UNIFORM: Automatic Alignment of Open Learning Datasets

UNIFORM: Automatic Alignment of Open Learning Datasets

Luca Cagliero, Lorenzo Canale, Laura Farinetti
Politecnico di Torino
Torino, Italy
(luca.cagliero, lorenzo.canale, laura.farinetti)@polito.it

***Abstract*—Learning Analytics aims at supporting the understanding of learning mechanisms and their effects by means of data-driven strategies. LA approaches commonly face two big challenges: first, due to privacy reasons, most of the analyzed data are not in the public domain. Secondly, the open data collections, which come from diverse learning contexts, are quite heterogeneous. Therefore, the research findings are not easily reproducible and the publicly available datasets are often too small to enable further data analytics. To overcome these issues, there is an increasing need for integrating open learning data into unified models.**

This paper proposes UNIFORM, an open relational database integrating various learning data sources. It presents also a machine learning supported approach to automatically extending the integrated dataset as soon as new data sources become available. The proposed approach exploits a classifier to predict attribute alignments based on the correlations among the corresponding textual attribute descriptions.

The integration phase has reached a promising quality level on most of the analyzed bechmark datasets. Furthermore, the usability of the UNIFORM data model has been demonstrated in a real case study, where the integrated data have been exploited to support learners' outcome prediction. The F1-score achieved on the integrated data is approximately 30% higher that those obtained on the original data.

***Index Terms*—Learning Analytics, Classification, Data Integration**

I. Introduction

The diffusion of Learning Management Systems in schools and universities today allows educational institutions to digitally archive students' data and to extract knowledge useful for gaining insights into the learning process. Learning Analytics (LA) focuses on analyzing learner-generated data in order to accomplish different tasks, such as to predict students' outcomes, identify students at risk of dropout, or measure the students' engagement level [1].

The recent advances in LA have produced a relevant number of learning systems and methodologies that rely, to a large extent, on learner-generated data. However, as mentioned in [2], the majority of the studies in this field have been carried out on datasets that are not publicly available. In fact, due to privacy restrictions or internal policies, many institutions do not provide their data outside. Hence, the achieved results are not reproducible. Besides, the publicly available datasets are quite heterogeneous in terms of the collected type of data. For example, some of them mainly focus on the students' interactions with the Learning Management System, others on the exam outcomes, still others on the student-teacher or peer-to-peer interactions. The dataset size and schema are also largely variable. As discussed in [3], the open datasets, considered individually, are often too small to ensure robust results for the research tasks. In this context, finding real benchmark data to reliably test learning analytics approaches and algorithms is quite hard. It is hence very difficult to define a coherent and comprehensive roadway to the future of the field, that goes beyond interesting and inspiring, but limited in scope, experiments.

To tackle the aforesaid issues, some attempts to perform manual integration of open learning datasets have been performed. An overview of the related literature is given in [4]. Producing an integrated database that includes learning data related to various field allows researchers to improve the robustness of the performed analyses. However, the integration process is extremely time-consuming. Hence, there is a need for (semi-)automatic solutions to integrate learning data acquired from various contexts.

The present paper proposes UNIFORM, a integrated relational database whose schema includes tables and attributes describing learning data from various open sources. UNIFORM is designed to handle heterogeneous data and to enable further extensions in a flexible way. The educational context, in fact, can vary depending on educational level, national policies, cultural characteristics and so on, thus the integration of new datasets could likely require a dynamic adaptation of the integrated schema.

Since manual alignment of open datasets to the UNIFORM schema is a time-consuming and not scalable task, the paper also proposes a machine learning supported methodology to automatically align new datasets without human intervention. To this aim, a classification model is trained on a partially aligned dataset version. The classifier predicts the most likely attribute alignments in the UNIFORM dataset based on a correlation analysis performed on the textual attribute descriptions.

The procedure of automatic alignment with UNIFORM has been successfully accomplished on 11 well-known, open learning datasets. The performance of the automated integration phase are quite promising: to predict attribute matches an established ensemble method (i.e., Random Forest Classifier [5]) has reached a F1-score equal to or above 70% on 6 out of 7 datasets.

To highlights the pros of using the integrated data model in the analytics process in place of the original data, we have addressed also, as representative case study, the task of early predicting exam outcomes. Specifically, we have trained classification models to predict students' outcomes based on data extracted the integrated version. The generated predictions are significantly more precise and sensitive (i.e., F1-score +30%) than those produced by similar models trained on the original data.

The paper is organized as follows: first, in Section II we describe the datasets on top of which the UNIFORM schema has been generated. Section III overviews the related scientific literature. Next, Section IV presents the UNIFORM schema and the manual alignment procedure, respectively. Then, Section V presents the automatic alignment algorithm and evaluates its performance. Finally, Section VI describes a real case study tailored to student outcome prediction, whereas Section VII draws the conclusions and presents the future research directions.

