BABEŞ-BOLYAI UNIVERSITY
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Mining students' behavioural features in multiple learning environments

PhD thesis summary

PhD student: Mariana-Ioana DINDELEGAN (căs. MAIER)
Scientific supervisor: prof. dr. Gabriela CZIBULA

2023

Keywords: Educational data mining, Behavioural mining, Machine learning, 21st century competencies, Students' performance analysis and prediction.

Contents

Thesis table of contents **2**

List of publications **5**

Introduction **7**

1 Background **12**

2 Mining behavioural features of school students **13**

3 Mining behavioural features of students in academic environments **14**

4 *IntelliDaM*: A machine learning-based framework for students performance analysis **15**

Conclusions and future work **16**

Bibliography **17**

Thesis table of contents

Glossary **4**

List of publications **9**

Introduction **11**

1 Background **16**
1.1 Educational data mining . . . 16
1.1.1 Students performance analysis and prediction . . . 17
1.1.2 Behavioural mining in education . . . 18
1.2 Students' profile analysis in the context of digitalisation . . . 19
1.2.1 21st century competencies . . . 19
1.2.2 Digital literacy in nowadays society . . . 20
1.2.3 Gamification . . . 21
1.2.4 Learning Computer Science concepts through gamification . . . 22
1.2.5 Learning Taxonomies . . . 23
1.2.6 Online learning and digitalisation . . . 24
1.3 Machine learning models used . . . 26
1.3.1 Unsupervised learning . . . 27
1.3.1.1 K-Means clustering . . . 27
1.3.1.2 Principal component analysis . . . 27
1.3.1.3 T-distributed Stochastic Neighbor Embedding . . . 28
1.3.1.4 Uniform Manifold Approximation and Projection . . . 29
1.3.1.5 Self-organizing maps . . . 29
1.3.1.6 Autoencoders . . . 30
1.3.1.7 Association Rules . . . 30
1.3.2 Supervised learning . . . 31
1.3.2.1 Linear discriminant analysis . . . 31
1.3.2.2 Linear Regression . . . 31
1.3.2.3 Polynomial Regression . . . 32
1.3.2.4 Logistic Regression . . . 32
1.3.2.5 Stochastic Gradient Descent . . . 33
1.3.2.6 Tweedie regression . . . 33
1.3.3 Feature selection . . . 34

2 Mining behavioural features of school students **35**
2.1 Statistical analysis methods used . . . 35
2.2 Mining 21st century skills at 4th grade students . . . 36

2.2.1 Experimental methodology . . . 36
2.2.2 Results and discussion . . . 37
2.2.3 Conclusions and future work . . . 41
2.3 Using unsupervised learning for mining behavioural patterns from data. A case study for the baccalaureate exam in Romania . . . 41
2.3.1 Methodology . . . 43
2.3.1.1 Research instrument . . . 43
2.3.1.2 Formalisation . . . 43
2.3.1.3 Analysis Methods . . . 44
2.3.2 Case Study . . . 45
2.3.2.1 Data set . . . 45
2.3.2.2 Data analysis . . . 46
2.3.3 Results and discussion . . . 46
2.3.3.1 Statistical analysis results . . . 46
2.3.3.2 Unsupervised learning-based results . . . 49
2.3.3.2.1 Association Rules . . . 49
2.3.3.2.2 Self-Organizing Maps . . . 50
2.3.4 Discussion . . . 52
2.3.5 Conclusions and future work . . . 52
2.4 Mining sorting concept across curriculum levels. A cyclic learning based approach 53
2.4.1 Mining sorting concept across curriculum levels . . . 54
2.4.1.1 Sorting Algorithms in Secondary School . . . 55
2.4.1.2 Sorting Algorithms in High School . . . 55
2.4.1.3 Sorting Algorithms in University . . . 56
2.4.2 Teachers' perception on approaching sorting concepts . . . 57
2.4.2.1 Secondary and high school teachers' perception on sorting . . . 57
2.4.2.2 University Computer Science teachers' perception on sorting 60
2.4.3 Conclusions and future work . . . 61

