Received: 28 April 2020 | Revised: 12 November 2020 | Accepted: 12 January 2021

DOI: 10.1002/widm.1403

OVERVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

Review on publicly available datasets for educational data mining

Marian Cristian Mihaescu | **Paul Stefan Popescu**

Department of Computer Science and Information Technologies, University of Craiova, Craiova, Romania

Correspondence
Marian Cristian Mihaescu, Department of Computer Science and Information Technologies, University of Craiova, Craiova, Romania.
Email: cmihaescu@software.ucv.ro

Edited by Frank Hoffmann, Associate Editor, and Witold Pedrycz, Editor-in-Chief

Abstract

The availability of a dataset represents a critical component in educational data mining (EDM) pipelines. Once the dataset is at hand, the next steps within the research methodology regard proper research issue formulation, data analysis pipeline design and implementation and, finally, presentation of validation results. As the EDM research area is continuously growing due to the increasing number of available tools and technologies, one of the critical issues that constitute a bottleneck regards a properly documented review on publicly available datasets. This paper aims to present a succinct, yet informative, description of the most used publicly available data sources along with their associated EDM tasks, used algorithms, experimental results and main findings. We have found that there are three types of data sources: well-known data sources, datasets used in EDM competitions and standalone EDM datasets. We conclude that the success of the future of EDM data sources will rely on their ability to manage proposed approaches and their experimental results as a dashboard of benchmarked runs. Under these circumstances, the reproducibility of data analysis pipelines and benchmarking of proposed algorithms becomes at hand for the research community such that progress in the EDM domain may be much more easily acquired. The most crucial outcome regards the possibility of continuously improving existing data analysis pipelines by tackling EDM tasks that rely on publicly available datasets and benchmarking data analysis pipelines that use open-source implementations.

This article is categorized under:

Application Areas > Education and Learning

Fundamental Concepts of Data and Knowledge > Big Data Mining

KEYWORDS

educational data mining, public data-sets

1 | INTRODUCTION

The Educational Data Mining (EDM) as research domain was well defined in Baker et al. (2010)as "the area of scientific inquiry centered around the development of methods for making discoveries within the unique types of data that come from educational settings, and using those methods to better understand students and the settings which they learn in."

All the progress in the area has been recently captured by Labarthe et al. (2018) by analyzing the relationship between EDM, Learning Analytics (LA) and AI for education (AIED) in terms of conferences, journals, professional societies and the community of reviewers. Therefore, throughout the paper, we shall use only the EDM acronym for both LA and AIED.

Among the first EDM review papers we mention Romero and Ventura (2007) and Baker and Yacef (2009) who has received the *Test of Time Award* at EDM2020 conference. Later, in Romero and Ventura (2010), authors describe EDM stakeholders and objectives along with types of data that may be mined. At the same time, PSLC DataShop Koedinger et al. (2010) was the main publicly available dataset making thus the EDM a well-established research area.

Later, Merceron (2015) points out the new LAK conference started in 2011 and appearance of "Big Data in Education" MOOC course by Ryan Baker offered on Coursera in 2013 and on EdX in 2015 and as MOOT[1] Baker and Wang (2015).

Of particular importance in the area of general EDM review papers they are the works of Romero and Ventura (2013), Romero and Ventura (2017) and Romero and Ventura (2020). The main event spotted in 2017 paper is appearance and well establishment of MOOCs: Coursera, edX or Udacity. Although MOOC challenges remain quite the same as before, they suffer some modifications brought by the new context. Thus, analyzing students' interactions becomes analyzing usage and engagement, FORUM data or in-video interaction. The paper from 2020 shows an increase to 9 conferences, 18 books—from which 14 are from last 5 years—a list of 11 top related journals and a list of top most cited papers. It also presents a link and a short description for a list of 9 tools and 13 publicly available datasets. Finally, a list of 16 most popular methods and 23 current applications is presented. Although this paper presents in short a list of 13 publicly available datasets, this review aims in getting into more details about 26 datasets in an attempt to present them in a way that is more informative for researchers that would wish to dive into the area of EDM quickly.

In the area of specific EDM review papers Slater et al. (2017) present software tools that are widely used, accessible, and compelling when dealing with research tasks. A review of data mining processes on how students solve programming problems has been addressed by Ihantola et al. (2015), which points out the lack of publicly available datasets as the main limitation for performing replication studies. In this area, we mention the latest development by Alon et al. (2019), which present a neural-based model for building code embedding's. We state that this progress brings the area of NLP close to computer programming with several exciting applications like automatic code review or API discovery. Further, a review of visual analytics of educational data has been published by Vieira et al. (2018) and the area of text mining in education has been reviewed by Ferreira-Mello et al. (2019).

Still, systems and progress must be evaluated, and this may be achieved only through publicly available datasets and reproducible research. In this matter, we observe that there is no systematic review on publicly available EDM datasets, as the only mention reduces to a table in Romero and Ventura (2020) which provides only the link to the publicly available dataset and a minimal description.

In this context, we hypothesize that the knowledge gap is represented by the *Data* ingredient. This paper investigates to what extent the EDM community has access to well documented and publicly available datasets. Furthermore, we investigate the structure of that data and tasks that were tackled. Aspects like data formats, data types, number of samples, number of features and others are investigated.

The intended audience of this paper regards new researchers that want to delve into the EDM area or experienced researchers that would progress much faster with all the reviewed information at hand. The researchers that do not have an e-Learning system running would be the primary beneficiaries of the review since they may directly start working on already defined EDM tasks given that the dataset is publicly available and well documented. Various solutions are implemented and well documented proposing alternative and hopefully, better solutions that many be compared with already existing ones. This approach represents an essential ingredient for the progress of EDM domain. On the other hand, EDM research groups which have e-Learning systems running would also benefit upon the review, as they may further decide what task should be further tackled and what data is necessary to be logged.

The approach consists of a systematic review of the literature, with the first step consisting of reviewing the general EDM review papers. Then, articles that present EDM review on specific aspects are also considered. Section 2 presents the related works on both general and specific EDM review papers. Section 3 presents the methodology of the review process, such that findings and further improvements may be achieved. Furthermore, Section 4 offers the primary EDM public datasets, repositories and data sources by providing information such as the number of citations, structure of the dataset, tasks with approaches, results and findings. Finally, the conclusions and future works are presented in Section 5.

2 | METHODOLOGY OF REVIEW PROCESS

We designed a custom systematic sources (i.e., research papers and websites) review method where the following questions were asked upon analyzing two types of data sources: (a) That describe publicly available datasets and (b) That use publicly available datasets in their experimental results. Data sources that only cite papers describing EDM datasets without using the dataset for running the experiments were not taken into consideration.

Q1. Which are the primary publicly available EDM datasets and which are the categories to which they belong? The deliverable should be a comprehensive, structured list of EDM datasets with minimal details for practitioners and researchers interested in this area.

Q2. Which are the primary data types that are available? This analysis would be very beneficial for data analysts as a first step in tackling particular problems.

Q3. Which are the main types of tasks that have been tackled? This would include the range of possibilities that researchers or practitioners may address.

Q4. Which are the requirements for successful future datasets? This would provide a roadmap for improving the status of already existing EDM datasets and especially for groups that take into account making datasets publicly available. These aspects may refer to datasets in general, not specifically to EDM related ones.

2.1 | Strategies for collection of bibliographic materials

We performed electronic searches in Google Scholar (GS)[2] and Google Dataset Search (GDS)[3] mainly using the query "educational data mining datasets" along with appended query words like "review papers," "public," "UCI ML, "Mendelay" and other specific words which narrowed down the search. The goal of this first step was to create a pool of publicly available datasets which was further split into the categories presented in Section 4. Once the categories and datasets were established, we started to describe them by reviewing bibliographic materials.

