1. **Metodologie**- în această secțiune descrieți cum ați selectat articolele pe care le parcurgeți, cum le-ați catalogat (dacă este cazul)

Am selectat 11 articole pe tema aleasa de pe IEEE Explore, Scopus si Science Direct

Mai sunt in plus inca 4 articole de pe Scopus si Web Of Science pe care inca nu le-am parcurs in detaliu si care raman ca backup

Mai este un articol de pe Science Direct care nu avea foarte multe detalii despre Feature Selection care ramane ca backup

Proces de selectare:

Feature selection, educational data mining, last 5 years:

* Science Direct, <https://www-sciencedirect-com.am.e-nformation.ro/>, Advanced search

Query: (“educational data mining” OR “edm”) AND (“feature selection” OR “variable selection” OR “attribute selection” OR “variable subset selection”)

Filters applied: 2018-2023

O imagine care conține text

Descriere generată automat

Rezultate = 187 -> filtrate manual la 30 (24 eliminand articolele din 2018 si 2019) (22 eliminand apoi articolele din 2020) (15 eliminand apoi articolele din 2021)

* IEEE Explore, <https://ieeexplore.ieee.org/>, Advanced Search

Query: ("All Metadata":"educational data mining" OR "All Metadata":"edm") AND ("All Metadata":"feature selection" OR "All Metadata":"variable selection" OR “All Metadata”:”attribute selection” OR “All Metadata”:”variable subset selection”)

Filters Applied: 2018-2023

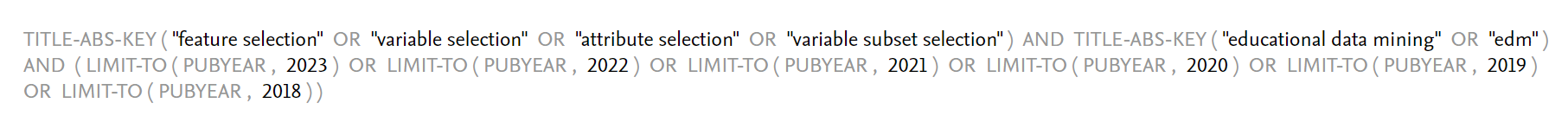
O imagine care conține text

Descriere generată automat

Rezultate = 32 -> filtrate manual la 20 (11 eliminand articolele din 2018 si 2019) (8 eliminand apoi articolele din 2020) (2 eliminand apoi articolele din 2021)

* Scopus, <https://www.scopus.com/home.uri>, Advanced search

Query: TITLE-ABS-KEY (“feature selection” OR “variable selection” OR “attribute selection” OR “variable subset selection”) AND TITLE-ABS-KEY (“educational data mining” OR “edm”) AND (LIMIT-TO(PUBYEAR, 2023) OR LIMIT-TO(PUBYEAR, 2022) OR LIMIT-TO(PUBYEAR, 2021) OR LIMIT-TO(PUBYEAR, 2020) OR LIMIT-TO(PUBYEAR, 2019) OR LIMIT-TO(PUBYEAR, 2018))



Rezultate = 139 -> filtrate manual la 25 (22 eliminand articolele din 2018 si 2019) (17 eliminand apoi articolele din 2020) (12 eliminand apoi articolele din 2021)

* Web Of Science, <https://www-webofscience-com.am.e-nformation.ro/>, Advanced search

Query: #4 OR #3 OR #2 OR #1

Refine By: Publication years: 2018 OR 2019 OR 2020 OR 2021 OR 2022 OR 2023

Where:

#1 Query: TI=((educational data mining OR edm) AND (feature selection OR variable selection OR attribute selection OR variable subset selection))

#2 Query: AB=((educational data mining OR edm) AND (feature selection OR variable selection OR attribute selection OR variable subset selection))

#3 Query: KP=((educational data mining OR edm) AND (feature selection OR variable selection OR attribute selection OR variable subset selection))

#4 Query: AK=((educational data mining OR edm) AND (feature selection OR variable selection OR attribute selection OR variable subset selection))

O imagine care conține text

Descriere generată automat

Rezultate = 143 -> filtrate manual la 11 (8 eliminand articolele din 2018 si 2019) (6 eliminand apoi articolele din 2020) (2 eliminand apoi articolele din 2021)

