Towards a Deep Analysis of High School Students' Outcomes

Adina Barila[1†] Mirela Danubianu[2††] and Andrei Marcel Paraschiv[1†]

[1†]Stefan cel Mare University of Suceava, Romania
[2††] Integrated Center for Research, Development and Innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for Fabrication and Control (MANSiD), Ștefan cel Mare University, Suceava, Romania

Summary
Education is one of the pillars of sustainable development. For this reason, the discovery of useful information in its process of adaptation to new challenges is treated with care. This paper aims to present the initiation of a process of exploring the data collected from the results obtained by Romanian students at the BBaccalaureate (the Romanian high school graduation) exam, through data mining methods, in order to try an in-depth analysis to find and remedy some of the causes that lead to unsatisfactory results. Specifically, a set of public data was collected from the website of the Ministry of Education, on which several classification methods were tested in order to find the most efficient modeling algorithm. It is the first time that this type of data is subjected to such interests.

Key words:
educational data mining, classification

1. Introduction

New information and communication technologies have penetrated massively into the education system in different forms. From learning management systems (LMS) such as Moodle, Canvas, Edmodo, Google Classroom, and live lesson support using Zoom, Google Meet or Teams applications (as lately happened due to the COVID-19 pandemic), to the analysis of the data collected through these systems in order to find useful information for future development processes. Although data-driven research has rapidly developed, education still face many issues, such as the difficulty to collect sufficient data for all cycles of education or the poor interaction between the stakeholders. They subsequently influence the quality of the educational act and the final students' outcomes [1].

In the Romanian education system, the completion of the high school cycle is marked by the Baccalaureate exam (formerly called "maturity"). The results obtained at this exam mark the completion of an important stage in the educational process, but are also the foundation for the transition to higher education, especially because many of the admissions in the higher cycle are based on the Baccalaureate results. In light of this, it is important to make an analysis of this exam in order to increase the percentage of those who successfully complete it.

We set out to start a project that would make an in-depth analysis of the results, often unsatisfactory, obtained in recent years at the Baccalaureate exam, an analysis that would be useful for decisions on how to adapt the requirements of high school education in Romania. Because the results of the statistical analysis provide useful information, but do not highlight in-depth the aspects of causality, we set out to make models that reveal new information that, analyzed, to allow the disclosure of the causes that led to the given situation. Data mining techniques are used for this purpose.

Its originality consists in the fact that it is for the first time when these techniques are applied with the objective of modeling the Baccalaureate exams. Data is collected from public sites transformed and integrated into a single dataset.

The rest of this paper is organized as follow. Section 2 describes the state of the art on the interest in data mining techniques applied to data from the educational environment, as a way to obtain useful information for its sustainable development. Section 3 refers to the Educational Data Mining (EDM) concept and process, Section 4 addresses the issue of finding appropriate classification methods for the data collected following the Baccalaureate exam, and Section 5 presents some conclusions.

2. State of the art

The widespread use of information and communication technology in education systems allowed the collection of important amounts of data that have provided a clearer picture of the educational processes and phenomena. The idea of analyzing these data is not new, but due to the specifics of the different levels of the education cycle, most projects addressed higher education.

Over time, several specific issues, materialized in questions such as: "What kind of courses attract more students?" "Can student performance be anticipated?" or "What factors affect student achievement and what is their influence?" found an answer by exploring this data.

An interesting work about how the ICT could be used to enhance Knowledge Management in higher education, that addresses multiple issues, including interrelations between higher education processes and knowledge management technologies, infrastructures and processes is presented in [2]. In [3] it is stated that the prediction of

Manuscript received June 5, 2021
Manuscript revised June 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.6.11

Community college students' transfer to four – year institutions through data mining techniques provides valuable information and "significant benefits for decision makers". A study on the application of data mining techniques on small data sets to predict whether students will be successful or not is presented in [4]. Most data used for analysis are related to students' profile and consider only final points for activities and exams. Finally, the study shows what a key influencer the final grade is for students A case study that mines students' data to analyze their learning behavior and to predict their results is presented in [5]. Recently educational data mining (EDM) emerged as a research area that aims to analyze data collected during teaching and learning activities.

The trend of concern for the use of EDM in order to develop a sustainable education is provided by the number of papers published last years on this topic. Fig. 1 shows this trend obtained by analyzing two international reference databases (Web of Science and Scopus) between 2002 and 2020.

