Visualizing the Educational Data Mining Literature

I. Papadogiannis, N. Platis, V. Poulopoulos, C. Vassilakis, G. Lepouras, M. Wallace,
G. Karountzou

***Abstract*—This article provides a visualization of a literature review in students' performance prediction using educational data mining (EDM) techniques for the period 2015-2019. The results of the review are presented concisely and simply with the use of diagrams. Various aspects of the literature are examined, such as the algorithms adopted, the type of results drawn, the educational setting of the application and the actual exploitation of the outcomes. Findings indicate that tertiary education dominates the EDM field; in contrast, the focus given to secondary and primary education is minimal.**

***Index Terms*—Visualization, Educational Data Mining, Student Performance, Literature review.**

I. Introduction

The use of data mining techniques in educational data has increased greatly in recent years. This has led to a huge increase in the amount of educational data now available. The introduction of information systems allows the recording and retention of large volumes of data in educational institutions. The development of modern as well as asynchronous distance learning has also increased the volume and type of data. Thus, the conditions have been for the application of data mining techniques in education and the educational data mining has been developed as a separate interdisciplinary discipline.

The prediction of academic performance of the students is a frequent choice of researchers in this scientific field. A significant number of studies have been published primarily for predicting the performance of students in higher education. The main goal is early detection of students with weaknesses to develop appropriate actions and policies on behalf of educational institutions. For the prediction of the students' academic performance, a large number of techniques and data mining algorithms have been applied and various types of explanatory factors have been used. As predictive variables have used demographic, socio-economic data, grades and other academic data.

We have recently prepared an extensive review of the relevant literature, soon to be published [1].

In this short paper we summarize this review and visualize its main findings. Readers interested in further details may seek the complete list of the referenced works and a discussion on the characteristics of each article in the original journal publication.

II. Visualizing the Literature Review

A. Articles' selection

The article selection process shown in the Figure 1:

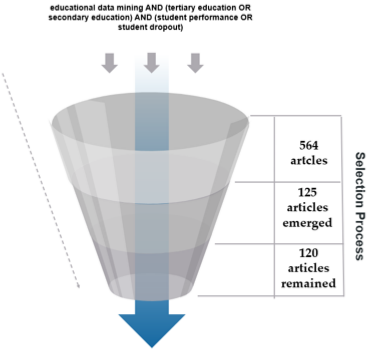

educational data mining AND (tertiary education OR secondary education) AND (student performance OR student dropout)
564 artcles
125 articles emerged
120 articles remained
Selection Process

Fig.1Selection process

In the first phase, we found 564 articles. Subsequently, the title and abstracts of the articles were sought, applying inclusion criteria. After that, 125 articles remained for a full study. After a thorough study and application of the inclusion and exclusion criteria, finally, 120 articles emerged. These articles have been widely criticized.

B. Literature review sources

The number of articles finally selected differs depending on the source. The Figure 2 shows in detail the percentages by a different source.

We observe that a large number of articles concerning journals appear less than two times. The majority of articles reviewed originated from IEEXplore (23%) Corresponding percentage articles derived from Springer (20%).

Published on December 19, 2020.

I. Papadogiannis, (e-mail: i.papadogiannis@uop.gr)
N. Platis (e-mail: nplatis@uop.gr)
V. Poulopoulos (e-mail: vacilos@uop.gr)
C. Vassilakis (e-mail: costas@uop.gr)
G. Lepouras (e-mail: gl@uop.gr)
M. Wallace (e-mail: , Wallace@uop.gr)
G. Karountzou (gkarountzou@uop.gr)
Authors are with the ΓAB LAB - Knowledge and Uncertainty Research Laboratory, Department of Informatics and Telecommunications University of the Peloponnese, Greece.