3 Mining behavioural features of students in academic environments 63
3.1 Mining academic data sets using unsupervised learning for student performance analysis . . . 63
3.1.1 Our approach . . . 64
3.1.1.1 Data set . . . 64
3.1.1.2 Experiments and setup . . . 65
3.1.1.3 Evaluation measures . . . 65
3.1.2 Results and discussion . . . 66
3.1.2.1 First experiment . . . 67
3.1.2.2 Second experiment . . . 69
3.1.3 Conclusions and future work . . . 71
3.2 Comparative assessment of academic performance in traditional and synchronous online learning environments . . . 71
3.2.1 The proposed approaches . . . 72
3.2.1.1 Theoretical model . . . 72
3.2.1.2 Data collection and preprocessing . . . 73
3.2.1.3 Performance evaluation metrics . . . 78

3.2.2 Using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance 79
3.2.2.1 Our approach 80
3.2.2.2 Analysis of the unsupervised learning results 81
3.2.2.3 Discussion 83
3.2.2.4 Conclusions and future work 89
3.2.3 Using self-organizing maps for comparing students' academic performance in online and traditional learning environments 89
3.2.3.1 Methodology 90
3.2.3.2 Results and discussion 91
3.2.3.3 Conclusions and future work 95

4 *IntelliDaM*: A machine learning-based framework for students performance analysis 96
4.1 Methodology 97
4.1.1 Formalisation 98
4.1.1.1 Unsupervised learning approach 98
4.1.1.2 Supervised learning approach 99
4.1.2 Feature analysis and selection 99
4.1.2.1 Statistical-based feature analysis 99
4.1.2.2 Feature selection 100
4.1.2.3 Feature sets quality analysis 100
4.1.3 Unsupervised learning-based analysis 101
4.1.3.1 Performance evaluation 101
4.1.4 Supervised learning-based analysis 101
4.1.4.1 Performance evaluation 102
4.2 Experimental results 104
4.2.1 Data sets 104
4.2.1.1 The data set $D_{2018-2020}$ depiction 104
4.2.1.2 The data set $D_{2020-2021}$ depiction 105
4.2.2 Experiments 106
4.2.3 Results 106
4.2.3.1 Experimental setup 107
4.2.3.2 Feature analysis 107
4.2.3.2.1 Statistical-based feature analysis 107
4.2.3.2.2 Feature selection 109
4.2.3.2.3 Feature sets quality analysis 109
4.2.3.3 Unsupervised learning-based analysis 110
4.2.3.4 Supervised learning-based analysis 111
4.2.4 Discussion 113
4.3 Conclusions and future work 115

Conclusions and future work 116

Appendix 117

Bibliography 121

List of publications

The ranking of publications was performed according to the CNATDCU (National Council for the Recognition of University Degrees, Diplomas and Certificates) standards applicable for doctoral students enrolled after October 1, 2018. All rankings are listed according to the classification of journals[1] and conferences[2] in Computer Science.

Publications in Web of Science - Science Citation Index Expanded

[MCD23] **Mariana-Ioana Maier**, Gabriela Czibula, Lavinia-Ruth Delean, *Using unsupervised learning for mining behavioural patterns from data. A case study for the baccalaureate exam in Romania*, Studies in Informatics and Control, 32(2), pp. 73-84, 2023. (**AIS Quartile Q4** according to JCR 2022)

Rank C, 2 points.

[CCML22] Gabriela Czibula, George Ciubotariu, **Mariana-Ioana Maier**, Hannelore Lisei, *IntelliDaM: A machine learning based framework for enhancing the performance of decision-making processes. A case study for educational data mining*, IEEE Access, Volume 10, pp. 80651–80666, 2022 (**AIS Quartile Q2** according to JCR 2021)

Rank B, 2 points.

[MCOM21] **Mariana-Ioana Maier**, Gabriela Czibula, Zsuzsanna-Edit Oneţ-Marian. *Towards using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance*, Mathematics, Special issue on Didactics and Technology in Mathematical Education, 2021, 9(22), 2870 (**IF Quartile Q1** according to JCR 2020)

Rank A, 8 points.

[OMCM21] Zsuzsanna Oneţ-Marian, Gabriela Czibula, **Mariana Maier**. *Using self-organizing maps for comparing students' academic performance in online and traditional learning environments* (2021). Studies in Informatics and Control, 30(4), pp. 1–11 (**IF Quartile Q3** according to JCR 2020)

Rank C, 2 points.