The task of collecting bibliographic materials reduced to finding research papers or Websites that either describe a public EDM dataset or use a public EDM dataset in their experiments. Therefore, we have used mainly GS along with DS as primary tools for finding needed research papers and information about datasets and their usages, as Gusenbauer (2019) states that "Google Scholar, with 389 million records, is currently the most comprehensive academic search engine". We have found a total number of 41 datasets—most of them from the last 10 years—that were described in six categories.

Among the most well-known academic search engines and bibliographic databases that were compared with GS, we mention CiteSeerX, EbscoHost, Microsoft Academic, ProQuest, Scopus and Web of Science. Since our task narrowed down to finding specific papers (i.e., that describe or use public EDM datasets) using exhaustively all available search engines and bibliographic datasets would have become very time consuming and potentially with the very same results as the ones obtained using GS and DS. Since the task of the review was to find the most representative and used EDM datasets and to outline their usage, we have not gone into particular detailed usage scenarios and results. Such detailed information may be further obtained by using search engines like Scopus, Elsevier ScienceDirect, WOK, IEEE Xplore Digital Library, ACM Digital Library, Wiley Online –Library, or Springer.

One limitation of this approach regards the difficulty in assessing the importance of the number of citations for a paper that describes a dataset. This limitation inherently occurs because an article that cites a paper describing a dataset may be in one of the following situations: (a) The citing paper uses the dataset in experiments, maybe along with other datasets or (b) The citing paper cites the article describing the dataset, without actually using the dataset in experimental results. As the number of citations is sometimes large (i.e., hundreds) and not all papers may be available, this becomes a somewhat difficult task to investigate in detail all the documents that cite the article describing the dataset. As an approach, reduced our investigations to the first page of citations returned by GS as these may be the most informative papers regarding the actual usage of the dataset. A particular situation occurs when the dataset has no associated scientific paper describing the dataset, and this will be discussed later in the appropriate section. Collecting the bibliographic materials reduced only to finding good research papers and assessing their impact by the number of citations, quality of publication (i.e., conference paper, journal, etc.) and publication year without taking into consideration other detailed scientometrics about authors, publication or citations.

The investigated research papers fell in following categories: (a) papers that perform review EDM in general or regarding a specific topic (i.e., tools, visual analytics, learning experience or text mining), (b) papers that describe public

datasets or EDM workflows on those available datasets, (c) articles that actually use in their data analysis pipeline an EDM dataset and tackle an EDM task and (d) papers that just cite a dataset (i.e., described in an article, used in a competition or standalone) without using it. Most of the items were published in the last 5 years, although firstly published EDM review papers are also presented.

3 | EDM PUBLIC DATASETS, REPOSITORIES AND DATA SOURCES

3.1 | UCI ML repository

On *UCI ML repository*[4] there are various educational datasets configured and ready to be used for experiments. Although these datasets are available from a long time and are among the most referenced datasets in the EDM community, they differ a lot in structure, logged features and usefulness in terms of citations, practical usages and tackled tasks.

University Data Set[5] is the first dataset with 285 instances and 17 attributes that has been donated in 1988. Unfortunately, there is no paper to describe the dataset and therefore we do not know citation or usage of the dataset, and this fact makes it of little interest.

Teaching Assistant Evaluation Data Set[6] is another early dataset described in Loh and Shih (1997) which gathered 1,340 citations, not mainly for the quality and extensive usage of the dataset but for the discussion on split selection methods. The dataset itself is relatively small, with 151 instances described by 5 attributes.

Student Performance Dataset[7] is described in Cortez and Silva (2008), has 394 citations and consists of 649 instances described by 30 attributes from which three can be used as target class. The dataset is divided into two files which can be merged because they have the same structure: one from math classes and one from Portuguese courses. Although there are three features which can be used as the target class, two of them can be used in the process of classification are positively correlated with the last. The attributes are mainly demographic and social, but some of them are school-related, and their usage is limited to predicting the student's final grade at the end of the semester or year. One thing that needs to be mentioned is that the final grade is from 0 to 20, resulting in a significant number of class values.

This dataset has also been uploaded on Kaggle[8] where 305 publicly available kernels perform exploratory data analysis. Unfortunately, there is not defined any task with specific validation metric such that there is no leaderboard publicly available.

User Knowledge Modeling Data Set[9] is described in Kahraman et al. (2013), has 96 citations and consists of 403 instances defined by five attributes from which there is only one class attribute discretized into four values. In the case of this dataset, the details are highly related to learning activities with no demographic or social information about the users. The dataset is feasible mainly for classification, clustering, and it can be used to predict the knowledge level of learners. In unsupervised context, the students are grouped in clusters representing different knowledge levels. Otherwise, the dataset has been used in the experimental results of newly proposed algorithms for testing new predictive validation metrics.

Educational Process Mining Dataset (EPM)[10] is described in Vahdat et al. (2015), has 42 citations and consists of 230,318 samples described by 13 attributes computed based on a group of 115 students. The dataset has been created by logging performed activities of the students while using an educational simulator. Based on this structure, the dataset has been used in classification, regression or clustering for predicting learning difficulties, analyzing structured learning behavior, self-organizing map clustering or discovering student behavior patterns.

Open University Learning Analytics dataset[11] is another reasonably well described standalone dataset in Kuzilek et al. (2017). Among the most tackled tasks within the 60 citations, we mention early identification of at-risk students, student engagement predictions or the role of demographics in on-line learning. The prediction tasks were addressed by classical ML algorithms (i.e., DT, J48, XGBoost, CART, SVM or Random Forest) and were validated by classical metrics. The greatest limitation is that the dataset consists of seven *.csv* files which have the design of a databases schema. Thus, using this dataset may require intensive preprocessing as feature values should be computed for items (i.e., student, course, assessment) that may be taken into consideration.

Student Academics Performance Data Set[12] is described in Hussain et al. (2018), has 43 citations and consists 300 instances described by 24 nominal attributes which are mainly demographic. The main tackled tasks regard prediction of student academic performance in a classification context.

3.2 | Mendeley Data Repository

Another important data source is Mendeley Data Repository,[13] which has been described in detail as a platform for research data management in Bhoi (2018).

Embeddings and topic vectors for MOOC lectures dataset[14] has been described in Kastrati et al. (2020) and consists of word embeddings and document topic distribution vectors generated from 12,032 MOOCs video lecture transcripts from 200 courses collected from Coursera learning platform. The most significant shortcoming of this dataset is that it does not include the transcripts themselves. The dataset is aimed for NLP tasks and consists of ten *.csv* files, six with data about topic vectors and four with data about word embeddings.

Data for: Effectiveness of flip teaching on engineering students' performance in the physics lab[15] has been described in Gómez-Tejedor et al. (2020) and consists of a sample of 1,233 students enrolled from 2013 to 2017 who completed the subjects of Physics and Electricity. The dataset contains laboratory grades and final grades both for traditional teaching methodology and flip teaching that is suitable for classification and regression tasks.

Dataset for factors affecting teachers' Burnout[16] has been described in Prasojo et al. (2020) and contains data about a sample of 876 teachers across three Indonesian provinces that completed a printed form of a questionnaire. The questionnaire, responses and factor analysis results are available, along with a final validated model and relationship testing.

Data of academic performance evolution for engineering students[17] has been described in Delahoz-Dominguez et al. (2020) and consists of data containing academic, social, economic information for 12,411 students described by 44 attributes in a *.csv* format. The dataset may be used for prediction, classification and evaluation models of academic and social variables.