II. Datasets description

To build the proposed UNIFORM schema, we started from 10 publicly available datasets that contain learning data and from Our Institution Dataset (hereafter denoted as OID). The public datasets are enumerated below.

- OULAD[1] (Open University Learning Analytics Dataset), which contains data about student interactions with a learning management system.
- HARVARDX[2] and MITX[3] Dataverse, which contain the descriptions of the student activities in one edX platform course.
- COURSERA Forums[4], which contain discussion threads presented in the forums of Coursera MOOCs.
- PORT dataset[5], which collects students' performance data in two secondary schools in Portugal.
- xAPI-Edu-Data[6] (XAPI), which consists of data about student behavior acquired in the University of Jordan.
- EPM[7], which contains information about student grades and behaviors during the interactions with online resources at the University of Genova.
- EDSA[8], which contains data about students' interactions with the online resources of the European Data science Academy portal.
- ISTM[9], which contains the students' answers to a set of survey questions about time management at Nottingham Trent International College.
- UoJ[10], which contains data about student performance. Despite the stored data are not real, we decided to include this dataset anyway because in this paper we are mainly interested in the dataset schema rather than on the data instances.

OUD (Our Institution Dataset) contains data about B.S. student performance and accesses to online educational resources, including video-recorded lectures.

Table I synthetically describes the characteristics of each dataset, outlining the type of contained data as well as some relevant statistics about data size and schema complexity (i.e., dataset size expressed in MB, number of tables). The content of the datasets is synthesised by exploiting the following data descriptors: (a) SPD (Student Personal Data), e.g. personal ID, age, gender, ethnicity; (b) SCD (Student Career Data), e.g. school degrees, entry test grades, educational modules enrollment; (c) EMD (Educational Module Data, e.g. available courses, course description, course prerequisites; (d) SAD (Student Assessment Data), e.g. exam grades, intermediate assessment evaluations; (e) ERA (Educational Resource Access), e.g. activities within a learning management system, online resources access, video-lectures streaming; (f) IAD (Interaction Activity Data), e.g. forum posts, peer-to-peer interactions, student-teacher interactions. The table clearly shows the heterogeneity of the analyzed open datasets, in terms of schema, focus, and complexity.

III. Related work

A large number of learning analytics research papers exploit educational data collected in proprietary datasets. They are not in the public domain, mainly due to privacy issues or internal policies. Papers often face research issues of universal importance (such as understanding learner engagement factors, preventing school drop-out, improving student performance, personalizing learning environment, and so on) but often outcomes are validated on datasets that only reflect a very specific learning context. The actual reusability of the research results is quite limited, because it is difficult to replicate and analyze the previously published findings.

Some literature works, however, base their findings on publicly available datasets. Table II resumes, for each of the datasets listed in Section II, scientific papers that used them for a specific learning analytics task. The tasks covered in these papers are enumerated below:

- Prediction of school dropout.
- Identification of at-risk students.
- Prediction of student performance.
- Analysis of student engagement levels.
- Study of the learning process (e.g. using learning analytics to experimentally validate pedagogical models).

Table II contains references to literature papers, when available; it contains "S" (for Suitable) when there is no published work (to the best of our knowledge), but the dataset is suitable for addressing such a task.

[1]https://bit.ly/2m4a0NF
[2]https://bit.ly/2FLEz3f
[3]https://bit.ly/314niIv
[4]https://bit.ly/2mVuOas
[5]https://bit.ly/2lmoFDC
[6]https://bit.ly/2lmp2y0
[7]https://bit.ly/2ltgwgU
[8]https://bit.ly/2mc0NTG
[9]https://bit.ly/2me1HyT
[10]https://bit.ly/2mxrq5L

In [6] the *prediction of school drop out* task is performed using the OULAD dataset, while in [7] the same task uses the HARVARDX and the MITX datasets. Sentiment analysis is applied in [8] and in [9] to the text information contained in COURSERA users posts for the prediction of student attrition rate. The COURSERA dataset is also used for *prediction of at-risk students* in [10], and in [11] the authors perform the same task using the OULAD dataset, in order to provide timely interventions to students.

OULAD data relative to student interaction with a Virtual Learning Environment in [6] are also used for *prediction of student performance*, while in [12] the authors use PORT data that describe student behavioral and lifestyle information as well as parent education level and occupation for the same task. xAPI-Edu-Data is used for *prediction of student performance* and for *student engagement analysis*: [13] considers features related to student learning behavior (e.g., raise hand in the classroom, participate to group discussions, access to online resources).

In [14], the authors focus on *student engagement analysis* and use HARVARDX and MITX datasets to evaluate student differences in completing STEM MOOCs, according to nationality and gender. The EPM dataset, which contains information about student grades and interaction with online resources, is analysed in [15] to study of the correlation between learning path and academic performance.

Our Institution Dataset (OID) is suitable for several research tasks, as it contains different categories of data (see Table I), while EDSA and ISTM focus on a specific area: the former contains information about user interaction with online resources, while the latter contains students' answers to survey questions about time management, so they are suitable for specific tasks only.