DOI: http://dx.doi.org/10.24018/ejers.2020.0.CIE.2306

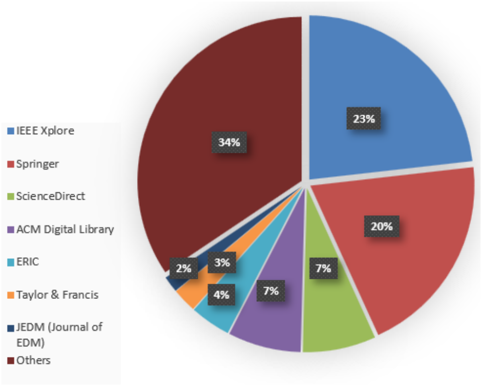


Fig.2 Literature sources

C. *Research field*

The vast majority of articles were on tertiary education. This may be due to better access to data through the development of Learning Management Systems (LMS) in higher education institutions, as well as the fact that more scientific experiments can be performed more easily in tertiary education. As shown in Figure3, the percentage of research related to universities or colleges reached 78.69 Research followed in secondary education at 14.75%, while less research conducted on online platforms.

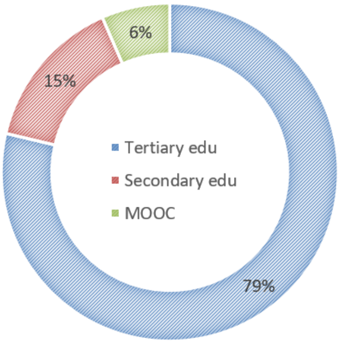


Fig.3 Research field

D. *Algorithms*

In most papers more than one algorithm was applied. We categorized the papers by methodological terms. As we present below the following categories was used:

Association rules. A category of machine learning algorithms that strive to extract interesting relationships between variables and create "if-then" statements.

Bayesian methods. Algorithms use Bayes' theorem to update the conditional probability for a hypothesis as more data become available.

Decision Trees. A category of non-parametric supervised learning methods that attempts to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Ensemble methods that combine particular base algorithms in order to compose a new optimal predictive model.

Instance-based learning. A family of learning algorithms that construct hypotheses directly from the training data, without any previous hypothesis.

Logistic regression is used to model the probability of a certain binary class or event. (e.g. pass/fail, win/lose etc).

Neural networks is a series of algorithms that efforts to identify underlying relations in a set of data through a process similar the way the human brain works. It is constructed as a combination of connected nodes and stimulates with the neurons of the biological organisms.

Support Vector Machine an algorithm that tries to discover a hyperplane in an N-dimensional space that clearly classifies the data in different categories and minimize some error mesure.

Linear regression is a statistic method for modeling the relationship between a dependent variable and one or more explanatory or independent variables.

The measures used for the evaluation were also recorded. The Figure 4 shows the frequencies per method.

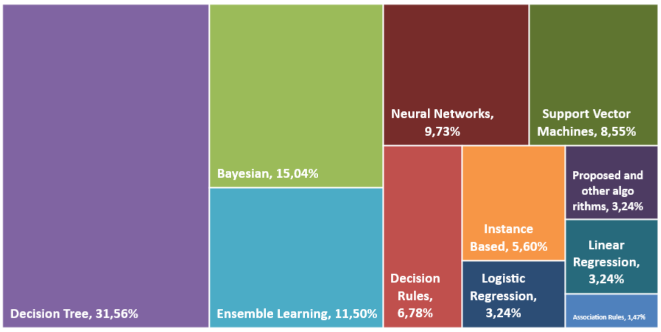


Fig.4 Frequencies per method

The accuracy was the most used evaluation tool. We observed differences in accuracy between different methods and datasets. It should be noted that in almost every article a dataset differed was used. Thus, the comparison of the accuracy of the algorithms only related in article level. Figure 5 shows the average accuracy per method.

We have observed high frequency and high accuracy score of Decision Tree algorithm. In high-frequency redounded the extensive use of the method in general and the huge number of algorithms accessible in familiar tools, such as WEKA. Other methods such as Bayesian algorithms, mainly Naive Bayes, have also been 45 times used. Despite its calculation speed and low resource consumption. In some cases, Naive Bayes was found to have the highest, while the average accuracy of the algorithm reached 0.7560.

