ISSN 1870-4069

Gathering, Integration and Preprocessing of Data in Educational Data Mining: Towards a Study of Self-efficacy in Learning

Mayra Mendoza, Yasmín Hernández,
Javier Ortiz, Alicia Martínez, Hugo Estrada

Tecnológico Nacional de México, Cenidet,
Mexico

{m21ce017, yasmin.hp, javier.oh,
alicia.mr, hugo.ee}@cenidet.tecnm.mx

Abstract. Educational Data Mining emerged to take advantage of growing educational data and it has been extensively applied to improve Learning Environments, such as Intelligent Tutoring Systems. We are developing an intelligent tutoring system to teach mathematical logic, and we want to model self-efficacy in the modelling of students driven by data. We conducted a study to gather data from several sources. To have a precise learning model, we have conducted several visualization and preprocessing tasks to prepare data for machine learning algorithms, such as data integration, data exploratory analysis, data normalization, data standardization, resampling, management of outliers, discretization, and annotating. We built several versions of the dataset to prepare data for machine learning algorithms. The results of the preparation data process are presented.

Keywords: Data mining, educational data mining, intelligent tutoring systems, intelligent learning environments, preprocessing data, self-efficacy.

1 Introduction

Electronic devices in daily life generates cumulus of data every second. The analysis of this data allows us to obtain information about processes, people, relationships, behaviors, and about ourselves, which in turn allows us to make decisions and take actions. Data mining emerged to take advantage of growing data. Data mining seeks to discover patterns in large volumes of data, to extract information and to transform it into an understandable structure for later use [1].

The advancement of technology has made possible to store and process a huge amount of data, and it allows having successful data mining applications in different fields, for example, commerce, banking, and health. However, education is one of the fields where data mining has arisen the most interest and research. Education is one of

pp. 69–79; rec. 2022-09-16; acc. 2022-09-30

the fields that has benefited the most from computers, therefore there are an incredible volume of data on the interaction of students with learning environments.

This is result of the increasing use of learning environments, such as intelligent tutoring systems (ITS), e-learning systems, educational games, learning management systems (LMS) and massive open online courses (MOOC), in addition to administrative computer-based systems. With these data, we could know the students, understand different aspects of the interaction of the students with the systems and comprehend the learning process itself.

Educational Data Mining (EDM) is an emerging discipline, interested in the development of methods to explore the exceptional data that comes from educational environments, and concerned in the use of these methods to understand students and the environments in which they learn [2]. The research on intelligent tutoring systems has taken advantage of EDM. These educational programs simulate the behavior of human tutors.

Namely, ITS teach students in the same way that a human tutor does [3]. To have a more precise and adaptive behavior, several elements have been integrated to ITS behaviors such as the recognition of emotions and personality, as well as the modeling of different cognitive states such as motivation, self-efficacy, and self-regulated learning. These elements allow a more personal and motivating interaction with ITS.

As known, some subjects are perceived as difficult; thus, students are apathetic, do not do homework, and do not study in their account, and consequently they have a low achievement in their studies. We want to promote motivation and self-efficacy in students, and to provide them with tools for self-regulated learning. These constructs have been identified as important players in learning [4].

With this aim, we are developing MateLog, an ITS to teach mathematical logic. Towards including self-efficacy in the student model, we conducted a study to gather data about performance of the students and about personal traits such as self-efficacy and procrastination. We are going to analyze this educational dataset to know the relationship of self-efficacy and learning. Before to apply machine learning algorithms, we conducted an exploratory analysis and preprocessing of data.

In this paper, we present the study we conducted to gather data and the results of the exploratory analysis and preprocessing of data. The paper is organized as following: Section 2 presents background and a brief review of literature on educational data mining; Section 3 describes the study we conducted to gather data; Section 4 depicts the tasks for visualization and preprocessing data. Finally, Section 5 outlines conclusions and future work.

2 Background and Related Work

Educational data can be used to improve the understanding of learning, of students, and to create a better, smarter, more interactive, engaging, and effective education. This requires advances in artificial intelligence and machine learning, human intelligence understanding and learning theories [5].

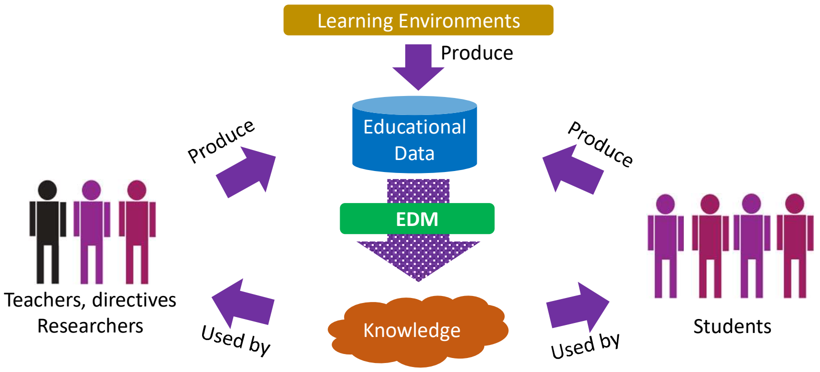


Fig. 1. Educational Data Mining knowledge discovery cycle [2].