[1] https://uefiscdi.ro/premierea-rezultatelor-cercetarii-articole
[2] http://portal.core.edu.au/conf-ranks/

Publications in Web of Science - Conference Proceedings Citation Index

[McA22] **Mariana Maier**, Camelia Serban, Andrei Moisin. *Mining Sorting Concept across Curriculum Levels: A Cyclic Learning Based Approach* (2022). The 4th International Workshop on Education through Advanced Software Engineering and Artificial Intelligence, workshop of ESEC/FSE conference, pp. 10-17

Rank A, 6 points.

[CCCD20] Liana Maria Crivei, Gabriela Czibula, George Ciubotariu, **Mariana Dindelegan**. *Unsupervised learning based mining of academic data sets for students' performance analysis* (2020). IEEE 14th International Symposium on Applied Computational Intelligence and Informatics, SACI 2020, Timisoara, Romania, pp. 457-462.

Rank D - CORE2020, 0.5 points.

Papers published in international journals and proceedings of international conferences

[Din18] **Mariana Dindelegan** (2018). *Digital and Coding Literacy for School Students*. Studia UBB Digitalia, Volume 63 (LXIII) 2018, June, Issue 1, 55-68. (**indexed Central & Eastern European Online Library**)

Rank D, 1 point.

[FPDP+14] Silvia Ferent Pipas, **Mariana Dindelegan**, Bogdan Padurean, Emilia Ciupan, Cornel Ciupan (2014). *Cost calculator for water jet, laser and plasma machining*. Acta Technica Napocensis Series: Applied Mathematics and Mechanics, Volume 57, 2014, March, Issue 1, 73-76. (**indexed Index Copernicus**)

Rank D, 0 points.

Publications score: 21.5 points.

Introduction

Educational and Behavioural Data Mining are the main research fields of our PhD thesis. Our PhD thesis is entitled "*Mining students' behavioural features in multiple learning environments*" and aims to get an overview of the most important characteristics of school and university students in different learning environments. *Educational data mining* (EDM) represents a substantial domain of research, where the main objective is to find significant patterns in data collected from various educational environments. *Behavioural mining* (BM) is a subfield of *Data Mining* (DM) focused on extracting behavioural patterns from data. Our thesis intends to put together these domains, in order to obtain an image of nowadays students' characteristics in different learning environments.

In educational environments, DM offers methods to support decision-making and thus provide decision support. Uncovering meaningful patterns and extracting knowledge from education-related data sets is a challenging and intensively investigated topic in the EDM literature, particularly in relation with Covid-19 pandemic [GSA21, CcMS+20]. A major target in EDM is to comprehend the students' learning process, predict students' learning outcome, provide a better comprehension of the education-related phenomena and help education institutions to understand and improve their education-related processes [BCR18]. Nowadays, academic institutions are more and more interested in improving their teaching methodologies, learning processes [MT13] and the academic performance of their students and instructors [JRHR15]. EDM addresses techniques to understand the learning processes and identify patterns in data, for supporting academic institutions in decision-making regarding university admission [Men20] or the influence of students' performance during their years of college [AS19].

Every education provider, and more generally every service provider, tries to offer suitable products to its beneficiaries. In this regard, providers must have an appropriate image of the clients' performance, so that the offered products or services may be adapted according to these performances. Given the rapid evolution of society, the need of a paradigm shift in education is required. Thus, the educational systems must take into consideration the available instruments, so that this shift can bring benefits to students, instructors, and educational institutions. For instance, relating to the Covid-19 crisis, education shifted to *online environments*, and the traditional teaching methods used by the educational institutions needed to adapt. Learning success (in both traditional and online contexts) is influenced by students' motivation and the effectiveness of the teachers. The teaching quality does not guarantee the students' motivation or vice versa, because the latter depends on other factors, intrinsic or extrinsic [Nas20]. In this context, there is an increasing interest in understanding how students learn and how to improve their academic performance.

Approached Problem

Mining behavioural characteristics of students from school and academic settings is the key problem addressed in our work, because understanding how students develop and enhance their results is a topic of growing interest for every education provider.