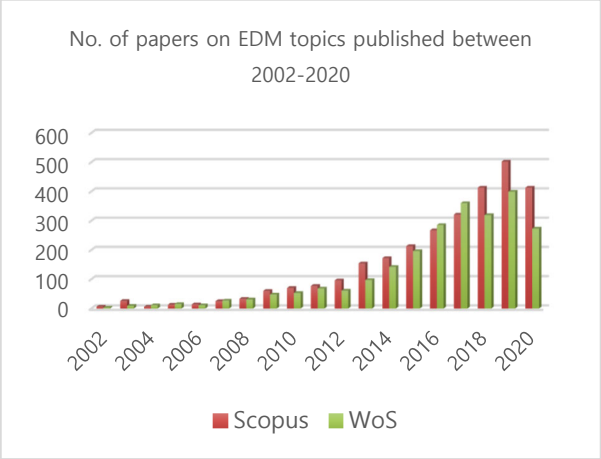

Fig. 1 Interest trend for educational data mining (based on data collected between 2002- 2020)

It is easy to notice an increasing trend, which denotes the expanded interest in conducting research in the field. A slight decrease is associated with 2020. This can be explained by moving the focus to COVID-19 topics. However, the health crisis and measures against the spread of the pandemic have forced the school at all levels to migrate online. In this context, important volumes of data, true "gold mines" for the discovery of valuable information for the adaptation of schools to the "education of the future", were generated and stored.

Despite the considerable number of projects and works in this field, it is found that an overwhelming percentage of research is directed towards higher education [6]. Sporadic, there are also references to secondary education. Our paper addresses aspects regarding the analysis of the results obtained for the Romanian High School graduation exam.

3. Educational Data mining (EDM)

Educational data mining (EDM) is an emerging field of research which aims to explore big amounts of data collected from educational environments in order to understand students' behavior and their results. It was defined in various ways. Calders et al. state that "EDM is both a learning science, as well as a rich application area for data mining, due to the growing availability of educational data. It enables data- driven decision making for improving the current educational practice and learning material" [7]. Romero and Ventura consider that EDM is the application of data mining (DM) techniques to datasets collected from educational environments to solve educational challenges [8] [9].

Concluding, EDM develops methods and techniques to build models from educational data for better understanding of students' behavior and their learning context. Obtaining such models is a complex process for which data mining is only the central stage.

The educational data used to build such models come from different sources, with different sizes and qualities. Event logs are often used, which indirectly track the interaction of each stakeholder with the platform. Educational data is also frequently found in LMS or other database systems. It can be information about student activities and can provide end-users with course-level statistics, through an aggregate view of all stakeholders' activity. In addition, summary information is presented on the partial and final grades and on each student's performance [10]. The educational data is often either sparse, noisy, inconsistent, or include several attributes that are irrelevant. Therefore, a data preparation step is required which involves cleaning of data, attributes transforming, features selection or dimensional reduction. This step is usually the one where the aim is to remove all unwanted data characteristics. For example, noisy data can be smoothed by binning. There are also used several techniques to overcome the problem of missing values. The incomplete records can be removed or completed with global constants or the most probable values [11]. However, the semantics of the missing values should be considered individually for all cases. EDM tasks may be categorized in five classes: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models [12] . The first three classes are traditional in data mining research, distillation of data for human judgment refers visualization and statistics analysis. Discovery with models is relatively new in EDM. It aims to develop a model to describe the

problem and to associate this model with another technique as a new component.
Obviously, the built models must be interpreted and evaluated to establish to what extent their quality is satisfactory.

4. Working methodology

4.1 Aim

In a broader project, we intend to conduct an in-depth analysis of causes for the results, often unsatisfactory, obtained in recent years at the Baccalaureate exam, an analysis that is useful for decisions on how to adapt to the requirements of high school in Romania.
This paper presents an incipient stage of research in which, on a dataset corresponding to the students' results obtained at the first Baccalaureate session in 2019, taken from the official page of the Ministry of Education, several methods of analysis are tested and compared. to decide which could be subsequently implemented. For this, complete Knowledge Discovery in Data (KDD) processes are developed in RapidMiner.

