DSLE: A Smart Platform for Designing Data Science Competitions

Giuseppe Attanasio, Flavio Giobergia, Andrea Pasini, Francesco Ventura, Elena Baralis,
Luca Cagliero, Paolo Garza, Daniele Apiletti, Tania Cerquitelli, Silvia Chiusano
Politecnico di Torino
Torino, Italy
(name.surname)@polito.it

***Abstract*—During the last years an increasing number of university-level and post-graduation courses on Data Science have been offered. Practices and assessments need specific learning environments where learners could play with data samples and run machine learning and data mining algorithms. To foster learner engagement many closed- and open-source platforms support the design of data science competitions. However, they show limitations on the ability to handle private data, customize the analytics and evaluation processes, and visualize learners' activities and outcomes.**

This paper presents Data Science Lab Environment (DSLE, in short), a new open-source platform to design and monitor data science competitions. DSLE offers a easily configurable interface to share training and test data, design group works or individual sessions, evaluate the competition runs according to customizable metrics, manage public and private leaderboards, monitor participants' activities and their progress over time. The paper describes also a real experience of usage of DSLE in the context of a 1st-year M.Sc. course, which has involved around 160 students.

***Index Terms*—Learning Analytics, Data Science, Data Analytics Challenges, Learning Platforms**

I. Introduction

Teaching data science fundamentals has become an established practice in several master degree and post-graduation courses related to computer science. Data science courses present the fundamental principles behind the automatic extraction of relevant knowledge from data. A key element of successful data science courses is the ability to stimulate learners' engagement. For this reason, the majority of them include laboratory practices, group works, and individual projects, where students can learn by doing practices. During data science practices, students usually experience on real and synthetic data, run machine learning and data mining algorithms, and evaluate and explore the achieved results in various ways [7].

Over the last decade, the Learning Analytics (LA) community has carried out a number of research projects tailored to higher-level courses [8]. Their common goal is to measure, collect, analyze, and report data about university-level students and the related learning contexts [9]. To enable the acquisition, storage, and analysis of learner-generated data, several intelligent systems have been proposed in literature. An overview of the existing environments is given in [10]. The conducted research has, for instance, deepened the knowledge about different learning mechanisms, explored the ways to stimulate students in order to improve their engagement, and early predicted the exam outcomes in order to minimize the probabilities of exam failure and students' dropout [11].

The LA community has paid a particular attention to non-conventional teaching and assessment methodologies. Under this umbrella, the research has demonstrated the positive effects of introducing competitions [12], [13] and serious games [14]–[16]. This work specifically addresses the use of competitions in higher-level data science courses. The advantages of introducing competitions in data science courses are manifold [10]: they foster the learning of best practices, stimulate the abilities in problem-solving, encourage creativity and group work, and give learners the chance to interact with new platforms and algorithms.

Several closed- and open-source platforms currently support the design of data science competitions. Each solution is designed to address specific needs. For example, Kaggle [1] supports the creation of single-phase challenges through a cloud-based workbench, but it does not allow handling private data. Other solutions (e.g., CodaLab [3], AutoComplete [2]) allow us to handle private data, but show limitations on the ability to customize the evaluation metrics. Recently proposed solutions (e.g., EvalAI [4], ParlAI [5]) are characterized by highly flexible environments, but are not designed to provide teachers with interactive interfaces to quickly collect usage statistics, generate result reports, and drill down on the activities of specific students. A thorough overview of the existing platforms is given in Section II.

This paper presents Data Science Lab Environment (DSLE), a new open-source platform to design and monitor data science competitions. To overcome the main limitations of existing platforms, DSLE is (i) easy to customize (even for the evaluation part), (ii) equipped with a multi-resolution interface for data analytics, visualization, and reporting, and (iii) open-source. The provided interface allows teachers to monitor usage statistics and explore students outcomes from different viewpoints. More specifically, DSLE includes specific modules devoted to (a) planning group works or individual sessions on public or private training data, (b) designing data mining and machine learning pipelines in a flexible way, (c) assessing the submitted results through various metrics (both conventional and not), (d) managing public and private leaderboards, (e)

TABLE I: Comparison between existing platforms

Feature	Kaggle [1]	AutoCompete [2]	CodaLab [3]	EvalAI [4]	ParlAI [5]	OpenML [6]	DSLE
Open-source platform	✗	✗	✓	✓	✓	✓	✓
Challenges on private data	✗	✓	✓	✓	✓	✓	✓
Multi-phase challenges	✗	✗	✓	✓	✗	✗	✓
Customize evaluation metrics	✗	✗	✗	✓	✓	✓	✓
Dashboards and reports	✗	✗	✗	✗	✗	✗	✓

monitoring participants' activities through interactive dashboards, and (f) producing interactive dashboards and reports on learners activities and competition results. In a nutshell, DSLE poses itself as an alternative, open-source solution to manage data science competitions. It is recommended in learning contexts, such as data science labs, in which there is a strong need for highly flexible solutions and integrated tools for learning data analytics and reporting.

The DSLE platform has been deployed and used in our university for a 1st-year M.Sc. course on data science. The course has involved approximately 160 students. As a case study, we present here the experience of usage, the collected statistics, and the feedback given by teachers and students, who have appreciated the platform usability and flexibility. The source code of the DSLE project is freely available[1]. We encourage interested readers to download and use the platform for learning purposes as well as to provide us with any useful comments or suggestions for future extensions.