The first concern is to identify *behavioural features of school students*, in order to help teachers in their instructional process. During a teacher's career (about 40 years long), there are different generations of students, so teachers must permanently adjust their methods of instruction to the current generation's profile, because each age has its motivation, goals, skills, etc., and students should be actively involved in learning. Mining the behavioural features of school students should give us a profile of nowadays students, with their interests and skills and thus, help instructors to adapt the teaching methods to their learners.

Mining concepts from Computer Science or another subject across curriculum levels should help institutions to design and adapt their curricula at the nowadays students, in order to facilitate the learning process. Nowadays students belong to Generation Z (those who are born between 1995 and 2010 [SG16], also named "digital natives"). The students of Generation Z are more technologically advanced and possibly more independent than learners of earlier generations [MCS19]. Although they are named "digital natives", they are not equipped in utilizing technological innovations for tactical purposes or build their careers. According to Shatto and Erwin [Sha16], the Zs have simple connection to streaming services, allowing them to study any subject whenever and wherever they want on a variety of gadgets. Their reliance on technology has a direct influence on their learning capability. If we mine the concepts needed to be learned for our society and take into account the profile of Zs, we can improve the instructional design, beginning with the school level and continuing with the college.

The second concern is related to the *behavioural features of students from academic level*. At this level, we aim to identify some tendencies while examining the students' performance both in traditional and online environments. We intend to use unsupervised learning techniques, to analyse and predict the students' performance and to compare the analysis' results in traditional and online environments. In the present context, when some teaching and evaluation activities, including lectures, assignments and examinations, are moved in online environments, there is an increasing interest in understanding the students' learning process, to improve their results. Nowadays, there is an accelerated development of these elements, as a consequence of several environmental factors. The Covid-19 epidemic, for example, changed every aspect of daily life, including schooling and *online learning* was a way for the educational providers' traditional teaching methods. The effectiveness of online learning is dependent on the standards of instruction and student engagement. This is a reason for us to compare the students' performance in traditional and online environments.

The third concern is to *develop a machine learning-based framework for students' performance analysis*. Such an instrument could help institutions of education in their decision-making processes. In this direction, we aim to introduce our framework and validate it on educational data sets from the academic level.

The challenge in our work is given by two major aspects: (1) the lack of open source data sets (we had to collect our own data sets and we noticed that people are very reserved when they have to share information about themselves, even if the answers are under anonymity), and (2) the difficulty to conduct comparison with the literature (the educational systems have important particularities, related to their country's politics).

Original Contributions

We have targeted our research on three major topics: (1) investigating behavioural features of school students; (2) mining behavioural features of students in academic environments; and (3) proposing a machine learning-based framework for students' performance analysis. For these directions, we focused on: *statistical analysis* (Chi-Square test, Wilcoxon signed-rank test, Pearson and Spearman correlation coefficients), *unsupervised learning* (UL) methods (k-means, principal component analysis, self-organizing maps, autoencoders, t-distributed stochastic neighbor embedding, uniform manifold approximation and projection for dimension reduction, and association rules), *supervised learning* (SL) methods (logistic regression, linear regression, linear discriminant analysis, polynomial regression, stochastic gradient descent, tweedie regression), and *feature selection* (ReliefF algorithm).

Therefore, Chapters 2, 3 and 4 provide our achievements and primary contributions in each of those aspects:

1. **Behavioural features of school students**

   We started our research from school environment, mining the behavioural features of students with age between 10 and 19 years. We proposed a statistical analysis to extract meaningful insights, *k-means* and *self-organizing maps* for clustering, and *association rules* to identify interesting relationships and dependencies between variables in our data sets. For this direction, the data sets were obtained by applying questionnaires to the targeted subjects. The results on this direction of research are as follows:

   (a) The first approach aimed to identify digital competences at 4th grade students, as behavioural features regarding the 21st century competences. Our proposed methodology and experimental results are described in the second section of the Chapter 2. The proposed approach has been published with the title "Digital and Coding Literacy for School Students" [Din18].

   (b) For mining behavioural features of school students from another level of study and age, we focused on high school students. We aimed to identify their preferences in choosing an exam item at the baccalaureate exam. The results are available at published in the article "Using Unsupervised Learning for Mining Behavioural Patterns from Data. A Case Study for the Baccalaureate Exam in Romania" [MCD23]. Our methodology and experimental results are presented in the third section of the Chapter 2.