The dataset integration issue has been less covered in literature: [4] analyses the state of the art of learning analytics with respect to data integration and concludes that the vast majority of scientific papers analyse data separately without integrating datasets. Very few of them use automatic tools for data integration: they include tools developed ad hoc for a specific research project [16], Business Intelligence software [17], SQL [18] and R scripts [19].

The scope is also limited: in [18] data sources belong the same institution, in [20] data integration is managed using only two e-learning platforms, in [21] the integration is limited to EdX MOOC data and other EdX data (Coursera MOOC data are not considered). [4] also underlines that collecting and merging multiple datasets is really challenging due to the different formats and structure. The reproducibility of the experiments is usually tough, due to the lack of descriptions about the process of data integration.

UNIFORM SCHEMA

Since most of open learning datasets are characterized by different schemas, we combine the separate data sources into a unique integrated schema, namely UNIFORM. UNIFORM generalizes the information provided by different learning institutions by integrating attributes describing similar concepts.

We initially integrate only part of the available open datasets. Specifically, we integrate the following seven datasets: OID, EPM, HARVARDX, OULAD, COURSERA, PORT, xAPI-Edu-Data. The remaining (randomly selected) datasets (i.e., EDSA, ISTM, MITX, UOJ) will be exploited to test the effectiveness of the automatic alignment procedure (see Section V).

After the manual alignment of the first seven datasets, the integrated schema consists of the tables reported in Table III. Notice that, to guarantee the generality and flexibility of the newly designed database, the attribute set in the original dataset is a superset of the union of the attributes in the original tables.

The USER table describes the personal characteristics (e.g., gender, age, place of birth) including also the free time activities (e.g. alcohol week consumption). Attribute *User_Type* discriminates between student users, teacher users, or *other*.

Course information (e.g number of credits) is stored in table COURSE, while the information related to specific course instances (e.g semester, start time, duration, language) are stored in table PRESENTATION. ASSESSMENT contains data related to student assessments, which depend on the assessment type (e.g, final exams, ongoing tests, etc.). The information about the exercises assigned during an assessment procedure are recorded in table LECTURE, while tables VIDEOLECTURE and FILE respectively describe the video-lectures and the other related teaching materials. Table FORUM, THREAD, POST contain data related to forums and posts. Finally, table ACTIVITY stores the generic activities of the users (e.g. clicks, mouse movements).

IV. MANUAL ALIGNMENT

To manually align the source datasets with the newly proposed UNIFORM, for each attribute in the source dataset we look for an approximated match with UNIFORM. If a match is not found, then a new attribute is created in the extended dataset version to represent the corresponding information. For example, the new tables MODULE and MODULE-PRESENTATION are added to the integrated schema.

Table IV (that is split in two parts for the sake of readability) indicates for each dataset the percentage of matched attributes per UNIFORM table, where the self-explanatory table names indicated in the left hand-side column describe the facet of the related attributes. The results show that UNIFORM integrates most of the original data attributes, but the percentage of matching per facet is relatively low due to the high heterogeneity of the input data.

The transposed version of Table IV was given as input to a popular density-based clustering algorithm, i.e., DBSCAN [5], in order to group together similar datasets. Figure 1 plots the cluster outcome on a 2D surface after reducing it using Principal Component Analysis [5] and considering the two principal components.

TABLE I
DATASETS FEATURES.

	OID (1)	**EDSA (2)**	**EPM (3)**	**HARV (4)**	**ISTM (5)**	**MITX (6)**	**OULAD (7)**	**COURSERA (8)**	**PORT (9)**	**XAPI (10)**	**UOJ (11)**
Data types	SPD, SCD, EMD, SAD, ERA	ERA	SAD, ERA	SPD, SCD, SAD, ERA, IAD	SPD, SCD, SAD	SPD, SCD, SAD	SPD, SCD, SAD, ERA	IAD	SPD, SCD, SAD	SPD, SCD	SPD, SCD
dimensions (MB)	122.6	7.7	19.3	70.2	0.2	12.5	464.4	70.5	0.1	0.1	5.0
number of tables	7	1	5	1	2	1	7	3	2	1	13

TABLE II
DATASETS TASKS.