KEEL dataset repository[18] has been shortly described in Hou et al. (2019) with a focus on KEEL Java software tool that can be used for a large number of different knowledge data discovery tasks. The archive dataset consists of many datasets that are also available with the standalone KEEL[19] project homepage.

Outcome based education attainment calculation (OBE dataset)[20] has no paper describing the dataset. Still, the dataset is fairly large consisting of 34,650,000 data entries for 20 programs, 1,872 courses, 4,800 course outcomes, 3,850 students and 9,800 assessments organized in 21 files of *.docx*, *.pdf* and *.csv* types. Although a detailed description of the data is not available, the contributors present the steps needed to be reproduced to properly investigate and understand the dataset.

As the datasets are relatively new and the current number of downloads is relatively small, we conclude that these datasets are not yet used EDM processes, but worth to be mentioned as references for researchers. Still, the fact that some datasets do not have an associated research paper that describes them will make them difficult to be spotted and used in EDM processes.

3.3 | Harvard Dataverse

Harvard Dataverse[21] is a repository for research data that currently contains 98,873 datasets organized on 13 subjects, such as Computer and Information Science, Engineering, Arts and Humanities, and so on. One particular useful key aspect is that Harvard Dataverse has implemented functionality data viewing and exploration, along with other valuable features.

HarvardX Person-Course Academic Year 2013 De-Identified dataset[22] has been described in Ho et al. (2014), a paper with 352 citations that presents detailed statistics about online courses that produced the data. The 338,223 logged instances are defined by 20 attributes and were used for understanding the progression of users, Examining access and usage patterns or predicting MOOC performance with week 1 behavior.

Canvas Network Person-Course[23] is a publicly available dataset from 2016 and has 2,766 downloads with not published paper describing the dataset. The only known usage is at LAVA hackathon, which is presented in a later subsection. The dataset consists of 325,000 aggregate records defined by 25 attributes. The dataset has plenty of missing values which leads to a necessity for preprocessing, and there is no explicit class attribute which makes it more feasible for clustering purposes. Each record represents one individual's activity in one of 238 courses.

Massively Open Online Course for Educators (MOOC-Ed) dataset[24] has been described in Kellogg and Edelmann (2015), a paper with 14 citations. The dataset is mainly designed to be used for Social Network Analysis, MOOCs courses dropout rate or exploring self-regulated learner profiles in MOOCs.

CAMEO Dataset: Detection and Prevention of" Multiple Account" Cheating in Massively Open Online Courses[25] has been described in Northcutt et al. (2016), a paper with 69 citations that addresses the problem of cheating as a big issue in MOOCs. Unfortunately, we could not find other papers that actually use the dataset for data analysis, although the citations tackle the very same problem. The dataset consists of four restricted files that may be accessed upon request, consisting of course listings, description and actual data in *.csv* format.

Nursing Student Data[26] as well as *Situated Academic Writing Self-Efficacy Scale Validation*[27] are two new datasets with the description in preparation as a unpublished doctoral dissertation of Mitchell Kim from Red River College. The first dataset has 255 observations defined by 84 variables, while the second one has 807 observations defined by 29 variables.

Early Reading and Writing Assessment in Preschool Using Video Game Learning Analytics[28] is also a very recent dataset authored by Amorim Americo, with no published paper describing the dataset in detail. The dataset contains data from 331 observations (i.e., students) represented by 25 variables, among which 20 regard phonological awareness, early reading and writing games, as well as their scores in a standardized word reading and word writing assessment.

3.4 | DataSchop@CMU data source

DataShop@CMU[29] has been described in Koedinger et al. (2010) and consists of a system that gathers fine-grained (i.e., very detailed), longitudinal (i.e., per semester or academic year) and extensive 40 datasets collected from ongoing courses or external courses. The data sets may be imported or exported as XML or tab-delimited text file format, and are available for download along with available applications/tools that may be used to support exploratory analyses. Taking into account the available built-in functionalities, such as visualizing student performance over time, hint use, latent knowledge, response times, various visualizations and many more we are in front of an integrated EDM environment that is built as a Web application which is available for free but is not open-source.

As the DataSchop@CMU gathered data from 6 courses in 2010, it continuously added and updated dataset versions and analysis/visualization tools. Thus, DataSchop@CMU gathers datasets about middle school math from ASSISTments, OLI (Online Learning Initiative) or other educational software (i.e., Andes, Cognitive Tutor, REAP, etc.). In general, about 3 datasets per year were regularly added or updated, some of them being used in EDM related data competitions or challenges. The dataset and competitions are further discussed in a distinct section.

Of particular importance, DataSchop@CMU has custom XML data format for data representation and tools for import/export data. Further, analysis and visualization tools such as dataset info, performance profiler and learning curve are available for exploratory data analysis. This approach makes DataSchop@CMU a very friendly data analysis ecosystem, with all ingredients (i.e., data and tools) available and integrated into one place. Still, a shortcoming of this approach regards the difficulty of building and using data models in other custom-developed applications.

With almost 400 citations, the DataSchop@CMU has been intensively used for tasks like sequence/temporal classification mining, automatic description of knowledge components, estimating the minimum number of opportunities needed for all students to achieve predicted mastery, predicting students' problem-solving performance. The employed algorithms are Naive Bayes classifier, hidden Markov prediction model, Bayesian Knowledge Tracing (BKT) prediction model, SVM, to mention only a few tasks and associated methods described in first papers that cite Koedinger et al. (2010) in GS. A useful feature that DataSchop@CMU exhibits are that it manages a list of topics of interest, for which related datasets and papers are presented.

3.5 | Datasets from EDM Competitions or within crowd-sourced platforms

Particular types of datasets are the ones made publicly available for EDM competitions or within the crowd-sourced platform such as Kaggle. This section presents EDM competitions that took place at top DM conferences under the supervision of prestigious professional organizations, competitions that run within a crowd-sourced platform or as hackathons. The most significant benefit of this approach is that it provides on the spot strong evidence about the quality of the proposed model/kernel and a ranking against other competitors.

KDD Cup 2010 Educational Data Mining Challenge[30] is the first EDM competition. Although it started 10 years ago Stamper et al. (2010) solutions are continuously added. The organizers offer two types of datasets: for development and challenges similar to Kaggle. The development datasets consist of two *Algebra I* datasets (2005–2006 and 2006–2007)

and one for *Bridge to Algebra* (2006–2007); these datasets consists of 575, 1840 and 1,146 students which are a quite large number. For the challenge, there are two datasets: *Algebra I* and *Bridge to Algebra*. These datasets have 3,310, and 6,043 students and each of these students for any of the datasets have 20 records. The evaluation of the competition was made using RMSE metric and based on the leaderboard.

RecSysTEL Datatel Challenge 2010[31] is a competition run within Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010) that was organized jointly by fourth ACM Conference on Recommender Systems (RecSys 2010). Unfortunately, the used dataset is not available, and there is no public leaderboard.

What Do You Know?[32] is a Kaggle competition launched in 2011. The task is to predict whether the student will answer a question correctly. The dataset is described by 16 features for training and testing files. Although there are no publicly available notebooks, the public and private leaderboards show 238 teams which is an indication of the quality and interest of the challenge. Unfortunately, there is no research paper describing the dataset or the notebooks from the leaderboards, so there is no indication regarding the approaches taken to tackle the problem,

The *Automated Essay Scoring*[33] and *Short Answer Scoring*[34] are two Kaggle competitions that were launched in 2012. Although currently there are 34 and 54, respectively kernels that are ranked according to the same objective criteria (i.e., quadratic weighted kappa error metric) there is no available public notebook with the solution. Unfortunately, the solutions may not have been described in scientific papers, so there is no possibility to study or improve any of the existing solutions. Unfortunately, the only thing that may be done is to create a notebook and start coding a kernel that will be ranked along with existing ones.