   (c) Another perspective was to mine concepts from Computer Science (CS) across curriculum levels using a cyclic learning based approach. We targeted the *sorting* concept, with the aim to connect the knowledge and competences gained in secondary and high school with the requirements from the academic level. The study was presented and published with the title "Mining sorting concept across curriculum levels: a cyclic learning based approach" [McA22]. The proposed methodology and our results are available in the last section of the Chapter 2.

2. **Behavioural features of students in academic environments**

   The second direction for our thesis was to mine behavioural features of students from academic environments. We used the statistical analysis to observe meaningful correlations between features, *unsupervised learning* methods for students' performance

analysis, and *supervised learning* methods to reinforce the results obtained with the *unsupervised learning* approach. For this direction, the data sets used in the research were obtained from the students' activity at several academic disciplines at Babeş-Bolyai University Cluj-Napoca (BBU), Faculty of Mathematics and CS. Our results on this direction of research are as follows:

(a) Firstly, we wanted to mine academic data sets for students' performance analysis using unsupervised learning methods at the *Data Structures and Algorithms course*. The study was presented and published with the title "Unsupervised learning based mining of academic data sets for students' performance analysis" [CCCD20]. Our methodology and the results are presented in the first section of Chapter 3.

(b) Comparative assessment of academic performance in traditional and synchronous online learning environments was another aim of our research, resulting from the Covid-19 epidemic and the shift of instructional activities to online settings. Our work is presented in the second section of Chapter 3 and is disseminated through two studies related to the performance of students at *Logical and Functional Programming* course, which are presented bellow.

i. In order to compare synchronous online learning with traditional classroom instruction for evaluating students' academic achievement, *unsupervised learning* methods were used in the first approach. The study is available at [MCOM21].

ii. Comparing students' academic achievement in online and traditional learning contexts using *self-organizing maps* was the second approach. The study is available at [OMCM21].

3. **A machine learning-based framework for students' performance analysis**

The third direction in our research was to build a framework for students' performance analysis. This framework is named *IntelliDaM* and has three main components: (1) a component for feature analysis and selection; (2) a component for *unsupervised learning*-based data analysis; and (3) a component including *supervised learning*-based predictive models. To evaluate the performance of *IntelliDaM*, authentic data have been utilised. The data sets were obtained from BBU, during three academic years, for a CS discipline. The study is published with the title "*IntelliDaM*: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining" [CCML22]. The proposed methodology and experimental results are described in the Chapter 4.

Thesis Structure

The thesis is organized in the following way.

Chapter 1 discusses the theoretical foundation and literature review. It starts with the presentation of EDM domain, with students' performance analysis and prediction, and then is outlined the BM applied in the educational field. The second aspect from this chapter is the students' profile analysis in the context of digitisation, in order to check the 21st century competences, the digital literacy nowadays, gamification as an educational strategy, learning taxonomies used in nowadays education, and online learning and digitisation for a new paradigm shift. The third section treats the machine learning (ML) models used in our

thesis: *unsupervised learning* models (*k-means* clustering, principal component analysis, t-distributed stochastic neighbor embedding, uniform manifold approximation and projection, self-organizing maps, autoencoders, and association rules), *supervised learning* models used to reinforce the results of unsupervised based-analysis (linear discriminant analysis, linear regression, polynomial regression, logistic regression, stochastic gradient descent, and Tweedie regressor), and *feature selection*.

In Chapter 2, we present our results from mining school students' features in their learning processes. The first section focuses on the skills necessary for 21th century competences at 4th grade students in five schools from Romania during the second semester of the school year 2017-2018. The methods used are *statistical analysis* and *k-means* clustering. In the second section, we want to find some behavioural features at high school students, for identifying their preferences in choosing an exam item at the baccalaureate exam. In this sense, we used statistical and unsupervised learning-based analysis. The third section is an incursion in the Romanian curriculum for CS, for the purpose of observing the approach of the *sorting* concept at several stages of knowledge and to propose a framework useful for students in learning sorting algorithms. We were interested in secondary and high school levels and their impact at the college level, such that we can introduce a framework for collecting behavioural features from those who want to learn about sorting.