Educational data mining is an emerging discipline that still has many pending solutions; but it has a potential to support the development of other fields related to education. EDM and, its technical basis, the machine learning techniques play an important role in augmenting and improving learning environments.

Machine learning is concerned with the ability of a system to acquire and integrate new knowledge through observations of users and with improving and extend itself by learning rather than by being programmed with knowledge [1]. These techniques organize existing knowledge and acquire new knowledge by intelligently recording and reasoning about data.

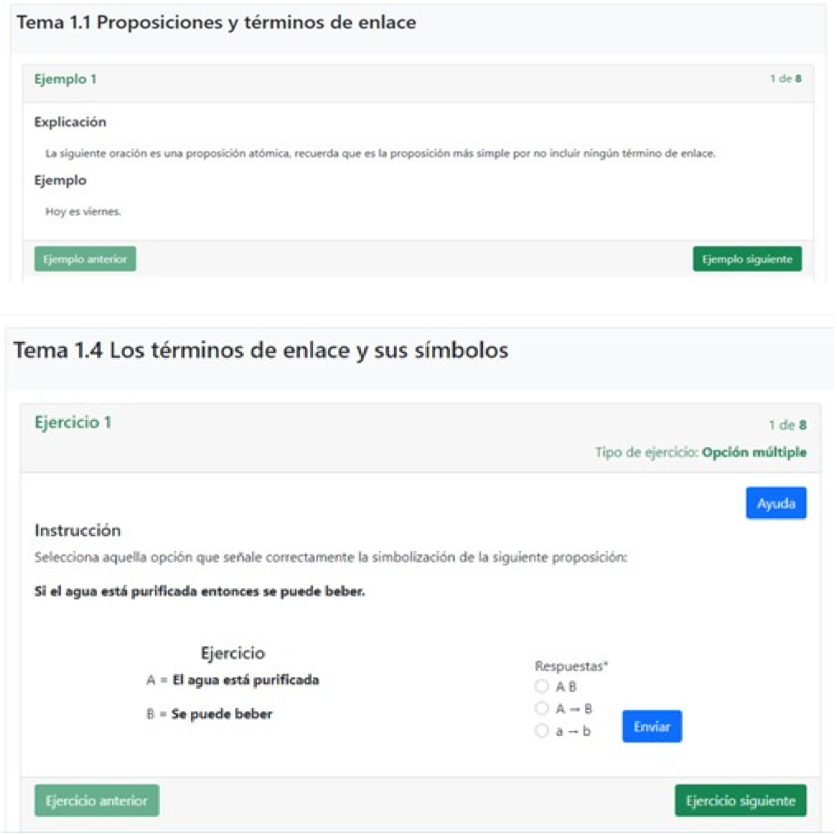
For example, observations of the previous behavior of students will be used to provide training examples that will form a model designed to predict future behavior. As in data mining, in EDM several computing paradigms and algorithms converge, such as decision trees, artificial neural networks, machine learning, Bayesian learning, logic programming, statistical algorithms, among others.

However, traditional mining algorithms need to consider the characteristics of the educational context to support instructional design and pedagogical decisions [2]. Educational data have meanings with multiple levels of hierarchy, which need to be determined by means of the properties of the data itself. Time, sequence, and context play an important role in the study of educational data [6].

EDM supports the development of research on many problems in education, since it not only allows to see the unique learning trajectories of individuals, but it also allows to build increasingly complex and sophisticated learning models [7].

The knowledge uncovered by EDM algorithms can be used not only to help teachers manage their classes, understand learning processes of their students, and reflect it in their own teaching methods, but also to support reflections of the student about the situation and give feedback to them [8]. Although one might think that there are only these two stakeholders in EDM, there are other groups of users, who see EDM from different points of view, according to their own objectives [2].

For example, education researchers, universities, course developers, training companies, school supervisors, school administrators, could also benefit from the knowledge generated by EDM [5]. Fig. 1 shows the interrelationships of educational environments, stakeholders and the EDM process.



Tema 1.1 Proposiciones y términos de enlace

Ejemplo 1 — 1 de **8**

Explicación

La siguiente oración es una proposición atómica, recuerda que es la proposición más simple por no incluir ningún término de enlace.

Ejemplo

Hoy es viernes.

Tema 1.4 Los términos de enlace y sus símbolos

Ejercicio 1 — 1 de **8**

Tipo de ejercicio: **Opción múltiple**

Instrucción

Selecciona aquella opción que señale correctamente la simbolización de la siguiente proposición:

Si el agua está purificada entonces se puede beber.