The *KDD Cup 2015*[35] tackled the problem of MOOC dropout prediction. The competition was run as a hackathon with dataset available only for contestants through a username and password and with a difficult way to visualize notebooks with kernels. Later, the dataset has been posted on a personal blog[36] without any other possibility of benchmarking new models.

Learning Analytics Hackathon[37] started in 2015. Since then every year a new competition was launched, the last one being finished in March 2020. Although most of the datasets are not publicly available for this competition, and there is no possibility to submit solutions further and get back evaluations, the proposed tasks are quite challenging. We mention several of the themes: building dashboards, explore practicalities, embedding learning analytics in pedagogic practice, multimodal learning analytics or integrating learning analytics in game-based assessment.

LAVA@UBC: Learning Analytics, Visual Analytics[38] is another hackathon that started in 2015, but which continued in 2017 and 2018. Although the first two editions do not provide publicly available datasets or information about the data models, we mention that the third edition uses the *Canvas Network Dataverse* which is presented in the previous subsection. This is a clear indication that EDM scientific community needs to work on publicly available data such that benchmarking becomes possible. The 2015 and 2017 hackathons had no specific predefined task; the 2018 hackathon was not organized as competition on two tracks: visualization (with Tableau) and development (with R and Python and Jupyther notebooks).

Students' Academic Performance Dataset[39] is described in Amrieh et al. (2016) and is available and ready to be used on Kaggle platform. The dataset has 460 instances described by 16 attributes and one class attribute, which has three possible values. The dataset consists of both demographic and scholar features, and it has 186 kernels that explore the data and offer results. Based on this dataset, there are 186 available kernels and have over 25,000 downloads. For this particular dataset, there is no task associated, and therefore there is no defined validation metric. Each user may define his task and start developing a public kernel as a notebook in python or R programming languages. The greatest limitation of this dataset consists of lacking a leaderboard in which kernels may be ranked according to a specific validation metric. The general task for which this dataset is designed is to predict the interval the students will fit at the end of the semester. The 185 kernels on Kaggle and 54 citations on GS show a reasonable large usage of the dataset with the EDM community.

NAEP 2017: ASSISTments Data Mining Competition[40] started under NAEP in 2017. The goals of the competition that made the available dataset are twofold: to do well at the task (mostly engineering-oriented) and to have the field learn from this competition (mostly science-oriented). Seventy-six attributes describe each of the 942,817 records from the dataset, and the dataset evaluation was made using both AUC and RMSE with the best result scoring 0.2618 for RMSE and 0.9551 for AUC. Regarding the publications related to this competition, there are eight prior studies which are presented on the competition site. The successful entries have been invited to submit their results at EDM2018 conference and to a special issue of the Journal of Educational Data Mining.

NAEP 2019: Educational Data Mining Competition[41] continues the competition first launched in 2017. The five winners will be featured at an AIED/EDM2020 conference workshop. Still, until then their solutions are briefly described

on the competition' site. For this dataset, it is a different approach having only seven columns for each student and 438,392 instances for training and for testing three datasets which logged the data at different time stamps. The solution evaluation was made using an aggregation score which is composed of two metrics: adjusted AUC and Adjusted Kappa. The winner solution used an XGBoost Regressor with an aggregated score of 0.5657. A significant difference compared to the previous competition is that this time, there is no previous publication that supports it presented on the site.

Kaggle: Students Performance In Exams[42] is an educational dataset available on the Kaggle platform. There are a total number of 8 attributes, five of them describing the instances and other three which can be fed to an algorithm along with the other five or be predicted as class attributes. The dataset was uploaded in 2018 on the platform and is not a big dataset offering the data logged for 1,000 instances. The only task defined on Kaggle regards the correlations between different attributes like *Lunch and Writing Score*. As the dataset is more engineering-oriented, it is not being supported by scientific publications. The dataset has 356 kernels (most of them about data visualization, exploratory data analysis and classification) and 14 discussions opened and therefore we conclude that it has been intensively used. Still, as it lacks evaluation methodology and therefore a leaderboard, we state that this is a limitation which deprives researchers of getting a better insight regarding the performance that can be obtained on a specific task.

Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)[43] is another dataset which is the subject of a competition described in Settles et al. (2018) and is mostly suitable for deep learning techniques. The data offered for the task comes from five Duolingo courses (three in 2018) with multiple translations in the following languages: Portuguese, Hungarian, Japanese, Korean and Vietnamese. For evaluation purposes during the competition, a weighted Macro F1 and the winning solution from 2018 used a Gradient Boosted Decision Tree (GBDT) and a recurrent neural network model (RNN). Solutions are presented in detail in Tetreault et al. (2018) as research papers. The competition from 2020 is *STAPLE: Simultaneous Translation And Paraphrase for Language Education*[44] and tackles a predictive task, that is to produce a high-coverage set of translations in the target language.

The CSEDM 2019 Data Challenge[45] is a competition that took place within second Educational Data Mining in Computer Science Education (CSEDM) Workshop within ninth International Learning Analytics and Knowledge (LAK19) Conference. The challenge was to build a student model that trains of previous students' programming process data and predicts if new unseen students succeed at a given programming task. The dataset[46] is from DataShop@CMU. Unfortunately, to the best of our knowledge, the results (i.e., source code of solutions, rankings, etc.) of the competition are not available, and further submission is not possible anymore.

2020 CSEDM Data Challenge[47] is a competition that took place within fourth Educational Data Mining in Computer Science Education (CSEDM) Virtual Workshop in conjunction with Educational Data Mining (EDM 2020) conference. The draft description of the data challenge[48] proposes CodeWorkout (CWO) dataset and other two possible secondary datasets—PCRS and OLI Python—for usage during the challenge. All datasets store information about programming process data: running attempt, compilation result, and errors obtained after submitting a solution (i.e., the source code) to the appropriate grading system. Unfortunately, there is no publicly available link to the datasets, results of the competition or papers regarding the proposed models.

EdNet: A Large-Scale Hierarchical Dataset in Education[49] is a new dataset described in Choi et al. (2019) which has also an competition associated with it. The dataset is based on data collected over 2 years by Santa; a multi-platform AI tutoring service with more than 780 K users in Korea available through Android, iOS and web. Four smaller datasets compose the dataset: KT1, KT2, KT3, and KT4 with different extents and each of the smaller datasets are very well documented. The datasets have a different approach offering the logged actions rather than a full set of features describing an instance, and it is more feasible for deep learning algorithms. This approach of offering a large amount of logged data allows users to have better flexibility in setting their tasks but implies more data preprocessing. In EdNet competition site[50] organizers offered only the fittest sub-dataset (KT1) which consist of students' question-solving logs, and the best results achieved 0.7368 accuracy and 0.7811 for AUC.

Riiid! Answer Correctness Prediction[51] is a very new dataset available on Kaggle platform and which has a competition assigned to it. The e-Learning context consists of lectures and questions stored in two *.csv* files. These files contain information about 418 distinct lectures mainly composed of 230 concepts and 170 solving questions. We mention that the questions are of multiple choice. The training file gathers recorded actions of 393,656 unique users that asked 13,523 distinct questions. The total of actions recorded consists of 99,271,300 asked questions and 1,959,032 lecture watching. The whole dataset has 5.45 GB and is very well described and structured. The public leaderboard gathers the scores from 1,521 teams which is a clear indication regarding the quality of the dataset. The task associated with the dataset is to build a knowledge tracing model that predicts how students will perform on future interactions, and the validation uses the classical area under the ROC curve metric.