	OID	**EDSA**	**EPM**	**HARV**	**ISTM**	**MITX**	**OULAD**	**COURSERA**	**PORT**	**XAPI**	**UOJ**
prediction of school drop out				[7]		[7]	[6]	[8], [9]			
prediction of at-risk students	S				S		[11]	[10]	S	S	
prediction of student performance	S				S		[6]	S	[12]	[13]	
student engagement analysis		S		[14]		[14]				[22]	
learning process insight	S		[15]								S

TABLE III
UNIFORM DATASET.

table name	**attributes**
INSTITUTE	**Institute_Id**, EduLevel, EntryGradeBase, FinalGradeBase, Name, Place, Type
USER	**User_Id**, AlchoolWeekendConsuption, AlchoolWorkdayConsuption, Birth_Place, Birth_Place_Type, Birth_Time, Disability, Education_Level, FamilyRelQuality, Familysize_Count, Father_Education_Level, Father_Job, FreeTimeQuantity, Gender, GoingOut_Duration, HealtStatus, ImdBand, InternetHomeAccess, Mother_Education_Level, Mother_Job, Nationality, NurseryAttendence, ParentStatus, Residence_Place, Residence_Place_Type, RomanticStatus
USER-INSTITUTE	**Institute_Id**, **User_Id**, Cds, ChoiceReason, Entry_Grade, ExtraEduSupport, Familysupport, Final_Grade, Guardian, HToSTravel_Duration, Higher, ParentAnsweringSurvey, ParentschoolSatisfaction, Registration_Time, StudentLevel, StudiedCredits, Unregistration_Time, User_Grade, User_Type
USER-COURSE	**Course_Id**, **User_Id**, Certified, DiscussionGroups_Count, Events_Count, Failures_Count, ForumPosts_Count, InteractingChapters_Count, InteractingDays_Count, MandatoryPosts_Count, PlayVideo_Count, ViewedAnnouncements_Count, ViewedCourseContent_Count, ViewedDashboard
USER-PRESENTATION	**Presentation_Id**, **User_Id**, Absences_Count, DiscussionGroups_Count, Events_Count, Explored, ExtraCVActivitiies, ExtraPaidClasses, ForumPosts_Count, Group, InteractingChapters_Count, InteractingDays_Count, LastInterction_Time, PartecipationSessions_Array, PlayVideo_Count, Registration_Time, Unregistration_Time, ViewedAnnouncements_Count, ViewedCourseContent_Count, ViewedDashboard, WeeklyStudy_Duration
COURSE	**Course_Id**, Credits, Institute_Id, Name, Typology
PRESENTATION	**Presentation_Id**, Course_Id, Duration, End_Time, Lang, Lectures_Count, Semester, Start_Time, User_Id
ASSESSMENT	**Assessment_Id**, Course_Id, Expiration_Time, GradeBase, Institute_Id, Lecture_Id, Presentation_Id, Start_Time, Type, Weight
USER-ASSESSMENT	**Assessment_Id**, **User_Id**, Grade, IsBanked, Submission_Time
USER-EXERCISE	**Exercise_Id**, **User_Id**, Grade
EXERCISE	**Exercise_Id**, Assessment_Id, GradeBase
LECTURE	**Lecture_Id**, Lecture_Type, Order, Presentation_Id, User_Id
USER-LECTURE	**Lecture_Id**, **User_Id** Participation, RaisedHands_Count
VIDEOLECTURE	**Videolecture_Id**, Lecture_Id, Presentation_Id, Recording_Time, User_Id
FORUM	**Forum_Id**, Course_Id, Depth, File_Id, Forum_Chain, Lecture_Id, OgForum_Id, Og_Forum_Title, ParentForum_Id, ParentForum_Title, Presentation_Id, Threads_Count, Title, TitleTags_Count, Users_Count, Videolecture_Id
THREAD	**Thread_Id**, Forum_Id, Views_Count
POST	**Post_Id**, NormalizedPost_Time, Order, ParentPost_Id, Post_Time, Thread_Id, User_Id, Votes_Count, Words_Count
FILE	**File_Id**, Course_Id, Format, Lecture_Id, Presentation_Id, Title, User_Id
ACTIVITY	**Activity_Id**, ActionType, Activity_Time, Assessment_Id, End_Time, Exercise_Id, File_Id, Forum_Id, Idle_Time, Keystroke, Lecture_Id, Mouse_Click_Left, Mouse_Click_Right, Mouse_Movement, Mouse_Wheel, Mouse_Wheel_Click, Post_Id, Start_Time, Sum_Click, Thread_Id, Type, User_Id, Videolecture_Id

Most of the datasets (i.e., HARVARDX, OULAD, PORT, XAPI, MITX, UOJ) are grouped together in the same cluster. EPM and ISTM are instead included in an apart cluster grouping assessment and training information, while COURSERA (i.e., forum data) and OID (i.e., Our Institution Data) are clearly uncorrelated with the preceding ones.

V. AUTOMATIC DATASET ALIGNMENT

Given the result of a partial manual dataset integration, the goal of this step is to automate the process of enriching UNIFORM with additional information extracted from other open datasets. To this purpose, we model the problem as a multi-label classification task.

A. Problem statement

Let U be the UNIFORM integrated dataset and let D be an arbitrary original dataset to be integrated. Let $\mathcal{S}(u)$=$\{u_1, u_2, \ldots, u_n\}$ be the set of attributes in the schema of U and let $\mathcal{S}(D)$=$\{a_1, a_2, \ldots, a_m\}$ be the set of attributes in the

TABLE IV
PERCENTAGE OF MATCHED ATTRIBUTES PER UNIFORM TABLE.