INITIAL

* Science Direct -> 187 articole
* IEEE Explore -> 32 articole
* Scopus -> 139 articole
* Web Of Science -> 143 articole

DUPA FILTRARE MANUALA

* Science Direct -> 30 articole
* IEEE Explore -> 20 articole
* Scopus -> 25 articole
* Web Of Science -> 11 articole

ELIMINAND ARTICOLELE DIN 2018 si 2019

* Science Direct -> 24 articole
* IEEE Explore -> 11 articole
* Scopus -> 22 articole
* Web Of Science -> 8 articole

ELIMINAND ARTICOLELE DIN 2020

* Science Direct -> 22 articole
* IEEE Explore -> 8 articole
* Scopus -> 17 articole
* Web Of Science -> 6 articole

ELIMINAND ARTICOLELE DIN 2021

* Science Direct -> 15 articole
* IEEE Explore -> 2 articole
* Scopus -> 12 articole
* Web Of Science -> 2 articole

DUPA INJUMATATIRE DUPA CITIRE ABSTRACT SI KEYWORDS

* Science Direct -> 8 articole
* IEEE Explore -> 1 articole
* Scopus -> 7 articole
* Web Of Science -> 2 articole

LISTA ARTICOLE INAINTE DE CITIRE INTEGRALA:

[1]

M. Kumar, C. Sharma, S. Sharma, N. Nidhi, and N. Islam, “Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance,” in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, Mar. 2022, pp. 1013–1017. doi: [10.1109/DASA54658.2022.9765236](https://doi.org/10.1109/DASA54658.2022.9765236).

[2]

A. Al-Zawqari, D. Peumans, and G. Vandersteen, “A flexible feature selection approach for predicting students’ academic performance in online courses,” *Computers and Education: Artificial Intelligence*, vol. 3, p. 100103, Jan. 2022, doi: [10.1016/j.caeai.2022.100103](https://doi.org/10.1016/j.caeai.2022.100103).

[3]

A. C. M. Yang, I. Y. L. Chen, B. Flanagan, and H. Ogata, “How students’ self-assessment behavior affects their online learning performance,” *Computers and Education: Artificial Intelligence*, vol. 3, p. 100058, Jan. 2022, doi: [10.1016/j.caeai.2022.100058](https://doi.org/10.1016/j.caeai.2022.100058).

[4]

S. Guzmán-Castillo *et al.*, “Implementation of a Predictive Information System for University Dropout Prevention,” *Procedia Computer Science*, vol. 198, pp. 566–571, Jan. 2022, doi: [10.1016/j.procs.2021.12.287](https://doi.org/10.1016/j.procs.2021.12.287).

[5]

H. Pallathadka, B. Sonia, D. T. Sanchez, J. V. De Vera, J. A. T. Godinez, and M. T. Pepito, “Investigating the impact of artificial intelligence in education sector by predicting student performance,” *Materials Today: Proceedings*, vol. 51, pp. 2264–2267, Jan. 2022, doi: [10.1016/j.matpr.2021.11.395](https://doi.org/10.1016/j.matpr.2021.11.395).

[6]

Y. Feldman-Maggor, R. Blonder, and I. Tuvi-Arad, “Let them choose: Optional assignments and online learning patterns as predictors of success in online general chemistry courses,” *The Internet and Higher Education*, vol. 55, p. 100867, Oct. 2022, doi: [10.1016/j.iheduc.2022.100867](https://doi.org/10.1016/j.iheduc.2022.100867).

[7]

J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, “Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization,” *Computers and Education: Artificial Intelligence*, vol. 3, p. 100066, Jan. 2022, doi: [10.1016/j.caeai.2022.100066](https://doi.org/10.1016/j.caeai.2022.100066).

[8]

W. Rujuan and W. Lei, “Research on E-learning Behavior Evaluation of Students Based on Three-way Decisions Classification Algorithm,” *Procedia Computer Science*, vol. 208, pp. 367–373, Jan. 2022, doi: [10.1016/j.procs.2022.10.052](https://doi.org/10.1016/j.procs.2022.10.052).