The paper is organized as follows. Section II describes the existing platforms and discusses pros and cons of each of them. Section III thoroughly describes the DSLE architecture and the standard workflow involving teachers and learners. Section IV presents a usage experience of the DSLE platform. Finally, Section V draws conclusions and presents the future research directions.

II. Existing Platforms for Designing Data Science Competitions

a) Kaggle: Kaggle [1] (https://kaggle.com/) is probably the most popular closed-source platform for hosting data science and machine learning competitions. It supports the creation of a cloud-based workbench that could be shared with other users. To set up a data science competition, Kaggle requires to find a public dataset first. Challenges on private data are not allowed. Challenges consists of a single phase, i.e., they cannot be split into consecutive steps. The results achieved by the challenge participants on a training data sample are evaluated according to predefined metrics and published on a ranked list, i.e., the public leaderboard. This allows participants to assess the quality of the currently developed solution. The final evaluation is performed on a test sample (not in the public domain). The challenge organizer is provided with the private leaderboard (i.e., the performance rankings obtained on the private data sample).

b) AutoCompete: AutoCompete [2] is another framework to design machine learning competitions. It integrates a subset of data transformation and feature selection methods, a selection of regression and classification algorithms, different strategies for parameter tuning, and some standard performance validation metrics. However, the platform is not open-source, does not allow us to customize the key components, and is not aimed at supporting teachers in statistics visualization and result exploration.

c) CodaLab: CodaLab [3] (https://competitions.codalab.org/) is an open-source alternative to Kaggle. The platform allows researchers to run machine learning experiments by maintaining the full provenance of an experiment from raw data to final results. Furthermore, it allows us to design machine learning competitions that involve code and result submission and sharing. The data analytics pipeline is described using ad-hoc using worksheets or notebooks. It provides users with an evaluation platform supporting hosting competitions and benchmarking through a public leaderboard. Despite the platform partly supports project customization using external languages, the platform is intended neither to monitor learners' activities and performance over time nor to perform human-in-the-loop evaluations.

d) EvalAI and ParlAI: EvalAI [4] (https://evalai.cloudcv.org/) and ParlAI [5] (https://parl.ai/) extend CodaLab functionalities by providing highly customizable back-end solutions, with a particular focus on the evaluation phase. More specifically, the extensions proposed by [4], [5] provide human-in-the-loop interfaces to evaluate answers and analytics outcomes as well as a more sophisticated architecture with modular, customizable components. For example, ParlAI supports integration with Amazon Mechanical Turk to evaluate dialog models.

e) OpenML: OpenML [6] (https://www.openml.org/) is an open, Web-based platforms to share datasets, analytics code, and experimental results. It is mainly intended to support real-time, large-scale collaboration among data science learners or researchers. Despite it allows users to explore the analytics results, its main purpose is to manage data science science competitions.

A. Comparison between DSLE and the existing platforms

Table I summarizes the main properties of the existing platforms. Three of them (i.e., Kaggle, AutoCompete, CodaLab) show limited flexibility, especially when private data are used and the outcomes of the analytics process need to be evaluated. Conversely, EvalAI and ParlAI are mainly research-oriented. They provide basic teacher and student interfaces, which are not designed to summarize the results and to monitor learners' activities over time. The DSLE platform extends the function-

[1]Link: https://zenodo.org/record/3666486

alities of the previous solutions, with a particular emphasis on activity monitoring and result exploration, visualization, and reporting.

III. The Data Science Lab Environment platform

The DSLE platform aims at designing and monitoring data science competitions. Specifically, it has been designed to manage short-lasting competition covering various data science tasks, among which classification, regression, and clustering. The design and implementation of DSLE are aimed at maximizing platform flexibility and usability.

Figure 1 depicts the high-level architectural blocks of the DSLE platform. It manages the competitions by interacting with learners and teachers at various levels. It processes the submitted solutions through a predefined workflow. The processing phase entails computing the performance of the submitted solutions, comparing them with each other and with the baseline methods, making intermediate and final results public, and supporting teachers in monitoring learner activities and the outcomes of the submitted works.

The main features of the DSLE platform are summarized below:

- **Task-specific data preparation**: DSLE enables the use of public and private data. To simplify the analytics and assessment procedures, data are transformed, sampled, and split into different datasets according to the specific task to be addressed.
- **Customization of the evaluation metrics**: Beyond traditional metrics, DSLE allows teachers to integrate additional evaluation strategies tailored to the specific task under analysis. The implementations of both the analytics and evaluation processes are provided as Python scripts. The assessment of the participants' submissions can be also based on a comparative strategy. More specifically, the teacher could (i) analyze the distribution of the results of all the participants to rate the the quality of the solution, or (ii) implement or select a subset of reference baseline methods, which are then compared with the other submissions.
- **Public and private leaderboard management**: DSLE manages the submissions of the participants, stores them into a submission database, and keeps the public and private leaderboard updated. The former ranking is mainly devoted to supporting participants during the competition, whereas the latter is maintained for evaluation purposes.
- **Management of the competition workflow**: DSLE monitors the entire competition workflow, including the access control policies, the selection of the submissions to be evaluated, and the result visualization.
- **Computation and visualization of Key Performance Indicators**: DSLE supports the definition of a set of Key Performance Indicators describing (i) the utilization level of the platform, (ii) the student activities (individual or group activities), (iii) the properties of the submitted results. DSLE provides teachers with visualization and reporting functionalities on top of the defined KPIs. Specifically, it supports the creation of interactive dashboards, synthetic reports, and the computation of ongoing systems usage and performance statistics.