Different methods such as Support Vector Machines, Ensemble Learning Methods and Neural Networks, have been applied to a minor extent and produced marginally lower accuracy. Logistic regression has been applied to a few articles, such as Instance-Based Learning. Next, we present in more detail the accuracy and the number of the most frequently applied algorithms.

Decision Trees algorithms accounted for the majority of the algorithms used with very high accuracy. Many algorithms included in this category, such as ID3, CART, C4.5 etc., so researchers had the opportunity to choose from many different approaches in the same paper. The C4.5 algorithm was the most accurate.

Logistic regression was shown a similar score in accuracy compared to Decision Trees proving that it is a powerful algorithm, although it was used in a small number of articles. Instance-based learning algorithms, SVM and Neural Networks showed lower average accuracy and used by fewer researchers.

DOI: http://dx.doi.org/10.24018/ejers.2020.0.CIE.2306

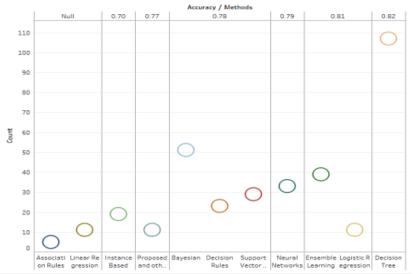


Fig.5 Average accuracy per method

Ensemble learning methods were used in a quite large number of articles. The most frequent algorithm was Random Forest which recorded an accuracy of 0.79. Only a few times Stacking and AdaBoost algorithms were used and showed better accuracy. We also noticed the very low frequency of use of Unsupervised learning techniques, such as K-means, in only two cases.

E. *Attributes*

The attributes used were tested with various measures. The Figure 6 presents the frequencies of difference attributes combinations of features used. The majority of papers we studied used student ranks as explaining variables (28.72%). Student demographics such as gender, profession of parents, age, etc were used at 23.40% and academic data at 21.81%. In smaller percentages, combinations of the previous attributes are used. Finally, other variables applied to a very minor proportion (behavioral data, internet logs, motivational data), while in two studies data from documents was used.

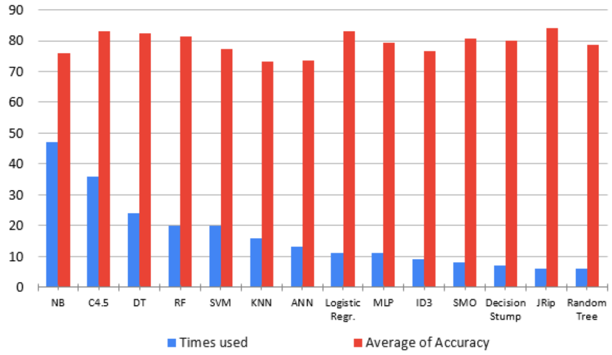


Fig.6 Average accuracy per algorithm

We also tested the average accuracy in terms of attributes used. We present only demographic, academic and grade data because of highest frequency of use. In Figure8, we observe that the highest accuracy has the use of grades as a predictive attribute (0.79476). Grades has also presented the smaller 95% confidence interval (0,74941 - 0,84011). The use only demographic and other academic data reduces accuracy (0.76905) but the 95% confidence intervals overlapped (0,70851 - 0,82959) and no statistically significant difference is displayed. The use of only demographic data alone was shown bigger accuracy (0, 94567) in only three cases and the use of academic information was shown an accuracy of 0,8828. As shown in the spider graph (Figure 8) the shape look like a normal heptagon and no combination showed particularly bigger accuracy.