[9]

L. S. Rodrigues, M. dos Santos, I. Costa, and M. A. L. Moreira, “Student Performance Prediction on Primary and Secondary Schools-A Systematic Literature Review,” *Procedia Computer Science*, vol. 214, pp. 680–687, Jan. 2022, doi: [10.1016/j.procs.2022.11.229](https://doi.org/10.1016/j.procs.2022.11.229).

[10]

S. Verma, R. K. Yadav, and K. Kholiya, “A Scalable Machine Learning-based Ensemble Approach to Enhance the Prediction Accuracy for Identifying Students at-Risk,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, pp. 185–192, 2022, doi: [10.14569/IJACSA.2022.0130822](https://doi.org/10.14569/IJACSA.2022.0130822).

[11]

M. F. Yacoub, H. A. Maghawry, N. A. Helal, T. F. Gharib, and S. Ventura, “An Enhanced Predictive Approach for Students’ Performance,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 879–883, 2022, doi: [10.14569/IJACSA.2022.01304101](https://doi.org/10.14569/IJACSA.2022.01304101).

[12]

S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, “Educational data mining to predict students’ academic performance: A survey study,” *Education and Information Technologies*, 2022, doi: [10.1007/s10639-022-11152-y](https://doi.org/10.1007/s10639-022-11152-y).

[13]

S. Begum and S. S. Padmannavar, “Prediction of Student Performance using Genetically Optimized Feature Selection with Multiclass Classification,” *International Journal of Engineering Trends and Technology*, vol. 70, no. 4, pp. 223–235, 2022, doi: [10.14445/22315381/IJETT-V70I4P219](https://doi.org/10.14445/22315381/IJETT-V70I4P219).

[14]

A. B. Urbina-Nájera and L. A. Méndez-Ortega, “Predictive Model for Taking Decision to Prevent University Dropout,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 205–213, 2022, doi: [10.9781/ijimai.2022.01.006](https://doi.org/10.9781/ijimai.2022.01.006).

[15]

S. Begum and S. S. Padmannavar, “Student performance prediction with BPSO feature selection and CNN classifier,” *International Journal of Advanced and Applied Sciences*, vol. 9, no. 11, pp. 84–92, 2022, doi: [10.21833/ijaas.2022.11.010](https://doi.org/10.21833/ijaas.2022.11.010).

[16]

M. Q. Memon, Y. Lu, S. Yu, A. Memon, and A. R. Memon, “The Critical Feature Selection Approach using Ensemble Meta-Based Models to Predict Academic Performances,” *International Arab Journal of Information Technology*, vol. 19, no. Special Issue 3A, pp. 523–529, 2022, doi: [10.34028/iajit/19/3A/12](https://doi.org/10.34028/iajit/19/3A/12).

[17]

“Chicken Swarm-Based Feature Subset Selection with Optimal Machine Learning Enabled Data Mining Approach-Web of Science Core Collection.” <https://www-webofscience-com.am.e-nformation.ro/wos/woscc/full-record/WOS:000823686500001> (accessed Jan. 03, 2023).

[18]

“Feature Selection with Optimal Stacked Sparse Autoencoder for Data Mining-Web of Science Core Collection.” <https://www-webofscience-com.am.e-nformation.ro/wos/woscc/full-record/WOS:000779564300024> (accessed Jan. 03, 2023).

GREEN – CITITE SI SELECTATE

YELLOW – CITITE, NU TRATEAZA SUBIECTUL DESTUL DE IN DETALIU, PASTRATE CA BACKUP

RED – CITITE, PREA PUTINE DETALII DESPRE PROCESUL DE FEATURE SELECTION, ELIMINATE

WHITE – NECITITE, PASTRATE CA BACKUP

REZUMAT ARTICOLE:

ART 1 BAZAT PE SISTEMUL EDUCATIONAL PORTUGHEZ, SCOLI; FEATURE SELECTION (WRAPPER METHODS)

ART 2 ONLINE LEARNING; DETAILED FEATURE SELECTION

ART 3 ONLINE; SELF ASSESMENT; SOME DETAILS ON FEATURE SELECTION

ART 4 COLUMBIA; FEATURE SELECTION (RANDOM FOREST)