3.6 | Standalone EDM datasets

Multimodal learning Math Data Corpus[52] has been described in Oviatt et al. (2013), a paper that has 26 citations. A particularity of the dataset is that it contains high-fidelity time-synchronized multimodal data recordings (speech, digital pen, images) on collaborating groups of students as they work together to solve mathematics problems. The tackled issues are expertise estimation, prediction of problem-solving in mathematics or prediction of participation style.

NUS Multi-Sensor Presentation (NUSMSP) Dataset[53] is described in Gan et al. (2015) and consists of data from 51 unique individuals (32 males and 19 females). The data itself consists of video, 6 Android sensor data, depth and audio, and has been used for providing automatic feedback to entry-level to students for developing basic oral presentation skills, detection of patterns in oral presentations or automated assessment of presentation. Still, the presentations cover a diverse set of topics. Therefore presentations are poorly connected to education, and therefore the findings themselves are not related to education, teaching or learning.

Learn Moodle August 2016[54] is also a standalone dataset which has no published paper describing the dataset. Which is hardly used in cited research, therefore its utility is rather challenging to be appraised. Still, the dataset has a detailed description as it gathers data from 6,119 students exported in six *.csv* files. The greatest shortcoming is that the data files contain raw information, such that data analysis may require heavy preprocessing to obtain well-formed train and test data.

Lix Puzzle-game Data Set[55] is shortly described in Vahdat et al. (2016) and contains 15 *.cvs* file, one for each participant at experiments of collecting data from gaming interactions during playing the game called Lix. Each file contains 11 features describing in detail the actions performed during gameplay. Currently, there are only two citations, and game monitoring is not strictly related to educational processes.

Student Life Dataset[56] is described in Wang et al. (2014). Later, Wang et al. (2017) used the dataset to assess mental health and academic performance on the data that encapsulates the actions collected from 48 undergraduate students during a 10-week long period. The dataset is very well maintained and documented having a list of 16 publications and eight presentations available on the site. The publications tend to analyze the dataset mainly from the psychological perspective, and one of the most interesting tasks is to predict the student's GPA based on the data collected from the phones. One valuable mention is that for this dataset there is also an R package[57] available which aims to help navigate and analyze. As the current number of citations is 577, we conclude that the dataset has been intensely used successfully in experiments. The particularity of this dataset is that it consists of sensor data from smartphones (i.e., accelerometer, microphone, light sensor, GPS) and self-reports to determine behavior (i.e., activity, conversation, sleep or location) along with other indicators like stress or mood.

Dataset for empirical evaluation of entry requirements into engineering undergraduate programs in a Nigerian university[58] is described in Odukoya et al. (2018). The dataset covers Engineering Education and gathers raw data from 2005 until 2009 and gather statistics regarding age and 4 scores with a target class. Among the tackled tasks from the 9 citations, we mention predicting the performance of first-year student or analysis of the relationship between students' first-year results and their final graduating grades.

MUTLA: A Large-Scale Dataset for Multimodal Teaching and Learning Analytics[59] dataset is described in Xu et al. (2019). The dataset is very well described and covers many academic subjects (i.e., Mathematics, English, Physics, and Chemistry) and gather data from three sources: user records at question level log of student responses, brainwave data and webcam data. Although the dataset has no any known usage or citation yet, the description and available data makes them very suitable for EDM processes.

4 | CONCLUSIONS AND FUTURE WORKS

The shift of open data movement has been recently addressed by Machado et al. (2019), which also mentions the "limited support for researchers in education to generate, access, and share experimental data using openly available digital education platforms." Still, this review paper is focused on the process of conducting experiments and communicating results. We conclude that the presented limitations of EDM publicly available datasets in conjunction with our conclusions regarding actions that need to be taken on public data sources and repositories represent a valuable road-map for the next generation datasets.

This paper presents the main publicly available datasets that have been used in EDM processes. We have identified three types of data sources:

1. General-purpose repositories in which educational datasets are uploaded. In this category, there are UCI ML, Mendeley and Harvard Dataverse data repositories. A particular case is the one of *DataShop@CMU*, which is a repository of datasets that belong only to the EDM domain and also has integrated tools for performing various EDM tasks. The strongest limitation of this approach regards the difficulty of benchmarking the results published by other researchers. The difficulty comes from the fact that research papers that describe the approach are scattered along with many conferences and journals and running proposed approaches may be technically challenging in terms of used programming languages, packages or even validation methodologies. The most suitable approach is trying to replicate the data analysis and findings from a research paper and try to improve the reported quality metrics and finally submit the results to the very same publication or conference. We observe that any research paper does not describe datasets from Harvard Dataverse and therefore, this may explain their low usage by EDM scientific community.

2. Datasets used in EDM competitions represent a category that becomes very popular in the last years. The greatest advantage of the competitive approach represents the quick benchmarking of the proposed solutions against other competitor's solution. The greatest limitation is that tackled tasks may not be as elaborate as they may be in real live e-Learning system's scenarios. Still, obtained models represent valuable information for researchers and developers that may want to integrate similar EDM workflows in their particular e-Learning systems. Another limitation regards the openness of data analysis pipelines notebooks and their documentation. The first aspect may be specified by the organizer of the contest. In contrast, the second regards the incentive of the author to describe the solution properly. The most desirable situation is when a complete data analysis pipeline is presented as a scientific paper into a good conference or journal. The hackathon option is a flavor of competition which runs for a small timeframe—usually two days—and whose results (i.e., datasets, kernels, validation metrics and findings) may not be all publicly available. Still, event in this situation it is up to the organizer to which extent the resources (i.e., the dataset and its description, models and their descriptions) are publicly available.

3. Standalone EDM datasets represent a small category of data sources. The dataset maintenance is at the disposal of the author, as well as the proposed solutions and their results. These datasets may be hard to find by the academic community, and there may be a little incentive in using them in academic research.

Therefore, the answer to the first research question is that we have found that there are 41 publicly available EDM datasets belonging to three categories (i.e., general-purpose, competitions and standalone). Along with these, there is *DataShop@CMU* dataset which is a dedicated system for datasets management with available integrated data analysis tools. Regarding the data types that may be found in EDM datasets we have found the following categories: (a) demographic and social data, (b) performed learning activities data, (c) text data, (d) multimodal (i.e., speech, digital pen, images, brainwave and webcam) data, and (e) multi-sensor data (i.e., video, Android sensor data, depth and audio). Accordingly, that tackled tasks are strictly correlated with available data and mostly fall into the area of learner's modeling and prediction of the final grade, knowledge level, learner's answer, learner's difficulties, dropout or engagement. More elaborate analysis is required for inferring learning behavior, behavior patterns. Lately, particular interest has been shown towards using NLP techniques for topic detection, detection of cheating, automatic assessment of essay, early reading/writing assessment or second language acquisition modeling. Finally, multimodal and multi-sensor data opened the way for finding patterns in oral presentations, automated evaluation of presentation or stress detection.

Taking into account the investigated data sources, we conclude with the following actions that may be taken into consideration by EDM actors (i.e., independent researchers, academic institutions, private companies, etc.) that are willing to make their datasets public.