A more detailed description of each feature is given below.

A. *Data preparation*

The input dataset of each competition is split into two partitions: *development* and *evaluation*. The *development* set is used by participants to design a solution for the proposed competition, whereas the *evaluation* set is used by DSLE to automatically compute the quality scores achieved by the participant submissions.

The data preparation phase is tailored to the challenge goal. For example, for the classification and regression tasks the *evaluation* set contains the value of the target attribute for each evaluation record (i.e., the ground truth values). The evaluation of the submitted solutions is in charge of the DSLE platform. For example, to evaluate the submissions to classification or regression tasks, it considers the ground truth values reported in the evaluation set, which are hidden to participants, and compares them with the values predicted by the student solution to compute accuracy, precision, and recall measures [17].

The *development* set is further split in two subsets: public and private data. Public data are used to generate a public leaderboard. Participants can explore it to compare the quality of their solutions with respect to the other students' solutions when the competition is open. Conversely, private data are used to generate the leaderboard for the final evaluation and it is disclosed only at the end of the competition. The split in public and private data allows us to penalizing the solutions based on overfitting models.

B. *Customizable evaluation process*

The submission platform collects the solution files produced by the learners and evaluates them according to the scoring function selected in the configuration setting. The custom evaluation block is separately applied on public and private data. For example, for the classification and regression tasks public and private predictions are evaluated against the ground truth provided in the evaluation solution set to rate the overall quality of the designed solution. Beyond the standard evaluation metrics, the DSLE platform supports metric customization. A customized metric can be either a variant of existing ones or a new one defined through a dedicated Python script.

The assessment of the participants' submissions can be also based on a comparative strategy. More specifically, the teacher could (i) analyze the distribution of the results of all the participants to rate the the quality of the solution, or (ii) implement or select a subset of reference baseline methods, which are then used to evaluate the other submissions.

C. *Leaderboard management*

The DSLE platform stores in a database all the submissions along with their private and public scores and maintains both a private and a public leaderboard. During the competition,

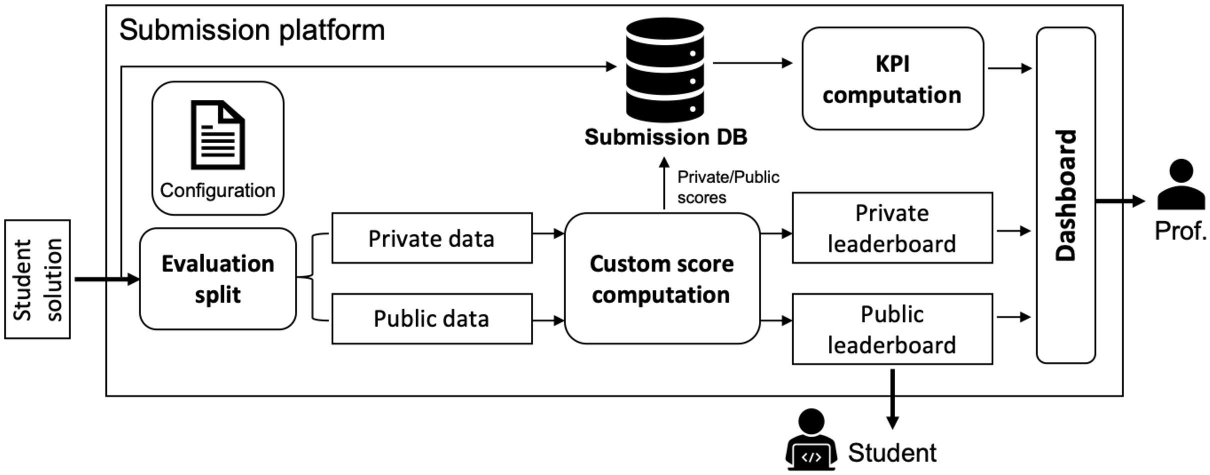


Fig. 1: Data Science Lab Environment: architecture.

learners can check the public leaderboard to compare the quality of their solutions with those of other submissions. Conversely, the access to the private leaderboard is restricted and allowed to teachers exclusively for evaluation purposes.

Based on the outcomes achieved on public data, for each competition learners pick their most promising solutions. The preferred solutions should not only maximize the performance on public data, but also be general enough to be potentially effective on private data as well. In case they do not specify any preferred submission, the private score for the final leaderboard will be computed by exploiting the solution with the highest public score.

Figure 2 shows the Web page through which learners select their submissions. Each row corresponds to a different submission. It contains the submission timestamp, public score, and a checkbox (to give the preference).

The leaderboard and the private baselines can be jointly used to assess the quality of the submitted solutions. Specifically, the final evaluation of the students can be performed by combining the ranking of the students in the leaderboard and the improvements with respect to the baselines. Hence, teachers can combine a competitive component, which is influenced by the quality of the competing students, with a non-competitive one, which depends on the overall quality of the submitted solutions compared to a baseline. Teachers could take both components into account during the evaluation process.

D. Management of the competition workflow

Figure 3 describes the general workflow of the competitions managed by the DSLE platform. At the beginning of the competition students are provided with (i) the text with the description of the competition, (ii) a dataset to be analyzed and (iii) a personal API key (one for each student). To control the access and the submissions to the DSLE platform, each participant is provided with a personal key. The personal key is requested to complete the submission and to log in the personal Web area. Each competition has peculiar characteristics (e.g.