4.2 Data Set

We collected data from public web sites containing Baccalaureate results at country level.
The Baccalaureate is the national exam which every student graduating a high school must take it, especially if he/she want to enroll in university. The exam consists in 2 or 3 oral examinations, 4 written examinations and a digital skills examination (similar to ECDL exam). The oral examinations are at Romanian Language and Literature, a foreign language studied during the high school years and the native language and literature (for the students belong to an ethnic group). The digital skills examination is similar to ECDL exam and can be equivalent with it if the student holds an ECDL certificate.
The written examinations are the following:

- Romanian Language and Literature;
- the native language;
- compulsory discipline/subject depending on the graduated academic program (profile) followed in high school;
- chosen discipline depending on the graduated academic program;

After each written examination, the student may file an appeal to request a reassessment of the paper. The finale grade for a discipline is the grade obtained after appeal, if there was an appeal, or the initial grade, if the student didn't request a reevaluation. The dataset attributes are about the high school's name and the academic program the student has graduated, graduation year, form of education, the name of the subjects/disciplines the student must choose, the grades/marks for all examinations including for the possible appeals, the final average of the marks, hierarchical position in the county, hierarchical position in the country and the final result. The types of the attributes in our dataset are:

- numerical
  - real – the grades for each written examination (the initial grade, the grade obtained after appeal, the final grade), the final average
  - integer - hierarchical position in the county, hierarchical position in the country
- polynomial – the grades for oral examinations, the name of the compulsory discipline, the name of the chosen discipline, final result etc. Fig. 2 shows a sample of the initial data set.

Limba m odernă studiată - competenţe	Nota	Disciplina obligatorie a profilului - scris			Disciplina la alegere - scris			Competenţe digitale	Media	Rezultatul final
		Nota	Contestaţie	Nota finală	Nota	Contestaţie	Nota finală			
LIMBA ENGLEZA	A2-A2--A1-A1	ISTORIE			LOGICA, ARGUMENTARE SI COMUNICARE			Utilizator nivel mediu	7.4	REUSIT
		7.9		7.9	7.35		7.35			
LIMBA ENGLEZA	A1-A2-A2-B1-B2	MATEMATICA TEHN			ANATOMIE SI FIZIOLOGIE UMANA, GENETICA SI ECOLOGIE UMANA				5.95	RESPINS
		5.7	5.85	5.85	5.15		5.15			
LIMBA ENGLEZA	B1-A1--A1-A1	MATEMATICA TEHN			BIOLOGIE VEGETALA SI ANIMALA			Utilizator nivel mediu	5.76	RESPINS
		5.6	5.8	5.8	7	6.4	6.4			
LIMBA ENGLEZA	A2-A2-A2-B2-B2	MATEMATICA TEHN			BIOLOGIE VEGETALA SI ANIMALA			Utilizator nivel mediu		RESPINS
		1.2		1.2	1.75		1.75			
LIMBA ENGLEZA	B1-B2-B1-B2-B2	MATEMATICA TEHN			FIZICA TEH			Utilizator nivel mediu	5.5	RESPINS
		5.85		5.85	5.65		5.65			
LIMBA ENGLEZA	A1-A2--A1-A1	ISTORIE			GEOGRAFIE				6.8	REUSIT
		8		8	7.4		7.4			
LIMBA ENGLEZA	A2-B1-A2-B2-B2	ISTORIE			PSIHOLOGIE					NEPREZENTAT
		5.6		5.6	-2		-2			

Fig. 2 Sample of the initial collected dataset

As seen in the above figure, data from the initial dataset were not in a proper form for modeling. After a preprocessing step we obtained a dataset with 1630 cases and 26 features. Table 1 presents some of these features (using English /Romanian term, as in the dataset).