1. The dataset should be uploaded into a well-known public repository such as UCI ML, Mendeley or Harvard, or on a crowd-sourced platform that manages data science competitions. Currently, we observe that there are several types of platforms: (a) for competitions in any application domain (i.e., Kaggle) in which there are EDM datasets and tasks; (b) for specific application-oriented competitions (i.e., Duolingo); (c) run by professional organizations (i.e., National Assessment of Educational Progress [NAEP]); (d) standalone competitions like EdNet; or (e) hackathons.
2. Dataset authors should also thrive on submitting scientific papers with proper description to good conferences or journals, such that the articles are indexed in international databases and therefore may be easily found by EDM researchers. A dataset is more successful as it is highly used in experimental results of research papers and therefore the results may be found as citations.

3. Public repositories and crowd-sourced platform should provide a dataset documenting and task benchmarking functionalities. A key functionality on next-generation dataset repositories should be a multifaceted dashboard with the list of scientific papers which have used the dataset in their experiments, tackled tasks, algorithmic approaches, used programming languages and libraries or packages, results of validation and findings. The main goal is that the scientific community to be able to obtain quickly comparative results for the same task with all needed documentation. For this, dataset authors must provide train and test datasets, detailed task description and validation metrics.

In conclusion, benchmarking EDM tasks performed on large longitudinal publicly available datasets may benefit from being set up in a simulation/competition environment. This approach represents the way to go for next generation of data repositories that might be used by other researchers in an attempt to automatically gain strong evidence that personal contributions represent progress for the tackled task.

CONFLICT OF INTEREST

The author has declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Marian Cristian Mihaescu: Conceptualization; data curation; methodology; resources; writing-original draft; writing-review & editing. **Paul Stefan Popescu:** Conceptualization; data curation; resources; validation; writing-original draft; writing-review & editing.

ORCID

Marian Cristian Mihaescu https://orcid.org/0000-0003-0350-0441
Paul Stefan Popescu https://orcid.org/0000-0003-4504-6144

ENDNOTES

[1] Big Data and Education, https://www.upenn.edu/learninganalytics/ryanbaker/bigdataeducation.html
[2] Google Scholar, https://scholar.google.com/
[3] Google Dataset Search, https://datasetsearch.research.google.com/
[4] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php
[5] University Data Set, https://archive.ics.uci.edu/ml/datasets/University
[6] Teaching Assistant Evaluation Data Set, https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation
[7] Student Performance Data Set, http://archive.ics.uci.edu/ml/datasets/Student+Performance
[8] Student Alcohol Consumption, https://www.kaggle.com/uciml/student-alcohol-consumption/
[9] User Knowledge Modeling Data Set, http://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling
[10] Educational Process Mining (EPM): A Learning Analytics Data Set, https://tinyurl.com/y27yduo3
[11] Open University Learning Analytics dataset, https://tinyurl.com/y6ysfant
[12] Student Academics Performance Data Set, https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance
[13] Mendeley Data Repository, https://data.mendeley.com/
[14] Embeddings and topic vectors for MOOC lectures dataset, https://data.mendeley.com/datasets/xknjp8pxbj/1
[15] Data for: Effectiveness of flip teaching on engineering, https://data.mendeley.com/datasets/68mt8gms4j/2
[16] Dataset for Factors Affecting Teachers' Burnout, https://data.mendeley.com/datasets/6jmv43nffk/2
[17] Data of Academic Performance evolution for Engineering Students, https://data.mendeley.com/datasets/83tcx8psxv/1
[18] KEEL dataset repository, https://data.mendeley.com/datasets/py4hhv3rb8/1
[19] KEEL, http://www.keel.es/
[20] Outcome based education attainment calculation (OBE dataset), https://data.mendeley.com/datasets/9zkfwdm8xf/1
[21] Harvard Dataverse, https://dataverse.harvard.edu/
[22] HarvardX Person-Course Academic Year 2013, https://doi.org/10.7910/DVN/26147
[23] Canvas Network Person-Course, https://doi.org/10.7910/DVN/1XORAL
[24] Massively Open Online Course for Educators (MOOC-Ed), https://doi.org/10.7910/DVN/ZZH3UB

[25] CAMEO Dataset, https://doi.org/10.7910/DVN/3UKVOR
[26] Nursing Student Data, https://doi.org/10.7910/DVN/MQ8EP0
[27] Situated Academic Writing Self-Efficacy Scale Validation, https://doi.org/10.7910/DVN/M07HQ7
[28] Early Reading and Writing Assessment in Preschool, https://doi.org/10.7910/DVN/V7E9XD
[29] DataShop@CMU, https://pslcdatashop.web.cmu.edu/index.jsp?datasets=public
[30] KDD Cup 2010 Educational Data Mining Challenge, https://pslcdatashop.web.cmu.edu/KDDCup/
[31] RecSysTEL Datatel Challenge 2010, http://adenu.ia.uned.es/workshops/recsystel2010/datatel.htm
[32] What Do You Know?, https://www.kaggle.com/c/WhatDoYouKnow/
[33] Automated Essay Scoring, https://www.kaggle.com/c/asap-aes
[34] Short Answer Scoring, https://www.kaggle.com/c/asap-sas
[35] KDD Cup 2015, https://biendata.com/competition/kddcup2015/
[36] Blog by Philippe Fournier-Viger, https://tinyurl.com/y4csoswg
[37] Learning Analytics Hackathon,https://lakathon.org/about/history/
[38] LAVA@UBC: Learning Analytics, Visual Analytics, https://blogs.ubc.ca/lava/events/
[39] Students' Academic Performance Dataset, https://www.kaggle.com/aljarah/xAPI-Edu-Data
[40] NAEP 2017: ASSISTments Data Mining Competition, https://sites.google.com/view/assistmentsdatamining/home
[41] NAEP 2019: EDM Competition, https://sites.google.com/view/dataminingcompetition2019/
[42] Students Performance In Exams,https://www.kaggle.com/spscientist/students-performance-in-exams
[43] Duolingo Shared Task on Second Language Acquisition Modeling (SLAM), https://sharedtask.duolingo.com/2018.html
[44] STAPLE: Simultaneous Translation And Paraphrase for Language Education, https://sharedtask.duolingo.com/
[45] The CSEDM 2019 Data Challenge, https://sites.google.com/asu.edu/csedm-ws-lak-2019/
[46] KC Modeling for Programming, https://pslcdatashop.web.cmu.edu/Project?id=294
[47] 2020 CSEDM Data Challenge, https://sites.google.com/ncsu.edu/csedm-ws-edm-2020/data-challenge
[48] 2020 CSEDM Data Challenge Draft Description, https://tinyurl.com/y3kfu426
[49] EdNet dataset, https://github.com/riiid/ednet
[50] EdNet competition site, http://ednet-leaderboard.s3-website-ap-northeast-1.amazonaws.com/
[51] Riiid AIEd Challenge 2020, https://www.kaggle.com/c/riiid-test-answer-prediction/
[52] Multimodal learning Math Data Corpus, http://mla.ucsd.edu/data/
[53] NUS Multi-Sensor Presentation (NUSMSP) Dataset, https://scholarbank.nus.edu.sg/handle/10635/137261
[54] Learn Moodle August 2016, https://research.moodle.org/158/
[55] Lix Puzzle-game Data Set,https://sites.google.com/site/learninganalyticsforall/data-sets/lix-dataset
[56] Student Life Dataset, http://studentlife.cs.dartmouth.edu/
[57] *studentlife* GitHub repository, https://github.com/frycast/studentlife
[58] Dataset for empirical evaluation of entry requirements, https://tinyurl.com/yxqf2v42
[59] MUTLA, https://tinyurl.com/SAILdata

RELATED WIRES ARTICLE

Educational data mining and learning analytics: An updated survey

REFERENCES

Alon, U., Zilberstein, M., Levy, O., & Yahav, E. (2019). code2vec: Learning distributed representations of code. *Proceedings of the ACM on Programming Languages (POPL)*, *3*, 1–29.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), 119–136.