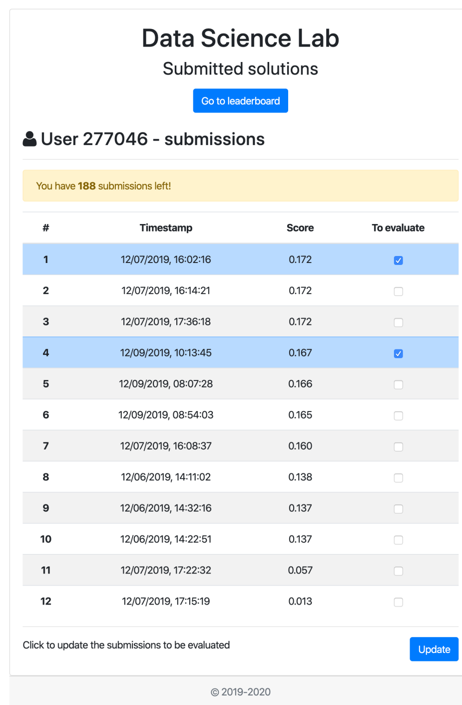


Fig. 2: Submission selection.

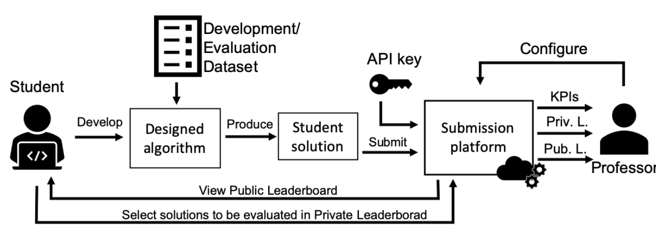
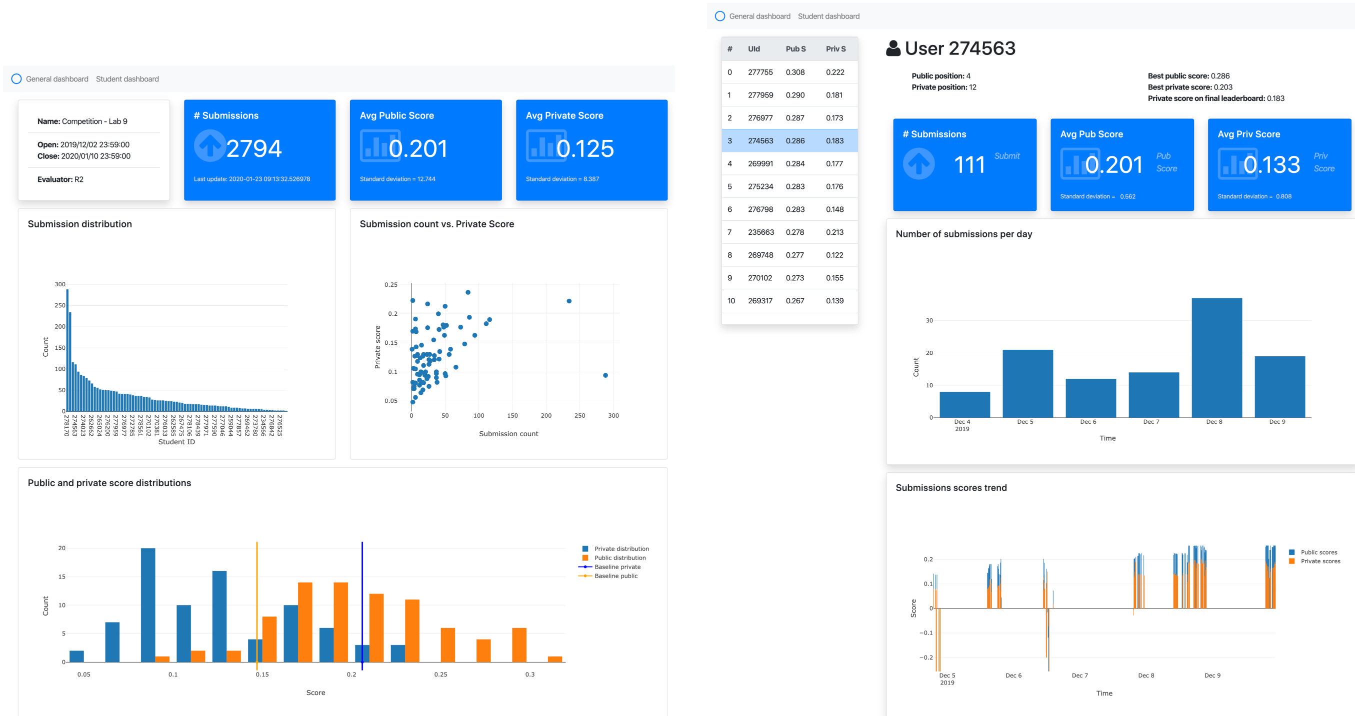


Fig. 3: Data Science Lab Environment: competition workflow.

data types, data structure, target variables, required tasks). Hence, the description of the competition includes all the



(a) Dashboard with high-level statistics on the competition.

(b) Dashboard with student-specific statistics.

Fig. 4: Data Science Lab Environment: dashboard panels accessible by the teacher.

necessary information such as the details of the dataset (e.g. structure, features meaning), the target of the competition (e.g. the target variable to be predicted, the objective of the unsupervised analysis), and the format of the solution file to be submitted. Before the beginning of the competition teachers have to set up the configuration files containing the details of the competition (e.g. start date, closing date, custom scoring function, format of the solution file). The configuration files are used by the platform to instantiate the competition.

