. Hughes, J. Smith, The efficacy of learning analytics interventions in higher education: A systematic review, British Journal of Educational Technology 50 (2019) 2594–2618. doi:10.1111/bjet.12720.

[4] M. Aruvee, A. Ljalikova, E. Vahter, L. Prieto, K. Poom-Valickis, Learning analytics to inform and guide teachers as designers of educational interventions, International Conference on Education and Learning Technologies (2018). doi:10.21125/edulearn.2018.0666.

[5] G. Siemens, R. Baker, S. D., Learning analytics and educational data mining: towards communication and collaboration, LAK '12: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (2012) 252–254. doi:10.1145/2330601.2330661.

[6] A. Schwendimann, B., M. J. Rodríguez-Triana, A. Vozniuk, P. Prieto, L., M. S. Boroujeni, A. Holzer, D. Gillet, P. Dillenbourg, Perceiving learning at a glance: A systematic literature review of learning dashboard research, IEEE Transactions on Learning Technologies 10 (2016) 30–41. doi:10.1109/TLT.2016.2599522.

[7] K. Verbert, X. Ochoa, D. Croon, R., R. A. Dourado, T. D. Laet, Learning analytics dashboards: the past, the present and the future, LAK20: Proceedings of the Tenth International Conference on Learning Analytics and Knowledge (2020) 35–40. doi:10.1145/3375462.3375504.

[8] D. A. Keim, T. Munzner, F. Rossi, M. Verleysen, Bridging information visualization with machine learning, Dagstuhl Reports 5 (2015) 1–27. doi:10.4230/DagRep.5.3.1.

[9] S. Govaerts, K. Verbert, E. Duval, A. Pardo, The student activity meter for awareness and self-reflection, CHI '12 Extended Abstracts on Human Factors in Computing Systems (2012) 869–884. doi:10.1145/2212776.2212860.

[10] M. Ezzaouia, E. Lavoue, Emoda: a tutor oriented multimodal and contextual emotional dashboard, LAK' 17: Proceedings of the Seventh International Learning Analytics and Knowledge Conference (2017) 429–438. doi:10.1145/3027385.3027434.

[11] H. He, B. Dong, Q. Zheng, D. Di, Y. Lin, Visual analysis of the time management of learning multiple courses in online learning environment, 2019 IEEE Visualization Conference (VIS) (2019). doi:10.1109/VISUAL.2019.8933778.

[12] B. Shneiderman, The eyes have it: A task by data type taxonomy for information visualizations, Proceedings 1996 IEEE Symposium on Visual Languages (1996).

[13] V. D. Maaten, Visualizing data using t-sne, Journal of machine learning research (2008).

[14] A. Pryke, S. Mostaghim, A. Nazemi, Heatmap visualization of population based multi objective algorithms, International Conference on Evolutionary Multi-Criterion Optimization

(EMO 2007) (2007) 361–375.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357. doi:10.1613/jair.953.

[16] Z. Mohseni, R. M. Martins, M. Milrad, I. Masiello, Improving classification in imbalanced educational datasets using over-sampling, 28th International Conference on Computers in Education (APSCE) 1 (2020) 278–283.