In Chapter 3 is presented our work related to mining the learning features of students at academic level. In the first section we used *Relational Association Rules* and *Principal Component Analysis* to analyse the performance of students at *Data Structures and Algorithms* course. The second section presents our comparative analysis of university students' performance in traditional versus online environment at *Logic and Functional Programming* course. In this regard, we used unsupervised learning techniques for the analysis such of *Autoencoders*, *T-distributed Stochastic Neighbor Embedding*, *Principal Component Analysis* and *Self Organizing Maps* and some supervised learning methods, such as *Logistic Regression*, *Linear Regression* and *Linear Discriminant Analysis* to validate our analysis.

Chapter 4 describes our proposed framework, named $IntelliDaM$, for students' performance analysis. Besides the proposed framework, the additional contributions envisaged by our research are: (1) to emphasise the effectiveness of $IntelliDaM$ on analysing students' performance related data; (2) to analyse and interpret, for the considered case study, the relevance of the patterns unsupervisedly mined from academic data and how these patterns are correlated with students' academic performance; and (3) to test whether or not the students' final performance prediction for a certain academic discipline is enhanced by incorporating in the prediction model their results achieved in previous CS courses from the curriculum.

After describing our work in these chapters, our conclusions and future work are presented, then the appendix and the list of the bibliographic resources.

Chapter 1

Background

This chapter presents the basic ideas relevant for the domain of our thesis, introduces the related work and the main computational models we are using in our research. The chapter is organised as follows.

The first section presents useful concepts from *Education Data Mining* (EDM), namely *Students performance analysis and prediction* and *Behavioural Mining applied in education.*

The second section addresses the issue of analysing the students' profile in our digitized society. It begins with a brief incursion on the *21st century competencies*, then it's presented the concept of *digital literacy* in our society. *Gamification* is described next, because it is more and more used in the instructional process nowadays. In what follows, principal *Learning Taxonomies* are presented, emphasizing on *Revised Bloom Taxonomy.* At the end of this section, we focus on *online learning and digitalisation.*

The last section describes the *Machine Learning* (ML) models helpful in our studies. It contains three parts: *unsupervised learning* (*k-means* clustering, principal component analysis, t-distributed stochastic neighbor embedding, uniform manifold approximation and projection, self-organizing maps, autoencoders, and association rules), *supervised learning* (linear discriminant analysis, linear regression, polynomial regression, logistic regression, stochastic gradient descent, and Tweedie regressor), and *feature selection* with its representative algorithms.

Chapter 2

Mining behavioural features of school students

This chapter is an incursion in behavioural features of school students, with the purpose to sketch the school student's learning profile. The chapter is structured as follows.

The first section presents the statistical analysis concepts utilised in our approaches: the *Chi-Square* test, *Z-score* and *F-value*.

In the second section, we look for the *21st century skills* at the 4th grade student's profile, through an educational project of "Dalia's Book" Association. Our results are published in the article "Digital and Coding Literacy for School Students" [Din18]. The purpose of the studt was, on one hand, to determine to what extent are the 4th grade students prepared to start the compulsory Computer Science (CS) classes in gymnasium, and, on the other hand, to find the learning profile of our students, identifying their competencies necessary for the challenges of our digitized society.

The third section continues the school *student's profiling* at high school level, for the purpose of obtaining the most important characteristics in choosing a subject for the baccalaureate exam in Romania. This section presents the study that was run for students from real sciences classes, i. e. specializations: *Mathematics-Computer Science*, *Mathematics-Computer Science intensive Computer Science*, and *Natural Sciences*. This study is published in the article "Using Unsupervised Learning for Mining Behavioural Patterns from Data. A Case Study for the Baccalaureate Exam in Romania" [MCD23]. One of the conclusions drawn is that high school students are very interested in choosing Biology and Computer Science for the baccalaureate exam.

After we have seen the great interest of high school students for CS in the third section, we propose to do some cross-sectional studies across curriculum levels for the most important concepts from CS. The last section presents our work regarding the approach of the *sorting concept across three curriculum levels*: secondary school, high school and university, with the main concern on the secondary and high school. This study follows a cyclic learning perspective and it was presented and published in under the title "Mining sorting concept across curriculum levels: a cyclic learning based approach" [McA22]. As an extension, we propose an application which will collect data from their behaviour as users and which will be useful in users profiling.