At the end of the analytics phase learners have to produce a solution file adhering to the competition guidelines. The submission guidelines, including the format of the result files, are specified in the competition description. The submission can be accomplished by accessing to the submission platform with the personal API key. The results uploaded by the learner during this phase will be automatically evaluated. Each leaner has a maximum number of submissions per competition. This hinder students to apply a *trial-and-error* approach. To check the number of submissions that are currently available, each learner can access the personal Web page.

The submission files sent out by the participants are stored in a relational database. The database stores also the evaluations that are automatically computed by the DSLE platform. In parallel, an historical knowledge base of events is created to support the efficient computation of Key Performance Indicators (see Section III-E).

E. Computation and visualization of Key Performance Indicators

During data science competition teacher would like to monitor how student performance evolve over time. For instance, if the achieved results get better then the quality of the adopted solution is likely to improve as well. This may indicate an improvement of the level of knowledge of the learners.

DSLE allows teachers to analyze the historical submission data through a subset of Key Performance Indicators [18]. They summarize the historical usage statistics and competition outcomes thus helping teachers to understand learners' comprehension and engagement levels.

KPI computation is expert-driven. Teachers could dynamically modify the KPI definitions to tailor them to their actual needs. KPI visualization entails generating a set of interactive dashboards describing the leaderboard, the platform usage statistics, as well as detailed statistics on participants' activities. Specifically, DSLE currently supports a multi-resolution dash-boarding interface: (i) a high level view of a competition by analyzing the general competition dashboard (see Figure 4a), and (ii) a detailed, student-centric, view by means of the student dashboard (see Figure 4b).

The high-level dashboard (Figure 4a) graphically summarizes the following information: (i) name of the competition, (ii) open and close timestamps, (iii) evaluation metric, (iv) total number of submissions, (v) average score (achieved on both public and private leaderboards), (vi) distribution of the total number of submission per student, (vii) distribution of the private scores normalized by the number of submissions, and (viii) private and public score distributions.

The global KPIs and charts ((iv)-(viii)) are mainly used for monitoring the competition during its opening period. They allow us to understand the complexity of the proposed task and to what extent the competition has been popular. They are expressed in terms of number of submissions, number of participants, distribution of the students per number of submissions, the presence of outliers (i.e., students who are significantly more active than others). The latter indicators allows teachers to detect anomalous situations, such as learners that are likely to apply a *trial-and-error* approach.

The submission count vs. private score scatter plot helps us to study the correlation between the scores obtained in the private leaderboard and the number of performed submissions.

The high-level dashboard provides also information about the rankings of the students with respect to the baselines for both public and private rankings (see the public and private score distributions in Figure 4a). For example, the public and private score distribution chart allows us to understand several interesting characteristics about the competition. First, we can figure out whether the per-student scores are uniformly distributed or if they are bounded on private and public data. Furthermore, the solutions that are too much tailored to the training data can be identified. For example, if the public score distribution is shifted on highest score values whereas the private score distribution is skewed on relatively low values, the submission is likely to suffer from data overfitting.

Teachers can also use the student dashboard to understand how students decided to tackle the competition (see Figure 4b). By selecting a student, her/his details are showed in the main section of the page. Specifically, the student dashboard provide the following information for each user: (i) ranks in the public and private leaderboards, (ii) best public and private scores, (iii) total number of submissions, (iv) average public and private scores, (v) number of submissions per day, and (vi) temporal trends of submission scores.

By analyzing the aforesaid user-specific statistics, teachers can understand, for instance, if the solutions submitted by one specific student are overfitting the public part of the evaluation set or if a specific student is submitting solutions that are general enough (high performance on both public and private subsets of the evaluation set). Moreover, these charts allow us explore the progress of the students over time. They indicate whether she/he has improved the quality of her/his solutions over time and the corresponding amount of effort (i.e., number of submissions and rankings over time). By analyzing these charts, teachers can also understand if a student is trying significantly different approaches, which are highlighted by significant increases or decreases in terms of scores of the submitted solutions over time. See for instance the positive and negative peaks in Figure 4b. They are representative of different approaches tested by a specific student over time.

To the best of our knowledge, a detailed set of statistics and charts about students' progress and activities over time like the ones computed by DSLE are not provided by the available platforms (at least not in the open source or free versions).

DSLE is developed entirely in Python and leverages lightweight server and database components. Along the three competitions hosted on DSLE, we achieved 100% up time, with no significant issues for students, on a single *t2.micro* Amazon AWS[2] EC2 instance.

IV. Use Case: an Experience of Data Science laboratories in a M.Sc. course

This section describes an experience of usage of the DSLE platform in a M.Sc. data science course. First, it describes the course organization and the competitions. Then, it provides some statistics about the usage of DSLE platform. Finally, it summarizes the results of a survey presented to the students of the course providing a comprehensive evaluation of the organization of the course, the usability of the platform, and the usefulness of the proposed competitions for learning purposes.

A. Overview

We used the DSLE platform in the context of an introductory data science course for students of a M.Sc. course in data science. The course is organized as follows: (i) 50 hours of frontal lectures, providing the basic concepts about machine learning and data mining algorithms, and an overview of the main python-based machine learning libraries, and (ii) 30 hours of practices in laboratory.