	OID (1)	EDSA (2)	EPM (3)	HARV (4)	ISTM (5)	MITX (6)
LECTURE	60.0%	0.0%	40.0%	0.0%	0.0%	0.0%
PRESENTATION	55.6%	0.0%	22.2%	22.2%	0.0%	33.3%
USER-EXERCISE	0.0%	0.0%	100.0%	0.0%	100.0%	0.0%
POST	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
ASSESSMENT	50.0%	0.0%	40.0%	40.0%	30.0%	40.0%
EXERCISE	0.0%	0.0%	66.7%	0.0%	100.0%	0.0%
THREAD	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
USER-ASSESSMENT	60.0%	0.0%	60.0%	60.0%	40.0%	60.0%
USER-LECTURE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
ACTIVITY	21.7%	26.1%	65.2%	0.0%	0.0%	0.0%
COURSE	80.0%	0.0%	40.0%	40.0%	60.0%	40.0%
VIDEOLECTURE	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
USER	19.2%	3.8%	3.8%	19.2%	15.4%	19.2%
FORUMs	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
USER_INSTITUTE	31.6%	15.8%	10.5%	10.5%	15.8%	10.5%
INSTITUTE	42.9%	14.3%	14.3%	14.3%	14.3%	14.3%
USER-COURSE	14.3%	0.0%	14.3%	21.4%	14.3%	14.3%
FILE	42.9%	28.6%	0.0%	0.0%	0.0%	0.0%
USER-PRESENTATION	19.0%	0.0%	14.3%	52.4%	9.5%	52.4%

	OULAD (7)	COURSERA (8)	PORT (9)	XAPI (10)	UOJ (11)
LECTURE	0.0%	0.0%	0.0%	40.0%	0.0%
PRESENTATION	33.3%	66.7%	22.2%	33.3%	55.6%
USER-EXERCISE	0.0%	0.0%	0.0%	0.0%	0.0%
POST	0.0%	100.0%	0.0%	0.0%	0.0%
ASSESSMENT	70.0%	0.0%	40.0%	0.0%	60.0%
EXERCISE	0.0%	0.0%	0.0%	0.0%	0.0%
THREAD	0.0%	100.0%	0.0%	0.0%	0.0%
USER-ASSESSMENT	100.0%	0.0%	60.0%	0.0%	60.0%
USER-LECTURE	0.0%	0.0%	0.0%	75.0%	0.0%
ACTIVITY	21.7%	0.0%	0.0%	0.0%	0.0%
COURSE	40.0%	80.0%	40.0%	40.0%	60.0%
VIDEOLECTURE	0.0%	0.0%	0.0%	0.0%	0.0%
USER	26.9%	3.8%	73.1%	19.2%	34.6%
FORUMs	0.0%	75.0%	0.0%	0.0%	0.0%
USER_INSTITUTE	21.1%	15.8%	42.1%	36.8%	10.5%
INSTITUTE	14.3%	14.3%	28.6%	14.3%	28.6%
USER-COURSE	21.4%	21.4%	21.4%	28.6%	14.3%
FILE	57.1%	0.0%	0.0%	0.0%	0.0%
USER-PRESENTATION	19.0%	14.3%	28.6%	33.3%	9.5%

schema of D. The automatic alignment task entails inferring the following mapping function $\mathcal{F}$:

$$\mathcal{F} : \mathcal{S}(D) \rightarrow \mathcal{S}(U)$$

To formalize the task as a classification problem, for each attribute pair (a_k, u_j) [$1 \leq k \leq n, 1 \leq j \leq m$] we estimate the probability

$$p(\mathcal{F}(a_k) = u_j)$$

In the single-label classification task, the output of the alignment procedure for attribute a_i is given by

$$arg_j \ max \ p(\mathcal{F}(a_k) = u_j)$$

If an attribute of the original dataset can be assigned to multiple attributes of the integrated schema (i.e., the multi-label classification task), then all the target attributes satisfying a minimum (user-specified) probability threshold are assigned.

B. Classifier training

To tackle the problem stated above, we trained a classification model on a training relational dataset consisting of a distinct record for each pair of attributes (a_k, u_j). A record is labeled as one if the two attributes are manually aligned, as zero otherwise.

To accurately predict attribute alignments, each record of the training dataset is characterized by a set of features denoting the cosine similarities between the following distances separately evaluated on textual attribute descriptions and attribute

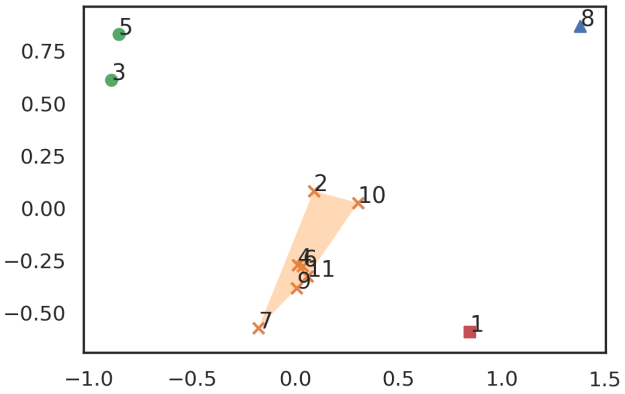


Fig. 1. Dataset clusters. DBSCAN settings: *eps*=1.2. *min_samples*=1.

labels: *TextRank* distance [23], BERT embedding vectors' similarity [24], Ensemble NERD [25], Fuzzywuzzy distance [11] with Token Set Ratio.