Baker, R., & Wang, E. (2015). *Big data and education*. New York, NY: Teachers College, Columbia University.

Baker, R., et al. (2010). Data mining for education. *International Encyclopedia of Education*, *7*(3), 112–118.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM—Journal of Educational Data Mining*, *1*(1), 3–17.

Bhoi, N. K. (2018). Mendeley Data Repository as a platform for Research Data Management. *Marching Beyond Libraries: Managerial Skills and Technological Competencies* (pp. 481–487). New Delhi, India: Overseas Press India Pvt. Ltd.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., ... Heo, J. (2019). *EdNet: A large-scale hierarchical dataset in education*. arXiv preprint arXiv:191203072.

Cortez, P., Silva, A. M. G. (2008). *Using data mining to predict secondary school student performance*.

Delahoz-Dominguez, E., Zuluaga, R., & Fontalvo-Herrera, T. (2020). Dataset of academic performance evolution for engineering students. *Data in Brief*, *30*, 105537.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *WIREs: Data Mining and Knowledge Discovery*, *9*(6), e1332.

Gan, T., Wong, Y., Mandal, B., Chandrasekhar, V., & Kankanhalli, M. S. (2015). Multi-sensor self-quantification of presentations. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 601–610). New York, NY: ACM. https://dl.acm.org/doi/10.1145/2733373.2806252.

Gómez-Tejedor, J. A., Vidaurre, A., Tort-Ausina, I., Mateo, J. M., Serrano, M. A., Meseguer-Dueñas, J. M., ... Riera, J. (2020). Data set on the effectiveness of flip teaching on engineering students' performance in the physics lab compared to Traditional Methodology. *Data in Brief*, *28*, 104915.

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, *118*(1), 177–214.

Ho, A., Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. In A. D. Ho, J. Reich, S. Nesterko, D. T. Seaton, T. Mullaney, J. Waldo & I. Chuang (Eds.), *HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No 1)* (p. 2014). Cambridge, MA: HarvardX. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263.

Hou, Y., Li, L., Li, B., & Liu, J. (2019). An anti-noise ensemble algorithm for imbalance classification. *Intelligent Data Analysis*, *23*(6), 1205–1217.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447–459.

Ihantola, P., Vihavainen, A., Ahadi, A., Butler, M., Börstler, J., Edwards, S. H., & Toll, D. (2015). Educational data mining and learning analytics in programming: Literature review and case studies. In *Proceedings of the 2015 ITiCSE on Working Group Reports*, 16, 41–63. New York, NY: ACM. https://dl.acm.org/doi/10.1145/2858796.2858798.

Kahraman, H. T., Sagiroglu, S., & Colak, I. (2013). The development of intuitive knowledge classifier and the modeling of domain dependent data. *Knowledge-Based Systems*, *37*, 283–295.

Kastrati, Z., Kurti, A., & Imran, A. S. (2020). WET: Word embedding-topic distribution vectors for MOOC video lectures dataset. *Data in Brief*, *28*, 105090.

Kellogg, S., & Edelmann, A. (2015). Massively Open Online Course for Educators (MOOC-E d) network dataset. *British Journal of Educational Technology*, *46*(5), 977–983.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of Educational Data Mining*, *43*, 43–56.

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, *4*, 170171.

Labarthe, H., Luengo, V., & Bouchet, F. (2018). Analyzing the relationships between learning analytics, educational data mining and AI for education. In *14th International Conference on Intelligent Tutoring Systems(ITS): Workshop Learning Analytics* (pp. 10–19).Quebec, Canada: HAL-CCSD. https://hal.archives-ouvertes.fr/hal-02015705.

Loh, W. Y., & Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*, 815–840.

Machado, J. S., Farah, J. C., Gillet, D., & Rodríguez-Triana, M. J. (2019). Towards open data in digital education platforms. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (Vol. 2161, pp. 209–211). Maceió, Brazil: IEEE. https://ieeexplore.ieee.org/document/8820885.

Merceron, A. (2015). Educational data mining/learning analytics: Methods, tasks and current trends. In *DeLFI WOrkshops* (pp. 101–109). CEUR-WS.org: CEUR Workshop Proceedings. https://dblp.org/rec/conf/delfi/Merceron15.

Northcutt, C. G., Ho, A. D., & Chuang, I. L. (2016). Detecting and preventing "multiple-account" cheating in massive open online courses. *Computers & Education*, *100*, 71–80.

Odukoya, J. A., Popoola, S. I., Atayero, A. A., Omole, D. O., Badejo, J. A., John, T. M., & Olowo, O. O. (2018). Learning analytics: Dataset for empirical evaluation of entry requirements into engineering undergraduate programs in a Nigerian university. *Data in Brief*, *17*, 998–1014.

Oviatt, S., Cohen, A., & Weibel, N. (2013). Multimodal learning analytics: Description of math data corpus for ICMI grand challenge workshop. In *Proceedings of the 15th ACM on International conference on multimodal interaction, Sydney Australia* (pp. 563–568). New York, NY: ACM. https://dl.acm.org/doi/10.1145/2522848.2533790.

Prasojo, L. D., Habibi, A., Yaakob, M. F. M., Pratama, R., Yusof, M. R., Mukminin, A., ... Hanum, S. (2020). Teachers' burnout: A SEM analysis in an Asian context. *Heliyon*, *6*(1), e03144. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6953709/.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs: Data Mining and Knowledge Discovery*, *3*(1), 12–27.

Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. *WIREs: Data Mining and Knowledge Discovery*, *7*(1), e1187.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs: Data Mining and Knowledge Discovery*, *10*(3), e1355. https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1355.

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018). Second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 56–65). New Orleans, Louisiana: Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-0506.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85–106.

Stamper, J., Niculescu-Mizil, A., Ritter, S., Gordon, G., & Koedinger, K. (2010). Bridge to algebra 2006–2007. development data set from kdd cup 2010 educational data mining. https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp.

Tetreault, J., Burstein, J., Kochmar, E., Leacock, C., & Yannakoudakis, H. (Eds.). (2018). *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications New Orleans*. Louisiana: Association for Computational Linguistics Retrieved from https://www.aclweb.org/anthology/W18-0500

Vahdat, M., Carvalho, M. B., Funk, M., Rauterberg, M., Hu, J., & Anguita, D. (2016). Learning analytics for a puzzle game to discover the puzzle-solving tactics of players. In *European Conference on Technology Enhanced Learning* (pp. 673–677). New York City: Springer. https://link.springer.com/chapter/10.1007%2F978-3-319-45153-4_89.

Vahdat, M., Oneto, L., Anguita, D., & Funk, M. (2015). Rauterberg M. 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* (pp. 352–366). New York City: Springer. https://www.springerprofessional.de/en/a-learning-analytics-approach-to-correlate-the-academic-achievem/2540078.

Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, *122*, 119–135.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., et al. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (Vol. 2014, pp. 3–14). New York, NY: ACM. https://doi.org/10.1145/2632048.2632054.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., et al. (2017). StudentLife: Using smartphones to assess mental health and academic performance of college students. In *Mobile health* (pp. 7–33). Cham: Springer. https://www.cs.dartmouth.edu/∼xia/papers/mobilehealth17.pdf.

Xu, F., Wu, L., Thai, K., Hsu, C., Wang, W., Tong, R. (2019). *MUTLA: A large-scale dataset for multimodal teaching and learning analytics*. arXiv preprint arXiv:191006078.