Table 1. Examples of features in data collected from Baccalaureate outcomes

Nr. Crt.	Attribute (English/Romanan)	Description	Domain
1	Student's school / Unit_înv	the name of the school	nominal : 43 distinct values
2	Previous graduated / Prom_ant	student graduated in previous years	binary: yes/no (da/nu)
3	Form of education Form_înv		binary: full-time/part-time (zi/ frecventa redusa)
4	Academic program / Specializare	the name of academic program the student graduated from	nominal : 43 distinct values
5	LLR-Skills / LLR-Competențe	the grade obtained by a student in Romanian Language and Literature – language skills test (oral examination)	Advanced user, Experienced user, Intermediate user, Absent (Utilizator avansat, Utilizator experimentat, Utilizator mediu, Neprezentat)
6	LLR-Writing / LLR-Scris	the grade obtained by a student in Romanian Language and Literature - written test	numeric: from -2 to 10 from 1 to 10 or -2 if the student was absent)
7	LLR- Contestation / LLR-Contestație	grade obtained after contestation	numeric: from 1 to 10
….			
24	Digital skills / Competențe digitale	the grade obtained by the student in digital skills test (a test similar to ECDL exam)	Advanced user, Experienced user, Intermediate user, Beginner user, Absent, NO (Utilizator avansat, Utilizator experimentat, Utilizator mediu, Utilizator incepator, Neprezentat, NU)
25	Average / Media	the average (arithmetic mean) is computed if the student obtained at least 5 in each test of the exam	numeric: from 5 to 10
26	Final result / Rezultatul final	final result: the student pass the exam if obtains an average scor at least 6.	3 values: Admis/Rejected/Absent (Reusit/Respins/ Neprezentat)

4.3 Modeling

For this stage, we established "final result" as target and we set out to develop classification models that would allow a better prediction of Baccalaureate results. These models allow to find those weaknesses in the basic training of students that lead to unsatisfactory Baccalaureate results and could be the foundation for decision-making on how to adapt the educational process in order to remedy them.

In addition to the issue of finding the best predictors, in this approach we faced the following challenge*: which method / algorithm provides the best classification performance?* To answer this question, we used the following classification methods: Deep learning, Logistic regression, Fast large margin, Generalized linear model, Nayve Bayes and Support Vector Machine.

Fig. 3 shows the level of models' performance in terms of the accuracy of the predictions, and Fig. 4 presents a comparative study of runtimes

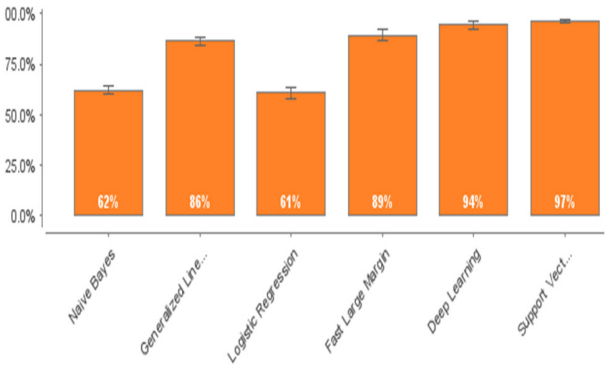

Fig. 3 Predictive models accuracy comparation

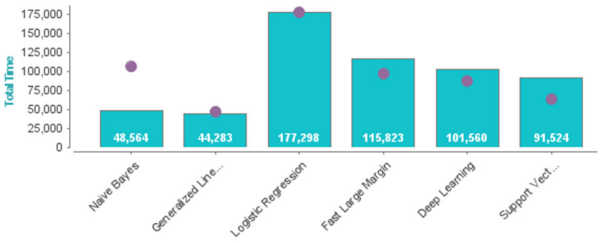

Fig. 4 Summary of runtimes

4.4 Discussion

Summarizing the above two characteristics, in Table 2, the following conclusions can be drawn:

- - the weakest performances, both in terms of accuracy and execution time are obtained for Logistic Regression;
- - the best accuracy for the data set considered is obtained for SVM;
- - the best time is provided by the Generalized Linear Model method.

Table 2. A comparative image of modeling methods performances

• Classification method	Accuracy [%]	Runtime [ms]
Nayve Bayes	62	48.564
Generalized Linear Model	86	**44.283**
Logistic Regression	61	117.298
Fast Large Margin	89	115.8923
Deep Learning	94	101.56
Support Vector Machine	**97**	91.524

Following these observations, and in accordance with the objective of this approach, we consider that in the near future we will deepen the research on the application of SVM for an extended data set with the results obtained at country level at the Baccalaureate exam.