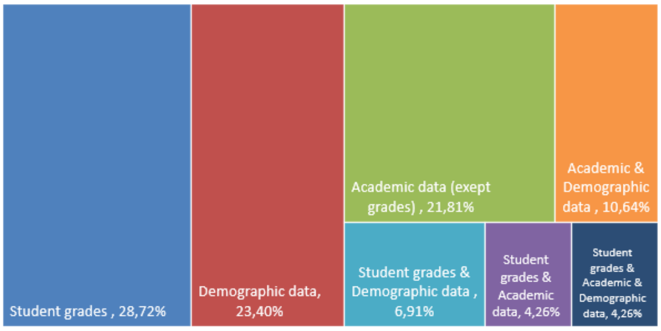


Fig.7 Attributes used

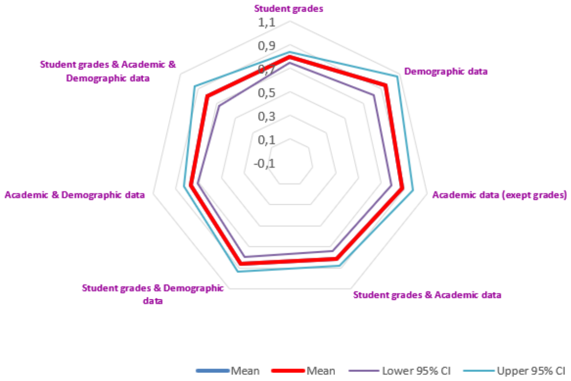


Fig.8 Accuracy per attribute

F. *Using the findings of the researches*

The using of research results for decision making can be very helpful for education authorities. Evidence- Based decision making enables strong decision support at local, regional, national and even supranational level. We studied the cases in which the results of the studies were used in decision making, according to what is mentioned in article. From the study of the articles we found that for the most part these are case studies in a limited sample. The main interest of the researchers was to evaluate the effectiveness of specific algorithms and techniques and not to use their results to conduct educational policy. However, there have been a few cases in which a tool has been developed for use by schools or universities. In Figure 9 we present six cases in which the paper targeting was the practical use of the findings.

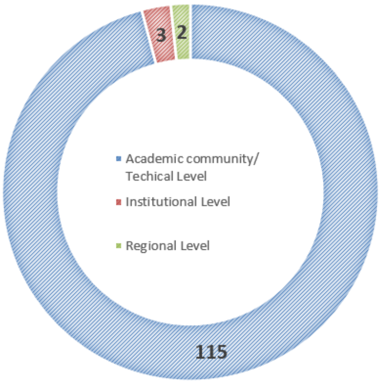


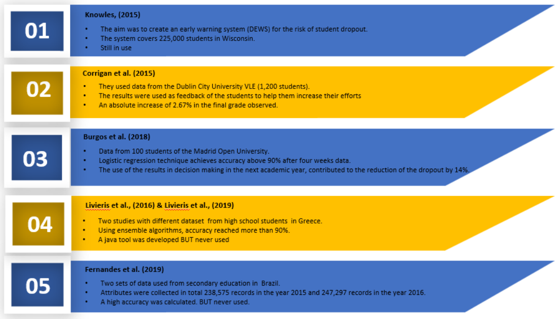
Fig.9 Findings target users

The reviewed papers are ruled by the studies of higher education, due to the higher availability of data and the closer connection that researchers have with organizations, allows more studies to carry out. On the other side, in Primary and Secondary Education, only a few studies have conducted although the field is greater and wider.

Many algorithms have been identified that allow for

DOI: http://dx.doi.org/10.24018/ejers.2020.0.CIE.2306

studies to evaluate their findings effectiveness in evaluating student performance. Our research has generally shown high accuracy levels. In only a few cases we observed low or very high accuracy and no statistically significant differences were found among the algorithms. No increase by use of more high-level algorithms was found. Decision trees are the majority of the methods adopted with satisfying accuracy levels. Naive Bayes, C4.5 and Random Forest algorithms were adopted more often and KNN was the only algorithms related to instance-based methods. Clustering (K-means) was used in only two papers. Ensemble methods have been implemented in fewer cases



01 Knowles, (2015)
- The aim was to create an early warning system (DEWS) for the risk of student dropout.
- The system covers 225,000 students in Wisconsin.
- Still in use

02 Corrigan et al. (2015)
- They used data from the Dublin City University VLE (1,200 students).
- The results were used as feedback of the students to help them increase their efforts
- An absolute increase of 2.67% in the final grade observed.

03 Burgos et al. (2018)
- Data from 100 students of the Madrid Open University.
- Logistic regression technique achieves accuracy above 90% after four weeks data.
- The use of the results in decision making in the next academic year, contributed to the reduction of the dropout by 14%.