Along the course period, we set up four competitions. Three of them were part of laboratory sessions, whereas the latter was included in the final examination. One of the laboratory competitions was set up using Kaggle [1], thus allowing students to compare the proposed DSLE platform with a renowned one. More specifically, for each main topic of the course, we set up two practices: a "standard" session, during which students had to solve a data science problem in collaboration with other students, and an individual competition. Hence, during the laboratory sessions, we organized three individual competitions: two competitions have been managed by using DSLE and one by using Kaggle [1].

The total number of enrolled students is 200. On average, half of them took part to the competitions (the participation to the competitions was recommended but not mandatory). Approximately 160 students attended the final exam competition. Table II reports the detailed statistics on platform usage, number of participants, and total number of submissions per competition.

TABLE II: DSLE usage statistics

Task	Platform	Participants	Submissions
Clustering	DSLE	118	7103
Classification	Kaggle	94	1372
Regression	DSLE	89	2808
Classification	DSLE	158	6784

[2]https://aws.amazon.com/

B. Organized competitions and associated tasks

The program of the course covers a large spectrum of data mining and machine learning concepts and algorithms. For example, it includes the most established techniques to perform frequent pattern mining, clustering, classification and regression. We have presented individual competitions concerning the following topics: (i) clustering, (ii) classification, and (iii) regression. The aim is twofold: explore various techniques, data types, and domains during the practices and stimulate lecture attendance (which is not mandatory) to learn theoretical concepts needed to solve the practical problems. The task considered in the examination was about classification.

The following subsections provide further details on each competition.

1) Clustering competition: It entails clustering textual news data based on the covered topic. We provided students with raw textual documents thus fostering the use of preprocessing techniques to prepare the input data. We deem such a preparation step as crucial, especially when coping with textual data.

The considered dataset was a sample of a larger benchmark dataset [19] including approximately 20,000 documents, grouped by topic in 20 newsgroups. We picked 4 representative topics and applied a uniform sampling to generate a heterogeneous document collection. The aim of the clustering task was to separate the documents into the original groups. Notice that the task is completely unsupervised since the number of topics and the news article labels are hidden.

The quality of the generated clusters was evaluated, against the ground truth, using the Adjusted Rand Index [17]. Since the participants have no labeled data, they could only estimate the quality of their results based on cohesion metrics or extract word clouds or word co-occurrences to identify similarities or patterns in the textual content.

The main issue encountered during the competition was the partial misalignment between the outcomes produced during the Rand index and those produced by the standard clustering quality metrics (e.g., Silhouette [17]). This negative effect was partly mitigated by the ability of DSLE to effectively handle multiple quality metrics. To comparatively assess students' results, we also generated a baseline solution. It uses the TF-IDF vectorization [20], a Singular Value Decomposition on the resulting matrix, and then applies the established *k*-means clustering algorithm [17].

Figure 5 shows several statistics from the DSLE dashboard. Specifically, from Figure 5a we can see that the majority of students outperformed the baselines, and that there was a significant number of participants who achieved an adjusted rand index close to 0.8. Figure 5b, instead, highlights an expected (yet not assured) behavior: the higher the submission count, the higher the average private score and the lower the variance. This also indirectly shows that the quality of the submissions has improved over time.

2) Classification competition: This competition was handled using *In-Class* Kaggle challenge with the goal of providing students with two competition environments and get their *a posteriori* evaluations. The proposed task was about the identification of uttered digits out of audio signals. The provided dataset was inspired by the Free Spoken Digit [21]. We sampled (with stratification) 2,000 recordings of numbers from 0 to 9 uttered by 4 distinct speakers with English pronunciation. Then, we provided 1500 samples as Development set, whereas the remaining 500 were used as evaluation set.

The competition task consists in designing and implementing a classification pipeline to label every audio signal in the Evaluation set with a number from 0 to 9. Students submissions were evaluated using the accuracy score [17]. In this case, the baseline was generated with a sliding window approach on normalized signals. For each time window, several statistical metrics were computed, such as the mean and standard deviation on amplitude values, and used to build a relational dataset. The latter was then used to train a Random Forest classifier.

The input data appeared to be quite heterogeneous, due to the characteristics of the recording environments or devices. Hence, the main issue was to properly choose the features used to well discriminate among the class labels.

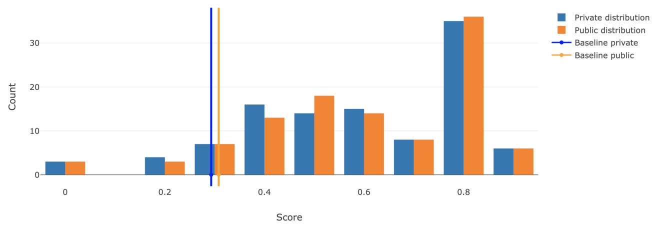
The number of participants keeps approximately stable compared to the previous clustering competition (see Table II). However, the number of submissions was significantly lower. This is mainly due the strict limitation enforced by Kaggle *In-Class* on the maximum number of submission per day available to each user (20). Unlike Kaggle, the DSLE platform allows teachers to configure the competition environment in a more flexible and effective way (e.g., they could set a cool-down time between two consecutive submissions).