Conclusions and future work

In this paper we conducted an initial research concerning a data mining system for analyzing the Baccalaureate exam in Romanian high schools, aiming to find some of unsatisfactory outcomes causes. We collected a data set from Romanian Ministry of Education on which we applied more classification algorithms to find to the most suitable for the proposed purpose. We used the following methods: Deep learning, Logistic regression, Fast large margin, Generalized linear model, Nayve Bayes and Support Vector Machine, and we found that, for our dataset the best accuracy is obtained from Support Vector Machine. As a result, for further validation we intend to use this algorithm on an extended data set, the results obtained for at least the last five years.

Acknowledgments

„This work is supported by the project *ANTREPRENORDOC,* in the framework of Human Resources Development Operational Programme 2014-2020, financed from the European Social Fund under the contract number 36355/23.05.2019 HRD OP /380/6/13 – SMIS Code: 123847."

References

[1] Wang, W.; Yu, H.; Miao, C. Deep Model for Dropout Prediction in MOOCs. In Proceedings of the 2nd International Conference on Cryptography, Security and Privacy, Beijing, China, 6–9 July 2017; pp. 26–32.

[2] W. Omona, T. van der Weide, and J. Lubega, 2010. "Using ICT to enhance knowledge management in higher education: A conceptual framework and research agenda", International Journal of Education and Development using Information and Communication Technology, vol. *6*(4), p.83-101, 2010.

[3] J. Luan, "Data mining and its applications in higher education" New directions for institutional research, Special Issue: Knowledge Management: Building a Competitive Advantage in Higher Education vol. *2002*(113), pp.17-36, 2002.

[4] S. Natek, and M. Zwilling, "Data mining for small student data set: Knowledge management system for higher education teachers" In Management, knowledge and learning international conference, Zadar. June *2013,*Vol. 1, p. 1379-1389, 2013.

[5] Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.

[6] Suhirman et al. Data Mining for Education Decision Support: A Review. **International Journal of Emerging Technologies in Learning (iJET)**, [S.l.], v. 9, n. 6, p. pp. 4-19, dec. 2014. ISSN 1863-0383. Available at: <https://online-journals.org/index.php/i-jet/article/view/3950> Date accessed: 14 May. 2021. doi:http://dx.doi.org/10.3991/ijet.v9i6.3950

[7] Calders T, Pechenizkiy M; Introduction to the special section on educational data mining. ACM SIGKDD Explor, 2011; 13(2): 3–6.

[8] Romero C, Ventura S; Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., 2010; 40(6): 601–618.

[9] Romero C, Ventura S; Data mining in education. Wiley Interdisc. Rev.: Data Min. Knowl. Discovery, 2013; 3(1):12–27.

[10] Yang, B.; Qu, Z. Feature Extraction and Learning Effect Analysis for MOOCS Users Based on Data Mining. Educ. Sci. Theory Pract. **2018**, 18, 1138–1149.

[11] Jin, C. MOOC student dropout prediction model based on learning behavior features and parameter optimization. Interact. Learn. Environ. **2020**

[12] Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

Adina BARILA (b. June 4, 1968) has obtained the B.S. degree in Computer Science from "Alexandru Ioan Cuza" University of Iași in 1990, and PhD degree in Computer Science from "Stefan cel Mare" University of Suceava in 2015. She is Lecturer of the Computers Department at "Stefan cel Mare" University of Suceava. Her current research interests include databases theory and implementation, data analytics, application of Data Science in education and economics.

Mirela DANUBIANU (b. July 13, 1961) has obtained the B.S. and M.S. degree in Computer Science from University of Craiova in 1985, and the PhD. degree in Computer Science in 2006 from "Stefan cel Mare" University of Suceava. She has also obtained the B.E. degree in Economics from University of Craiova in 2001. Currently, she is Associate Professor and Head of the Computers Department at "Stefan cel Mare University" of Suceava. She is the author/co-author of 5 books, 7 chapters and more than 100 papers which have been published in journals and presented at different conferences. Her current research interests include databases theory and implementation, modern data architectures, data analytics, application of Data Science in economics, education and healthcare.

Andrei Marcel PARASCHIV (b. June 28, 1982) received his BSc in European Studies (2005), MSc in European Studies (2007) from the Romanian-American University. Currently he is an IT Service Manager and Team Leader in the European Chemicals Agency (ECHA) in Helsinki, Finland. Previously he has been working as the Head of ICT in the Ministry for Research and Innovation in Bucharest, Romania. His current research interests include different aspects of Big Data applied in Project and Service Management.