ART 5 UNIVERSITATE ITALIANA; TOO FEW DETAILS ON FEATURE SELECTION

ART 6 UNIVERSITATE DIN ISRAEL; ONLINE; VARIABLE SELECTION USED IN DECISION TREE ALGORITHM

ART 7 SLOVAKIA; FEATURE SELECTION DETAILED

ART 8 SUPERSTAR PLATFORM; ALMOST NO DETAILS ON FEATURE SELECTION

ART 9 SOUTH KOREA; MIDDLE AND HIGH SCHOOL; DETAILED METHODOLOGY; ALMOST NO DETAILS ON FEATURE SELECTION

ART 10 INDIA; FEATURE SELECTION ENOUGH DETAILED

ART 11 FEATURE SELECTION (SELECTKBEST)

ART 12

ART 13

ART 14 VERY DETAILED FEATURE SELECTION

ART 15 PORTUGUESE; FEATURE SELECTION BSPO

ART 16 BEIJING; FEATURE SELECTION (FILTER AND WRAPPER METHODS) – VERY DETAILED

ART 17

ART 18

SELECTATE IN FINAL:

11 articole

* 1 articol IEEE Explore
* 5 articole Science Direct
* 5 articole Scopus

The main motivation for this work is the observation that constructing a fixed set of features to predict the students’ academic performance is not straightforward.

The main objective of the present paper is to predict the academic performance of students with higher accuracy.

The second one is to find an effective methodology to predict students’ performance [4]. It is very crucial to apply the feature selection to discover high influence features that need to be improved the dropout rate and enhance student performance.

The purpose of this paper is to propose a predictive model for students’ performance prediction. This is achieved by exploring various classification techniques.

In order to answer research questions explained in Section 1.1, the PICO protocol [25] was adopted. This emphasizes the definition of four key elements: population, intervention, comparison, and outcome. Concerning the work presented, ”Population” refers to the service related to student work. a ”Population” refers to sets of predictive techniques and models used to predict student performance. ”Comparison” refers to variability in performance, and in the predictor variables of the valuation models. ”Outcome” refers to the accuracy of the approaches, predictor variables, as well as job performance. The initial research took a sample of articles that were cited multiple times in the review that was carried out regarding existing works in the area. Based on this set of articles, several search terms were tested through the search tool of the Scopus database and the IEEE database. After multiple interactions, the following search terms were selected: (”K-12” OR ”high school” OR ”secondary school” OR ”student” ) AND (performance or academic performance) AND (”artificial intelligence” OR ”machine learning” OR ”deep learning” OR ”learning analytics”) The research was carried out in October 2021, and publications from 2013 onwards were considered, since, according to [17], this is considered the period in which the EDM area began to receive greater notoriety.

From this initial search, the articles found were concatenated, resulting in 4544 works. Then, only articles written in English were filtered and duplicate works were removed using a combination of article title, DOI and year of publication, reducing the number to 4519. Then the remaining articles had their titles and abstracts read and were selected according to the inclusion criteria discussed. In all, 172 articles were selected for a deeper analysis, at this point only articles outside the scope proposed by this research were removed. During the last phase, 24 works were selected. In short, they deal with the prediction of academic performance in the context of high school or elementary school, either by using classical machine learning methods or deep learning and may also concern the selection of attributes relevant to the prediction of student performance.

• Q1A: How is academic performance defined? Academic performance in the delimited scope of this work is defined in the literature in 4 aspects: A performance range, such as pass/fail or in concepts such as (A,B,C,D and F) that assign a value to the performance shown by the student, a grade that represents a discrete value, ranging from a lower limit to an upper limit, assigned to exam performance or in an assessment period, or school subject. Also, a discriminant of retention/dropout, which indicates whether the student remained or left the educational institution.

• Q1B: What are the metrics used to assess the prediction of academic performance? In works that predict academic performance as a range of performance, models are generally evaluated in terms of their accuracy, but not only. Many works use more than one evaluation metric, however, because the problem of academic performance prediction is usually associated with unbalanced sets, the accuracy metric tends to hide problems in the classification [9]. Because of this, other articles choose to use the area under the receiver operating characteristic curve (AUC ROC), although it was previously presented by [30] that the area under the precisionsensitivity curve (AUC PR) is the most suitable – only the work presented by [20] used such metric.