Chapter 3

Mining behavioural features of students in academic environments

Chapter 3 continues the mining of behavioural features for students from academic level. For this direction, we focus on students' performance prediction and analysis using *unsupervised learning*-based methods (Relational Association Rules, Principal Component Analysis, Autoencoders, t-SNEs, and Self-Organizing Maps). Some *supervised learning*-based methods (linear regression, logistic regression, linear discriminant analysis) were used for supporting the results obtained with unsupervised learning-based techniques. We used real data sets, collected from the Faculty of Mathematics and Computer Science, Babeș-Bolyai University, Romania.

The first section introduces the *relational association rules* and *principal component analysis* for students' performance analysis at *Data Structures and Algorithms* course. The study is entitled "Unsupervised learning based mining of academic data sets for students' performance analysis" [CCCD20].

The second section was inspired by the switch of the educational activities in online throughout the pandemic of Covid-19. In this section, we run two comparative analysis of students' performance in traditional versus synchronous online environments for *Logic and Functional Programming* course. These studies are:

- "Towards using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance" [MCOM21].
- "Using Self-Organizing Maps for Comparing Students' Academic Performance in Online and Traditional Learning Environments" [OMCM21].

Chapter 4

IntelliDaM: A machine learning-based framework for students performance analysis

With the goal of enhancing the results of students' performance analysis, we have introduced *IntelliDaM* [CCML22], an ML based framework for mining students' performance data. *IntelliDaM* offers three types of data analyses components designed for: (1) feature analysis and selection; (2) *unsupervised learning*-based data analysis; and (3) *supervised learning*-based predictive models. For evaluating the performance of *IntelliDaM*, we used authentic data obtained from Babeş-Bolyai University (BBU), Romania, during three academic years, for a CS discipline. Besides the proposed framework, the additional contributions envisaged by our research are: (1) to emphasise the effectiveness of *IntelliDaM* on analysing students' performance related data; (2) to analyse and interpret, for the considered case study, the relevance of the patterns unsupervisedly mined from academic data and how these patterns are correlated with students' academic performance; and (3) to test whether or not the students' final performance prediction for a certain academic discipline is enhanced by their results achieved in previous CS courses from the curriculum. Even if it is empirically evaluated on academic data, the proposed *IntelliDaM* framework is a general one, and it may be applied for any data analysis task.

The rest of the chapter is organised as follows. The first section discusses the methodology utilised for developing *IntelliDaM* framework and presents its main components. The second section presents the experiential assessment of our proposal of framework on an EDM case study, describing the data sets, the experiments and their results, whilst the third section describes the results obtained. The last section offers the conclusions of the research and some ideas for future work.

The obtained results of this study are presented in the article "*IntelliDaM*: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining" [CCML22].

Conclusions and future work

The present document describes the original results obtained for our PhD Thesis entitled "*Mining students' behavioural features in multiple learning environments*" with the goal to develop and implement DM techniques in issues belonging to the educational domain. Applying DM techniques in education [BCR18] is nowadays an interesting and active research domain in which the primary goal is building techniques to obtain relevant information from educational data in order to deeper comprehend students' learning processes and provide extra perspectives into education-related phenomena. The thesis described our work in this domain.

We have presented the problems which concerned us and our original contributions so far. Also, we outlined the directions we want to further continue our research in the EDM domain.

The first direction we targeted was the use of unsupervised learning and statistical methods for the analysis of nowadays school students' behaviour, to find which are their skills, weaknesses, or influencing factors in their instructional process. We want to correlate these results with the current situation from the educational system, to help instructors in adapting their teaching at present.

The second direction presented is the use of learning models at academic level, for students' performance prediction and students' performance analysis. The first concern was to mine the academic data set in order to identify some tendencies while examining the students' performance. The second concern was to compare students' performance in online and traditional learning environments, because one of the great challenges in latter years was the switch between traditional and online learning.

The third direction was to develop an ML-based framework for students' performance analysis, named $IntelliDaM$. This framework consists of components for feature analysis, unsupervised and supervised learning-based mining, useful for enhancing the performance of data mining tasks. Together with the first and the second directions, it will help education providers in their decision-making processes, for facing successfully the present challenges from education.