04 Livieris et al., (2016) & Livieris et al., (2019)
- Two studies with different dataset from high school students in Greece.
- Using ensemble algorithms, accuracy reached more than 90%.
- A java tool was developed BUT never used

05 Fernandes et al. (2019)
- Two sets of data used from secondary education in Brazil.
- Attributes were collected in total 238,575 records in the year 2015 and 247,297 records in the year 2016.
- A high accuracy was calculated. BUT never used.

Fig.10 Tools developed

In the papers we studied, student marks were most frequently used as prognostic factor. Combinations of the students' grades with demographic and academic data were frequently applied. No combination led to higher accuracy.

Our results confirmed a satisfying level of accuracy. However, many different users of the findings have been almost neglected. In a few studies, a tool has been created or the results provided to the institutions to improve the education provided. In some cases, the practical implementation of the study findings has been reported. Such was the practical application of methods for early diagnosis of student failure and the improvement of student results through feedback were the only factual findings.

III. Conclusions

This review recognized a rich number of analysis methods as well as an obsession with their use within the academic community. There has been insufficient utilization of data mining studies that support educational policy-making and institutional decision-making. We believe that the expansion of research aimed at their utilization in the daily teaching life but also in support of decision making at the educational policy level should be an option. Internal feedback with thriving cases of using different algorithms and techniques, in the long run, can lead to the withering away of the scientific field. In contrast, the practical application of the findings will extend the scope of research and will be useful for the research and the educational community.

References

[1] I. Papadogiannis, M. Wallace, and V. Poulopoulos, “A critical review of data mining for education: what has been done, what has been learnt and what remains to be seen,” *International Journal of Educational Research Review,* vol. 5, no. 4, pp. 353-372, October 2020.

Ilias Papadogiannis is born in Tripolis in 1975. He holds two BA’s, a BA in Management from Athens University of Economics and Business (Management Dept), Athens, Greece, 2007, a BA in Accounting from The Technological Institution (TEI) of Patras (Accounting Dept), Patras, Greece, 1996, and a postgraduate degree (Msc) in Governance from the University of Peloponnese, Tripolis, Greece, 2015.He is a PHD Candidate in the field of Educational Data Mining in the University of Peloponnese, Tripolis, Greece.

He has worked in the financial sector ας security broker for Eurobank Security (2000-2005). .He has 15 years of experience in administrative positions in educational fields. He is currently HEAD OF THE ICT DEPARTMENT of the Regional Education Directorate of Peloponnese, which resides in Tripolis, Greece and an Erasmus+ promoter in the region of Peloponnese. He has also participated as a lecturer in seminars concerning educational issues and he has published in educational conferences. Some of his previous publications include: Digital citizenship in Greek Primary schools in Peloponnese, Citizenship Education: a problematic concept or a myth, Warsow, Poland, CiCeA, 2018. His main scientific interest is Data Mining, Statistics and Research Methodology, and he has participated in several educational research projects.

Nikolaos V. Platis was born in Athens, Greece, on 28 May 1973. Since February 2017 he is an Assistant Professor at the Department of Informatics and Telecommunications of the University of the Peloponnese, specializing in Computer Graphics and Visualization. He received his BSc in Mathematics from the University of Athens in 1995, his MSc in Information Technology from University College London in 1996, and his PhD in Computer Science, in the area of Computer Graphics, from the Department of Informatics and Telecommunications of the University of Athens in 2005. He has taught courses on Computer Graphics, Visualization, Multimedia and Programming in undergraduate and postgraduate level, mainly at the Universities of Athens and of the Peloponnese. He has also worked as a programmer and software analyst and has extensive experience with many current software, programming and application technologies. His research interests lie in the areas of Graphics and Visualization, with emphasis on simplification techniques and multiresolution methods for the effective processing of three-dimensional models and visualization of multi-dimensional data, as well as the visualization of various types of information.