Furthermore, the total number of words in common between the two Bag-Of-Word textual representations and the total number of common Wikipedia entities recognized in the text are included as input features as well.

We train two different classification models: MultiLayer Perceptron and Random Forest classifier [5]. To tailor the algorithm configuration settings on the analyzed data distribution, we perform a grid search and we select the best configurations To train the classification models we use the Keras[12] and the algorithm implementations available in the Scikit Learn library [26].

C. Classifier evaluation

To evaluate classifier performance in predicting attribute alignment we carried out a 70%-30% hold-out validation with oversampling of class 1 (i.e., the minority class) to face the issue of imbalanced classes. The evaluation was conducted on the following (manually aligned) datasets: OID, EPM, HARVARDX, OULAD, COURSERA, PORT, XAPI.

Table V reports the results of the classifier evaluation in terms of (i) classifier accuracy (i.e., percentage of correctly classified records[13]), (ii) precision of class 1 (i.e., the number of records correctly classified as 1 over the total number of records classified as 1), (iii) recall of class 1 (i.e., the number of records correctly classified as 1 over the total number of records), and (iv) F1-score of class 1 (i.e., the harmonic mean of recall and precision).

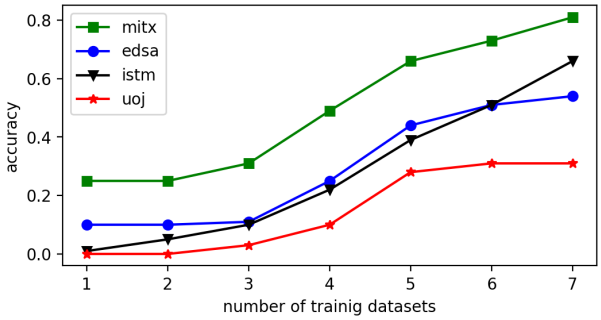
The Neural Network model is slightly more accurate than Random Forest, but the precision is fairly low. Hence, the performance of Random Forest is globally superior in terms of F1-score.

D. Automatic alignment of new open datasets

To evaluate the ability of the classifiers in automatically aligning new datasets, we trained the classifier on the seven aligned datasets and we tested them separtely on each of the four datasets excluded from the previous evaluation (i.e., MIXT, EDSA, ISTM, UOJ). We also simulated the enrichment of the UNIFORM dataset by training the classification models on a subset of aligned datasets. To avoid bias due to random dataset selection, we run the experiment multiple times by shuffling the considered training datasets and we averaged the results.

Figure 2 shows the accuracy values achieved by the Random Forest classifier on each test dataset by varying the number of aligned datasets in the training set. As expected, the accuracy increases while enriching the classification model with newly labeled data. An 80% accuracy was reached by using all the seven aligned datasets in the training set. The detailed results are summarized in Table VI.

Fig. 2. Classifier accuracy (%). Random Forest Classifier.



VI. Case Study

We present a real case study to show, in a practical example, that coping with the integrated dataset version has allowed us to achieve a better performance than that obtained using the original data.

The considered case study is focused on predicting the upcoming exam outcomes of a set of university-level students. We extract student and exam information from two examples of data sources, i.e., Our Institution Dataset (OID) and from HARVARDX. We have chosen the aforesaid datasets because they are approximately consistent in terms of teaching methodologies and covered subjects.

For evaluation purposes, we first split both the original datasets in training and test samples thus producing the following partitions: *OID-Train* (1078 records), *OID-Test* (270 records), *HARVARDX-Train* (1769 records), *HARVARDX-Test* (443 records). Next, we apply the semi-automatic alignment procedure described in Section V to generate a unified dataset version, i.e., UNIFORM (2847 samples), integrating the knowledge provided by OID and HARVARDX in a common data model. Then, we train a binary classifier on part of the integrated dataset (*UNIFORM-Train*) to predict, for a subset of students, the exam outcomes (0=fail, 1=pass). The achieved predictions are compared with those produced by the classifiers trained separately on each original training dataset (*OID-Train* and *HARVARDX-Train*). The idea is to

[11] https://github.com/seatgeek/fuzzywuzzy
[12] https://keras.io/
[13] The frequency counts are weighted by the relative class frequencies.