3) Regression competition: This competition was based on a real relational dataset acquired in 2019 by Airbnb[3]. Each dataset record corresponds to a different rental listing and contains various contextual information such as the rental price, the number of reviews, and the geographical position associated with the rental. Students were asked to build a robust regression model to predict the price of the rental listing. The quality of the results was qualitatively evaluated using the coefficient of determination (R^2) [17]. Notice that we did not apply any sampling or transformations of the original dataset to foster the use of data cleaning approaches, feature selection, and engineering. In this case, the baseline pipeline consisted in normalizing the numerical attributes, one-hot encoding the categorical ones and applying a Random Forest regressor [22].

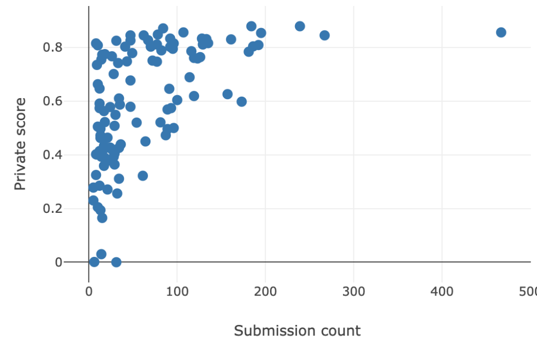
Figure 4a visualizes the submission statistics through the DSLE administrator dashboards. As for the clustering assignment, considering all the participants, a higher number of submissions led to a higher average R^2, while decreasing the variance.

Figure 4b shows the statistics of a representative student. She/he has mainly performed the submissions on the second day and in the last two days of the competition. The chart on the lower part of the figure shows the public score has

[3]http://insideairbnb.com/



(a) Distribution of private and public scores.



(b) Private scores in function of the submission count.

Fig. 5: Data Science Lab Environment: dashboard charts from the clustering competition.

increased over time, but still some low-quality solutions have been generated (unsuccessful trails with negative R^2 scores). The student has probably explored various analytic pipelines during the first half of the time, while has put the effort on tuning the performance of the most promising strategy in the second half.

4) Exam competition: The exam of the winter session has included a competition, managed by the DSLE platform, on a classification problem.

The task focused on sentiment analysis of textual reviews. Specifically, we crawled from the TripAdvisor website (https://www.tripadvisor.it/) around 40,000 textual reviews of hotels and apartments, written in the Italian language. Each review is associated with a binary label, expressing the reviewer sentiment (i.e., positive or negative). Labeled data are rather imbalanced (67% positive, 33% negative). As for the clustering competition, the baseline pipeline consisted in using the TF-IDF vectorization, followed by an SVD on the resulting matrix. The preprocessed data were then used to train a Random Forest classifier.

Participants achieved an average weighted F1-score [17] equal to 0.966 on private data (standard deviation = 0.011). Since the competition was mandatory for passing the exam, as expected, the number of participants was significantly higher than those of the previous competitions.

C. Survey outcomes

At the end of the course we presented a survey consisting of 18 multiple-choice questions and a comment box to provide further comments. The aim was to evaluate the level of satisfaction of the students about the course, with particular emphasis on the competition and on the interactions with the DSLE platform. Notice that in our country using competitions in the practices and in the final examination is quite uncommon. Hence, we would like to test the feeling of the students about the use of non-traditional learning methodologies.

68 students completed the survey. Table III reports a survey extract including questions related to the practices carried out in the laboratories, the competitions, and the usability of the platforms (DSLE and Kaggle). For each question, the distribution of the student answers are also given.

More than 89.7% of the students attended at least 50% of the laboratory sessions and 89.7% of them took part to at least one competition. 44.1% of the students deemed the participation to the internal competitions “definitely useful” whereas the percentage of students who answered “definitely useful” for the traditional practices is only 33.8%. Alternating “traditional” labs with internal competitions was considered at least a good choice by 64.7% of the students (48.5% good and 16.2% excellent, respectively).

Regarding the usability of the proposed platform, the students considered DSLE as highly usable (57.4% good and 22.1% excellent). The comparison with Kaggle is substantially a tie (76.4% of the students consider the platforms comparable). These results indicate that, from the users’ viewpoint, DSLE and Kaggle are quite similar and interchangeable. However, DSLE improves the teacher experience thanks to the teacher-oriented interface that provides various features, allowing deeper explorations of platform usage statistics and competition results.

V. Conclusions and discussion

In this paper we have presented a new open-source platform for organizing data science competitions. The DSLE platform has been specifically designed for managing and evaluating competitions among higher-level students. DSLE extends existing platforms by allowing teachers to gain insights into students’ activities and submitted results. Specifically, it produces dashboards and reports based on various Key Performance Indicators, which can be particularly useful for teachers who are interested in monitoring learners’ activities, comprehension of theoretical concepts, and problem solving attitudes.

We have conducted an experience of usage of the DSLE platform in a real university-level educational context. The experience has allowed us to compare the usability of DSLE with that of a renowned platform (i.e., Kaggle). DSLE has

TABLE III: Survey outcomes

Question	**Answers**			
How many laboratories have you attended?	None (1.5%)	From 1 to 4 (8.8%)	From 5 to 8 (29.4%)	From 9 to 10 (60.3%)
Considering the three labs associated with an internal competition, how many of them have you attended?	None (10.3%)	1 out of 3 (14.7%)	2 out of 3 (29.4%)	3 out of 3 (45.6%)
Was traditional lab attendance useful?	No (14.7%)	Partly (51.5%)	Definitely yes (33.8%)	
Was the participation to internal competitions useful?	No (11.8%)	Partly (44.1%)	Definitely yes (44.1%)	
Please rate the choice of alternating "traditional" labs with competitions on specific data mining tasks.	Bad (5.9%)	Fair (29.4%)	Good (48.5%)	Excellent (16.2%)
Please rate the usability of the internal platform used for handling submissions.	Bad (2.9%)	Fair (17.6%)	Good (57.4%)	Excellent (22.1%)
Please rate the usability of the platform used for handling submissions compared to those provided by Kaggle.	Worse (11.8%)	Comparable (76.4%)	Better (11.8%)	

shown to be more flexible than Kaggle because it handles private data, applies customized evaluation measures, and visualizes summarized charts and reports on students' activities and progresses.