Vasileios Poulopoulos has born in Kalamata in 1982 received his diploma from the Computer Engineer and Informatics Department of the University of Patras in 2005. I obtained my MSc and PhD from the same department in 2007 and 2010 accordingly in data mining and analysis from heterogenous sources of the web and especially big data Recently elected as an Assistant Professor at the department of the Digital Systems of the University of Peloponnese, while being a member of Knowledge and Uncertainty Research Lab of the University of Peloponnese from 2017. In 2019 I completed my post-doc performing research on the role of Big Data in Cultural Informatics, continuing the research on it.

He has worked for CTI-DIOPHANTUS (Research Institute) on several EU projects from 2002-2010 when I decided to turn to the private sector and especially decided to follow the “greek startup wave”. Being a founding member of Hellenic Startup Association and working for several projects, companies and startups from 2010 until 2015 a year that found me back to the University classes teaching for the Technological Educational Institute of Peloponnese.

His research interests include among others: data mining, knowledge extraction, big data, cultural heritage, personalization techniques, clustering techniques, categorization techniques as well as innovative web and mobile applications that could make our everyday life easier and better. He has published more than 50 papers in the aforementioned fields. Detailed information about the publications can be found in Google Scholar

DOI: http://dx.doi.org/10.24018/ejers.2020.0.CIE.2306

Costas Vassilakis was born in Arta, Greece in 1968. He is professor in the Department of Informatics and Telecommunications of the University of the Peloponnese, in the subject of Information Systems. He has received his degree from the Department of Informatics of the University of Athens in 1990 and his PhD from the same department in 1995. During the period 1991-1995 he received a scholarship from the Greek State Scholarships Foundation. He has conducted research in the subjects of information systems, systems software and netcentric systems, having published over 130 relevant papers in international scientific journals and conferences. He has participated in more than 25 international and national research and development projects. His scientific interests include information systems, semantic web, service-oriented architectures and information presentation issues.

George Lepouras is currently a Professor at the Department of Informatics and Telecommunications, University of Peloponnese. He holds a first degree in Mathematics from University of Athens, MSc in Information Technology from University of Strathclyde, Scotland and a PhD in Human Computer Interaction from University of Athens. Dr Lepouras is a senior member of ACM and has served as the chairman of the Greek ACM SIGCHI. His research interests include personal and task information management, multilingual interfaces, web based interfaces, virtual reality and augmented reality applications as well as cultural technologies. He is an author and co- author of more than 120 papers, of which more than 40 appear in international journals. He has participated in many national and European research and technological development projects, as a researcher or coordinator, among which are FP5 SmartGov, FP6 DELOS, FP7 Experimedia and H2020 CROSSCULT

Manolis Wallace was born in Athens in 1977. In 2001 I received a diploma in electrical and computer engineering and in 2005 a PhD in intelligent knowledge-based systems in uncertain environments, both from NTUA's School of Electrical and Computer Engineering. Since 2007 he is a faculty member at the Department of Informatics and Telecommunications of the University of Peloponnese, while at the same time and up to 2013 also a senior researcher at the Foundation of the Hellenic World. Before that, He was at the Athens Campus of the University of Indianapolis, where I served as the chair of the Department of Computer Science. His research interests lie in the meeting of computing and humans, specifically in areas such as cultural informatics, educational informatics, smart cities, personalization and so on, and since 2002 He have (co-) authored around 150 papers in these fields. Most of them can be found here or (uncurated) at Google Scholar. He serves, or have served in the past, as associate or guest editor in numerous journals and as general, local, program committee or publicity chair in numerous conferences.

Georgia Karountzou holds a PhD in Intercultural Education, a BA in Pedagogic from the National and Kapodistrian University of Athens, a MSc in Human Rights from the UCL University of London and a MEd in Educational Studies from the Hellenic Open University (HOU). She has conducted research in the fields of Intercultural Education, stereotypes and prejudices as well as in bullying and cyber-bullying. She is the former Director of Scientific and Pedagogical Guidance in Primary Education, Peloponnese School Counsellor, where she guided and supported the work of all School Consultants of Peloponnese. She is currently an Education Teaching Coordinator.

DOI: http://dx.doi.org/10.24018/ejers.2020.0.CIE.2306