• Q1C: What attributes are used to predict academic performance? The attributes used to predict academic performance, in the context of elementary and high schools, are similar to those found by [15] and [23], in predicting academic performance in programming courses and in predicting objectives of learning, respectively. That is, they are attributes that permeate the student’s dimensions, such as age, gender, personality, and the predominance of healthy habits; they are also attributing that relate to school factors, like presence, study time, engagement in classes, grades from previous periods; in addition, there are socio-economic attributes, such as parents’ profession, parents’ marital status and the number of cohabitants in the same house. Still, there are geographic attributes, such as distance from the student to the school and the location itself. It is noteworthy that the most direct attributes, such as student grades, are reported as the main predictors of academic performance, however, significant gains were reported when incorporating more indirect attributes into the generated model ([1] and [32])

• Q2: What is the quality of academic performance prediction works? Some points were raised in relation to the quality of the reviewed works. It is worrying that only 4% of the studies discuss threats to the validity of the studies since most of them (75%) did not have their model applied to a second population. It is reiterated that the problem of reproducing EDM studies had already been addressed by [18]. In his review [18] also pointed out the problem that studies predicting academic performance (in turn, in the context of teaching programming) showed unreliable results due to low reproducibility and the “evaluation of simplistic metrics”.

Thus, the systematic review carried out points to quality problems in the works, mainly in terms of the reproduction of experiments, delimitation of the scope of work, description of the variables used, and the need for public bases to carry out more effective comparisons between models.

• Q3: How are forecasting models applied in specific cases? Only one work was identified that presents an application, which uses a trained performance prediction model and seeks to alert an educational system about the student at risk of dropping out, given the context of high school and elementary school [20]. In this case, the proposed algorithm was applied to a national educational information system in South Korea, which is a centralized system for public education management. The work of [20] aims to improve the current model, however, there is little information about the architecture used and how the predictions of the early warning system are given as a whole. Finally, it is understood that the examples found do not answer the established research question, which sought greater detail regarding the applications that made use of the algorithms that are being developed in the area.

3.1. Challenges and Weaknesses of existing models In the proposed systematic review, several challenges and unexplored fields were identified in the prediction of academic performance in the context of elementary and high schools. Future work should pay attention to the following points:

• Academic performance prediction is an unbalanced problem by nature, so it is necessary to observe methods for balancing the data set and evaluation metrics that are relevant to this type of problem, such as the area under the precision-sensitivity curve.

• The use of data sets from more than one population in order to reinforce the validity of predictive models. It is noteworthy that the data set must have a large number of samples, otherwise, it is not possible to draw significant conclusions.

• There is a research gap in papers that use unsupervised models to predict academic performance

• Academic performance prediction using deep learning is being well explored in the field of universities, but there is this research gap in primary and secondary education.

• The development of applications to monitor the predicted performance of students still needs to be explored in elementary and high schools. These applications tend to be more useful as they become more interpretable [6]

4. Conclusion This systematic literature review investigated the issue of predicting the academic performance of students in the context of high school and elementary school. The search of two databases provided a synthesis of 24 works, which were within the determined scope. The articles were peer-reviewed and published between 2012 and October 2021. To the best of our knowledge, this is the first work that summarizes the effort of the educational data mining area to predict the academic performance of high school and elementary school students. Academic performance is observed in literature in multiple forms and definitions, although it could be distinguished into mainly four objectives that concern retention or some form of success measured by grades. Those are definitions that fit well when defining machine learning classification problems, which are the objectives of most of the analyzed work. When analyzing the metrics used to evaluate the proposed models, attention is drawn to the fact that multiple works are evaluating models trained on imbalanced data without using a metric that accounts for it, explicitly addressing this problem. This is concerning, especially, because it put into question the model’s actual performance. Also, there is a challenge concerning the proper description, and sharing of used datasets, to understand the used methodology. The observed challenges in the field include the adoption of larger and representative datasets of more than one population, the use of techniques that deal with the problem of data imbalance inherent to the problem, the adoption of appropriate evaluation metrics for unbalanced classes, the use of deep learning models and the creation of early warning systems based on interpretable models.