According to the outcomes of a comprehensive survey, we got positive feedback from both teachers and students on the DSLE platform. On the one hand, teachers were able to identify interesting behavioral patterns (e.g., students who worked more intensively at the competition start, or close to the end, or who worked consistently day by day). They have also identified critical situations (e.g., students with consecutive poor scores, due to methodological issues and implementation errors) and have monitored the students' progress over time. On the other hand, students have positively rated the usability of the platform and the usefulness of the proposed data science challenges.

As future work, we plan to extend DSLE with additional dashboards and interactive tools. For example, we plant to monitor and visualize the temporal distributions of the public and private scores. Since the submission database already contains low-granularity data, new dashboards and reports can be easily introduced. In the near future, we will foster contributions from external teams and single users to keep improving the visual analytics interface of the open-source platform. To this aim, we encourage interested readers to download and use the platform and give us useful comments or suggestions for future extensions.

References

[1] R. Tatman, "Setting up your public data for success," in *2018 IEEE International Conference on Big Data (Big Data)*, Dec 2018, pp. 3261–3262.

[2] A. Thakur and A. Krohn-Grimberghe, "Autocompete: A framework for machine learning competition," *CoRR*, vol. abs/1507.02188, 2015. [Online]. Available: http://arxiv.org/abs/1507.02188

[3] P. Liang, I. Guyon, E. Viegas, S. Escalera, X. B. Solé, and E. Carmichael, "Codalab," 2020. [Online]. Available: https://competitions.codalab.org/

[4] D. Yadav, R. Jain, H. Agrawal, P. Chattopadhyay, T. Singh, A. Jain, S. B. Singh, S. Lee, and D. Batra, "Evalai: Towards better evaluation systems for ai agents," 2019.

[5] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston, "Parlai: A dialog research software platform," *arXiv preprint arXiv:1705.06476*, 2017.

[6] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "Openml: Networked science in machine learning," *SIGKDD Explor. Newsl.*, vol. 15, no. 2, p. 49–60, Jun. 2014. [Online]. Available: https://doi.org/10.1145/2641190.2641198

[7] S. Slater, S. Joksimovic, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for educational data mining: A review," *Journal of Educational and Behavioral Statistics*, vol. 42, no. 1, pp. 85–106, 2017. [Online]. Available: https://doi.org/10.3102/1076998616666808

[8] Y.-S. Tsai and D. Gasevic, "Learning analytics in higher education — challenges and policies: A review of eight learning analytics policies," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, ser. LAK '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 233–242. [Online]. Available: https://doi.org/10.1145/3027385.3027400

[9] G. Siemens, "Learning analytics: The emergence of a discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, 2013. [Online]. Available: https://doi.org/10.1177/0002764213498851

[10] B. K. Daniel, "Big data and data science: A critical review of issues for educational research," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 101–113, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12595

[11] G. Siemens and R. S. J. de Baker, "Learning analytics and educational data mining: towards communication and collaboration," in *Second International Conference on Learning Analytics and Knowledge, LAK 2012, Vancouver, BC, Canada, April 29 - May 02, 2012*, 2012, pp. 252–254. [Online]. Available: https://doi.org/10.1145/2330601.2330661

[12] L. Benotti, M. C. Martnez, and F. Schapachnik, "A tool for introducing computer science with automatic formative assessment," *IEEE Transactions on Learning Technologies*, vol. 11, no. 2, pp. 179–192, April 2018.

[13] Y. Jiugen, X. Ruonan, and L. Zhihuo, "Process model construction of the college students' competition data mining," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, March 2017, pp. 207–210.

[14] M. Ibanez, A. Di-Serio, and C. Delgado-Kloos, "Gamification for engaging computer science students in learning activities: A case study," *IEEE Transactions on Learning Technologies*, vol. 7, no. 3, pp. 291–301, July 2014.

[15] Y. Wang, P. Rajan, C. S. Sankar, and P. K. Raju, "Let them play: The impact of mechanics and dynamics of a serious game on student perceptions of learning engagement," *IEEE Transactions on Learning Technologies*, vol. 10, no. 4, pp. 514–525, Oct 2017.

[16] M. Stephenson, E. Piette, D. J. N. J. Soemers, and C. Browne, "Ludii as a competition platform," in *2019 IEEE Conference on Games (CoG)*, Aug 2019, pp. 1–8.

[17] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining (2nd Edition)*, 2nd ed. Pearson, 2018.

[18] M. Golfarelli and S. Rizzi, *Data Warehouse Design: Modern Principles and Methodologies*, 1st ed. USA: McGraw-Hill, Inc., 2009.
[19] K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 331–339.
[20] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press, 2008.
[21] Z. Jackson, C. Souza, J. Flaks, Y. Pan, H. Nicolas, and A. Thite, "Jakobovski/free-spoken-digit-dataset: v1.0.8," Aug. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1342401
[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.