	OID		EPM		HARV		OULAD		COURSERA		PORT		XAPI	
	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF
Accuracy	0.90	0.61	0.76	0.91	1.00	0.60	0.96	0.42	0.97	0.73	0.86	0.94	0.87	0.53
F1-score(1)	0.07	**0.74**	0.35	**0.94**	0.08	**0.72**	0.03	**0.58**	0.06	**0.81**	0.35	**0.97**	0.12	**0.70**
Precision(1)	0.04	0.94	0.18	0.96	0.04	0.90	0.02	0.92	0.03	0.91	0.22	1.00	0.06	1.00
Recall(1)	0.90	0.61	0.76	0.91	1.0	0.60	0.96	0.42	0.97	0.72	0.86	0.94	0.87	0.53

TABLE V
CLASSIFIER EVALUATION SCORES. HOLD-OUT VALIDATION ON THE MANUALLY ALIGNED DATASETS.

	EDSA		ISTM		MITX		UOJ	
	NN	RF	NN	RF	NN	RF	NN	RF
Accuracy	0.58	0.54	0.68	0.66	0.84	0.81	0.30	0.31
F1-score(1)	0.02	**0.60**	0.02	**0.72**	0.16	**0.85**	0.03	**0.42**
Precision(1)	0.01	0.69	0.01	0.8	0.09	0.91	0.02	0.65
Recall(1)	0.57	0.53	0.68	0.65	0.82	0.80	0.29	0.31

TABLE VI
CLASSIFICATION EVALUATION SCORES FOR THE AUTOMATIC ALIGNMENT OF NEW OPEN DATASETS.

verify whether the data integration phase has improved the quality of the training data so that the prediction models would yield better performance.

We test also many variants of integrated models, including either part of the original data (e.g., integrate in UNIFORM only OID data *UNIFORM(OID)-Train*) or the whole set, i.e., *UNIFORM(HARVARDX+OID)-Train*.

We run the tests using the Scikit Learn implementation [26] of the established Random Forest Classifier [5]. Classifier parameters have been set up via grid search. The best configuration setting is reported in Table VII. Classifier evaluation has been performed in terms of the following metrics: (i) accuracy, (ii) precision of class 1, (iii) recall of class 1, and (iv) F1-score of class 1.

Table VIII summarizes the achieved results. The first line reports the prediction outcomes achieved using the original OID data for both training and test, while the second line indicates the result of a similar experiment carried out after integrating OID into UNIFORM. As expected, integrating the same data in the UNIFORM schema does not affect prediction performance.

The third and fourth lines respectively indicate the classifier performance achieved by integrating HARVARDX data exclusively in the training phase and in both training and test sets. The results show that the integration process significantly enhances classifier performance, because the enriched model is now able to capture new predictive patterns, which would remain undisclosed in the OID training data. The improvement in terms of F1-score is around 7% while testing exclusively OID test samples, while is approximately 32% while testing a stratified sample of OID and HARVARDX test records.

The aforesaid preliminary results show that integrating additional, consistent data from other sources could be particularly beneficial in real cases in which the problem of exam outcome prediction is particularly challenging, (e.g., when the number of training samples is relatively low).

TABLE VII
REAL CASE STUDY. RANDOM FOREST CLASSIFIER. CHOSEN PARAMETER SETTINGS.

Parameter	Value
bootstrap	True
ccp_alpha	0.0
class_weight	None
criterion	gini
max_depth	10
max_features	3
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0.0
min_impurity_split	None
min_samples_leaf	5
min_samples_split	8
min_weight_fraction_leaf	0.0
n_estimators	100
n_jobs	None
oob_score	False
random_state	None
verbose	0
warm_start	False

VII. CONCLUSIONS AND FUTURE WORK

The paper proposes a new data model integrating various open learning datasets. The integrated schema can be semi-automatically enriched with new open datasets as soon as they become available. The further integrations are machine learning supported. Specifically, a classification model is trained to predict the most likely feature dependencies among original and integrated data.

The performance of the automated integration phase are promising: by training the Random Forest Classifier, we have reached a F1-score on the minority class equal to or above 70% on 6 out of 7 datasets. Furthermore, we made a preliminary attempt to use the integrated data to solve a real Learning Analytics problem, i.e., exam outcome prediction. The results show that training the predictive models on the integrated dataset version could be particularly convenient, especially when the original dataset has limited cardinality.

We plan to shortly release the open dataset, the machine learning supported integration system, as well as the related documentation.

The current project leaves room for further extensions. For example, we plan to include multimodal data in the integrated data model (e.g., video-lectures, slides). This opens new research challenges regarding to way to process and automatically integrate data sources in different formats and acquired from different medias. Furthermore, we plan also to

TABLE VIII
CLASSIFICATION SCORES FOR STUDENTS'S OUTCOME PREDICTION.