How to cite this article: Mihaescu MC, Popescu PS. Review on publicly available datasets for educational data mining. *WIREs Data Mining Knowl Discov*. 2021;11:e1403. https://doi.org/10.1002/widm.1403

APPENDIX A: Information about publicly available EDM datasets

Dataset	Description
Automated Essay Scoring and Short Answer Scoring[33,34]	2012, AES: 13K instances (i.e., essays), 11 attributes with 153 participating teams; SAS: 5 attributes, 151 participating teams.
CAMEO Dataset[25]	2015, 4 *.csv* restricted files with statistics about student actions in MITx and HarvardX courses.
Canvas Network Person-Course[23]	2016, 325,000 instances (i.e., performed activities), 25 attributes (i.e.,) from 238 courses.
Data for: Effectiveness of flip teaching on engineering students[15]	2019, 1,233 instances (i.e., students), 2 attributes (i.e., laboratory and final grades) on Physics and Electricity subjects over 4 years of flip and traditional teaching methods.
Data of Academic Performance evolution for Engineering Students[17]	2020, 12,411 instances (i.e., students), 44 attributes with academic, social and economic information.
Dataset for Factors Affecting Teachers' Burnout[16]	2020, 876 instances (i.e., teacher's responses), 23 features (i.e., fields) from a questionnaire whose aim is to predict teachers' burnout.
Dataset for empirical evaluation of entry requirements into engineering programs in a Nigerian University[58]	2018, 1,445 instances (i.e., undergraduates) for 5 years, 5 features (i.e., age and scores), used to predict the academic performance in engineering programs.
Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)[43]	2018, data from five Duolingo courses with multiple translations in the following languages: Portuguese, Hungarian, Japanese, Korean and Vietnamese
Early Reading and Writing Assessment in Preschool Using Video Game Learning Analytics[28]	2020, 331 instances (i.e., students), 25 attributes (among which 20 regarding phonological awareness, early reading and writing games, and scores in a standardized word reading and word writing assessment.
EdNet: A Large-Scale Hierarchical Dataset in Education[49,50]	2019, large-scale hierarchical dataset divided in 4 sections consisting of student interaction logs collected over more than 2 years from Santa platform.
Educational Process Mining Dataset (EPM)[10]	2015, 230.318 instances (i.e., recordings from 115 subjects), 13 integer attributes, 42 citations, used for predicting student difficulties.
Embeddings and topic vectors for MOOC lectures dataset[14]	2019, the dataset consists of word embeddings and topic distributions of 12,032 video lectures from 200 courses from Coursera.
HarvardX Person-Course Academic Year 2013[22]	2014, 338,223 instances (i.e., student-course records), 20 attributes (i.e., identification, demographics and activities) that has been used for effort and student success.
KDD Cup 2010 Educational Data Mining Challenge[30]	2010, 3 development datasets based actions gathered from 3,561 students and 2 challenge datasets based on actions gathered from 9,353 students which took algebra courses.
KDD Cup 2015[35]	2015, The dataset was available only for contestants and the task was to predict the MOOC dropout of users.
KEEL dataset repository[18,19]	Consists of 908 datasets that may be used for various ML tasks by KEEL software tool.
LAVA@UBC: Learning Analytics, Visual Analytics[38]	2015, 2017 and 2018, with the task run on *Canvas Network Dataverse* which is previously presented.
Learn Moodle August 2016[54]	2016, data from 6,119 students is organized in 6 *.csv* files and holds data about 4 weeks of activity during a course activity.
Learning Analytics Hackathon[37]	Yearly, since 2015, each year has its own topic (i.e., open dashboards, interoperability, LA in pedagogy practice, multimodal LA.)
Lix Puzzle-game Data Set[55]	2016, 15 instances (i.e., participants), 11 features describing in detail the actions performed during playing Lix game.
Massively Open Online Course for Educators (MOOC-Ed)[24]	2015, *.csv* files with detailed information on the communications between learners stored as a graph (i.e., Nodes/learners and Edges/ communication) that may be used for Social Network Analysis.

(Continues)

Dataset	Description
Multimodal learning Math Data Corpus[52]	2013, multimodal data (i.e., speech, digital pen, images) while solving mathematics problems.
MUTLA: A Large-Scale Dataset for Multimodal Teaching and Learning Analytics[59]	2019, multimodal dataset, 11 features (i.e, 7 for student records, and. 4 for brainwave/webcam data)
NAEP 2017: ASSISTments DM Competition[40]	2017, 942K instances, 76 features, 523.1 MB available on request.
NAEP 2019: Educational Data Mining Competition [41]	2019, 5 datasets for train and 1 for test 438,392 instances.
Nursing Student Data[26]	2020, 255 instances (i.e., students), 84 attributes (i.e., demographics and writing efficacy metrics) that may be used for assessing nursing student education.
NUS Multi-Sensor Presentation (NUSMSP) Dataset[53]	2015, multi-sensor data (i.e., video, 6 Android sensor data, audio) from 51 individuals during oral presentations.
Open University Learning Analytics dataset[11]	2015, consists of 7 *.csv files,* 60 citations, used for early identification of at-risk students or student engagement predictions.
Outcome based education attainment calculation (OBE dataset)[20]	2019, zip archive with 3 folders and 19 *.xls* data files for 20 programs, 1872 courses, 4,800 course outcomes by 3,850 students for 9,800 assessments and 34,650,000 quantified and mapped entries.
RecSysTEL Datatel Challenge 2010[31]	2010, There is no information about used dataset, results or documentation.
Riiid! Answer Correctness Prediction[52]	2020, Large dataset (5.45 GB) with almost 100M asked questions and 2M lecture watching performed by about 400K unique users.
Situated Academic Writing Self-Efficacy Scale Validation[27]	2020, 807 instances (i.e., undergraduate and graduate students), 29 attributes (i.e., demographics, English language and writing status)
Student Academics Performance Data Set[12]	2018, 300 instances, 22 nominal attributes, 43 citations, used for predicting student academic performance.
Students' Academic Performance Dataset[39]	2016, 460 instances (i.e., students), 16 attributes (i.e, demographic and scholar), 186 kernels, has no associated task
Student Life Dataset[56]	2014, dataset (53GB) consists of activities (continuous sensor data, 32,000 self-reports and pre-post surveys) from 48 undergraduates over the 10 week spring term.
Student Performance Dataset[7,8]	2014, 649 instances (i.e., students), 33 integer attributes, 394 citations, available both on UCI and Kaggle.
Students Performance In Exams[42]	2018, 8 features, 1,000 instances, 467 submitted notebooks for two tasks
Teaching Assistant Evaluation Data Set[6]	1997, 151 instances, 5 categorical and integer attributes, 1,340 citations mainly for the proposed method and not for the usage of the dataset.
The CSEDM 2019 Data Challenge[45]	2019, data is based on 89 students which produced 25,796 transactions in 56.02 h based on Math/Computer Science subject
The CSEDM 2020 Data Challenge[48,47]	2020, 70K records (i.e., submissions in Java) of 410 students to 50 problems. Dataset stores information about programming process data: running attempt, compilation result, and errors obtained after submitting a solution (i.e., the source code) to appropriate grading system.
University Data Set[5]	1988, 285 instances (i.e., universities), 17 categorical and integer attributes, no citation or known usage.
User Knowledge Modeling Data Set[9]	2013, 403 instances, 5 integer attributes, 96 citations, used for adapted federated learning, validation of newly proposed algorithms.
What Do You Know?[32]	2012, 179K instances (i.e., learners) for training, 16 attributes, 238 teams on private leaderboard with the task to predict whether a student will answer a question correctly.