We consider these directions very important for the current situation in the educational field and we are confident they will offer valuable answers for anyone interested in EDM.

Future work will extend our research through developing new ML models with the target to uncover other meaningful patterns in our data sets. Another purpose is to enlarge our methods for collecting data and so, to increase our resources with data from students in different stages of learning. Also, we intend to consider more research instruments and to cooperate with specialists from psychology, pedagogy, or sociology who could offer their perspective in our research settings and interpretation.

Bibliography

[AS19] A. I. Adekitan and O. Salau. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5(2):e01250, 2019.

[BCR18] A. Bogarín, R. Cerezo, and C. Romero. A survey on educational process mining. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 8(1), 2018.

[CCCD20] L. M. Crivei, G. Czibula, G. Ciubotariu, and M. Dindelegan. Unsupervised learning based mining of academic data sets for students' performance analysis. In *IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI 2020)*, pages 11–16. IEEE Hungary Section, 2020.

[CCML22] Gabriela Czibula, George Ciubotariu, Mariana-Ioana Maier, and Hannelore-Inge Lisei. *IntelliDaM*: A machine learning based framework for enhancing the performance of decision-making processes. A case study for educational data mining. *IEEE Access*, 10:80651–80666, 2022.

[CcMS+20] C. Coman, L. G. Ţîru, L. Meseşan-Schmitz, C. Stanciu, and M. C. Bularca. Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective. *Sustainability*, 12(24), 2020.

[Din18] Mariana Dindelegan. Digital and coding literacy for school students. *Studia UBB Digitalia*, 63(1):55–68, 2018.

[FPDP+14] S. Ferenţ Pipaş, M. Dindelegan, B. Pădurean, E. Ciupan, and C. Ciupan. Cost calculator for water jet, laser and plasma machining. *ACTA TECHNICA NAPOCENSIS Series: Applied Mathematics and Mechanics*, 57:73–76, 3 2014.

[GSA21] Ram Gopal, Varsha Singh, and Arun Aggarwal. Impact of online classes on the satisfaction and performance of students during the pandemic period of covid 19. *Education and information technologies*, pages 1–25, 2021.

[JRHR15] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque, and Rashedur M. Rahman. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, 2(1):1, Mar 2015.

[McA22] Mariana Maier, Camelia Şerban, and Moisin Andrei. Mining sorting concept across curriculum levels. a cyclic learning based approach. pages 10–17, 2022.

[MCD23] M. Maier, G. Czibula, and L. Delean. Using unsupervised learning for mining behavioural patterns from data. a case study for the baccalaureate exam in romania. *Studies in Informatics and Control*, 2023.

[MCOM21] Mariana-Ioana Maier, Gabriela Czibula, and Zsuzsanna-Edit Onet-Marian. Towards using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance. *Mathematics, Engineering Mathematics - special issue on Didactics and Technology in Mathematical Education*, 9(22):2870, 2021.

[MCS19] J. B. Mosca, K. P. Curtis, and P. G. Savoth. New approaches to learning for generation Z. *Journal of Business Diversity*, 19(3), 2019.

[Men20] Hanan Abdullah Mengash. Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, 8:55462–55470, 2020.

[MT13] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97:320 – 324, 2013. The 9th International Conference on Cognitive Science.

[Nas20] S. minda Nasution. Online-Learning and Students' Motivation: A Research Study on the Effect of Online Learning on students' motivation in IAIN Padangsidimpuan. *Asian Social Science and Humanities Research Journal (ASHREJ)*, 2(2):9–16, 2020.

[OMCM21] Z. Onet-Marian, G. Czibula, and M. Maier. Using self-organizing maps for comparing students' academic performance in online and traditional learning environment. *Studies in Informatics and Control*, 30(4):1–11, 2021.

[SG16] Corey Seemiller and Meghan Grace. *Generation Z Goes to College*. Jossey-Bass; 1st edition, 2016.

[Sha16] K. Shatto, B.and Erwin. Moving on from millennials: Preparing for generation Z. *Journal of Continuing Education in Nursing*, 47(6):253–254, 2016.