Training Dataset	Test Dataset	Accuracy	F1-score(class 1)	Precision (class 1)	Recall (class 1)
OID-Train	*OID-Test*	0.60	0.71	0.56	0.98
UNIFORM(OID)-Train	*UNIFORM(OID)-Test*	0.59	0.70	0.55	0.98
UNIFORM(OID+HARVARDX)-Train	*UNIFORM(OID)-Test*	**0.64**	**0.76**	**0.62**	0.97
UNIFORM(OID+HARVARDX)-Train	*UNIFORM(OID+HARVARDX)-Test*	**0.93**	**0.94**	**0.91**	0.96

integrate learning data written in different languages and to exploit Deep Natural Language Processing techniques to align multilingual data models.

REFERENCES

[1] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, ser. LAK '12. New York, NY, USA: ACM, 2012, pp. 252–254.

[2] M. Oi, M. Yamada, F. Okubo, A. Shimada, and H. Ogata, "Reproducibility of findings from educational big data: a preliminary study," 03 2017, pp. 536–537.

[3] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Transactions on Learning Technologies*, vol. 10, no. 1, pp. 17–29, Jan 2017.

[4] J. Samuelsen, W. Chen, and B. Wasson, "Integrating multiple data sources for learning analytics—review of literature," *Research and Practice in Technology Enhanced Learning*, vol. 14, no. 1, p. 11, 2019.

[5] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005.

[6] N. I. Jha, I. Ghergulescu, and A.-N. Moldovan, "Oulad mooc dropout and result prediction using ensemble, deep learning and regression techniques," in *CSEDU*, 2019.

[7] J. K. Tang, H. Xie, and T.-L. Wong, "A big data framework for early identification of dropout students in MOOC," in *International Conference on Technology in Education*. Springer, 2015, pp. 127–132.

[8] D. S. Chaplot, E. Rhim, and J. Kim, "Predicting student attrition in MOOCs using sentiment analysis and neural networks." in *AIED Workshops*, vol. 53, 2015, pp. 54–57.

[9] S. Shatnawi, M. M. Gaber, and M. Cocea, "Automatic content related feedback for MOOCs based on course domain ontology," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2014, pp. 27–35.

[10] J. He, J. Bailey, B. I. Rubinstein, and R. Zhang, "Identifying at-risk students in massive open online courses," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[11] M. Hlosta, Z. Zdrahal, and J. Zendulka, "Ouroboros: early identification of at-risk students without models based on legacy data," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 2017, pp. 6–15.

[12] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[13] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using x-api for improving student's performance," in *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*. IEEE, 2015, pp. 1–5.

[14] S. Jiang, K. Schenke, J. S. Eccles, D. Xu, and M. Warschauer, "Cross-national comparison of gender differences in the enrollment in and completion of science, technology, engineering, and mathematics massive open online courses," *PloS one*, vol. 13, no. 9, p. e0202463, 2018.

[15] M. Vahdat, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, "A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator," in *Design for teaching and learning in a networked world*. Springer, 2015, pp. 352–366.

[16] D. Di Mitri, M. Scheffel, H. Drachsler, D. Börner, S. Ternier, and M. Specht, "Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, ser. LAK '17. New York, NY, USA: ACM, 2017, pp. 188–197. [Online]. Available: http://doi.acm.org/10.1145/3027385.3027447

[17] S. M. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative," *Journal of Learning Analytics*, vol. 1, no. 1, pp. 6–47, 2014.

[18] C. E. López Guarín, E. L. Guzmán, and F. A. González, "A model to predict low academic performance at a specific enrollment using data mining," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 10, no. 3, pp. 119–125, Aug 2015.

[19] M. N. Giannakos, K. Sharma, I. O. Pappas, V. Kostakos, and E. Velloso, "Multimodal data as a means to understand the learning experience," *International Journal of Information Management*, vol. 48, pp. 108–119, 2019.

[20] K. Mangaroska, B. Vesin, and M. Giannakos, "Cross-platform analytics: A step towards personalization and adaptation in education," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, ser. LAK19. New York, NY, USA: ACM, 2019, pp. 71–75. [Online]. Available: http://doi.acm.org/10.1145/3303772.3303825

[21] Z. A. Pardos and K. Kao, "moocRP: An open-source analytics platform," in *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, ser. L@S '15. New York, NY, USA: ACM, 2015, pp. 103–110. [Online]. Available: http://doi.acm.org/10.1145/2724660.2724683

[22] I. F. Siddiqui, Q. A. Arain *et al.*, "Analyzing students'academic performance through educational data mining," *3C Tecnologia*, 2019.

[23] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *CoRR*, vol. abs/1602.03606, 2016. [Online]. Available: http://arxiv.org/abs/1602.03606

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[25] L. Canale, P. Lisena, and R. Troncy, "A novel ensemble method for named entity recognition and disambiguation based on neural network," in *ISWC 2018, International forum for the Semantic Web and Linked Data Community, 8-12 october 2018, Monterey, CA, USA / Also published in LNCS, Vol.11136*, Monterey, UNITED STATES, 10 2018. [Online]. Available: http://www.eurecom.fr/publication/5564

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.