Ejercicio

A = **El agua está purificada**

B = **Se puede beber**

Respuestas*

○ A B
○ A → B
○ a → b

Fig. 2. Topics of mathematical logic included in MateLog, an example (left) and an exercise (right) are shown.

2.1 MateLog

MateLog is an intelligent tutoring system to learn mathematical logic for undergraduate and graduate students. MateLog consists of two main modules: the expert module and the student model. The expert module contains the knowledge about mathematical logic in the form of lessons, topics, examples, and exercises. The student model stores personal and academic information, and performance of the student in the course. In the Fig. 2 an example and an exercise are shown.

2.2 Self-Efficacy and Procrastination

Self-efficacy is a personal judgment of how well or poorly a person can cope with a given situation based on the skills they have and the circumstances they face. Self-efficacy affects every area of human endeavor. By determining the beliefs, a person holds regarding their power to affect situations, self-efficacy strongly influences both

Table 1. Extracts from scales for general self-efficacy, academic self-efficacy, and procrastination.

Scale	Items
General self-efficacy	I can find a way to get what I want, even if someone opposes me Puedo encontrar la forma de obtener lo que quiero, aunque alguien se me oponga
	I can solve difficult problems if I try hard enough Puedo resolver problemas difíciles si me esfuerzo lo suficiente
Academic self-efficacy	I consider myself capable enough to successfully face any academic task Me considero lo suficientemente capacitado para enfrentarme con éxito a cualquier tarea académica
	I think I have enough ability to understand a subject well and quickly Pienso que tengo bastante capacidad para comprender bien y con rapidez una materia
Procrastination	When I must do a task, I usually leave it until the last minute. Cuando tengo que hacer una tarea, normalmente la dejo para el último minuto
	I usually prepare in advance for exams Generalmente me preparo por adelantado para los exámenes

the power a person actually has to face challenges competently and the choices a person is most likely to make [9].

Also, self-efficacy has been recognized as a key trait of self-regulated learners [10]. Self-efficacy for self-regulated learning refers to the beliefs that individuals hold in their capabilities to think and behave in ways that are systematically oriented toward or associated with their learning goals.

Students with a robust sense of efficacy in their self-regulatory capabilities believe they can manage their time effectively, organize their work, minimize distractions, set goals for themselves, monitor their comprehension, ask for help when necessary, and maintain an effective work environment [11]. This action of avoiding, promising to do homework later, excusing or justifying delays, and avoiding guilt in the face of an academic task, refers to Academic Procrastination [12, 13].

In this situation, the student displays behaviors to voluntarily postpone activities that must be completed at a set time, either due to early family influence that has affected their self-esteem and tolerance to frustration; by the current choice of activities that guarantee an immediate achievement [13]; due to inadequate information processing [14] and of an irrational nature, or due to carrying out activities with more rewarding consequences in the short term than in the long term [13].

Evidence indicates that in the medium and long term, procrastination affects the academic life of people [15], since it is the first step towards other academic difficulties such as failures in the process of regulating academic behavior, Procrastination is related to less effective learning strategies which can affect the academic training and subsequently the professional performance of the person.

There are several instruments to know the self-efficacy and procrastination. A recognized scale for general self-efficacy is the proposed by Baessler y Schwarcer [16]. An extract of a Spanish scale for academic self-efficacy is proposed by Del

Table 2. Attributes in the dataset.

	Name	Description
1	Sex	Gender
2	Group	Group (A, B, C)
3	SP_Recoursing	Student is retaking the subject
4	SC_GeneralSelfEfficacy	Ten answers for each item of general self-efficacy scale
5	SC_AcademicSelfEfficacy	Ten answers for each item of academic self-efficacy scale
6	SC_Procrastination	Twelve answers for each item of procrastination scale
7	SP_Attendance	Attendances to class
8	SP_PretestGrade	Pretest exam grade
9	SP_Exercise1Grade	History of classical physics exercise grade
10	SP_Exercise2Grade	Unit conversion exercise grade
11	SP_Exercise3Grade	Introduction to vectors exercise grade
12	SP_Quiz1Grade	First term exam grade
13	SP_Unit1Grade	First term grade
14	SP_Quiz2Grade	Second term exam grade
15	SP_Unit2Grade	Second partial grade
16	SP_Exercise4Grade	Video exposition exercise grade
17	SP_Quiz3Grade	Third term exam grade
18	SP_FinalProject	Final project grade
19	SP_Unit3Grade	Third partial grade
20	SP_FinalGrade	Final grade in the subject.
21	ML_PreviousKnowledge	Previous knowledge about mathematical logic
22	ML_PreviousCourse	Student attended a previous course in mathematical logic
23	ML_CourseProgress	Progress in the mathematical logic course
24	ML_Topics	Topics completed
24	ML_Examples	Examples studied
28	ML_ExamplesSeconds	Time spent in examples
29	ML_Exercises	Exercises solved
30	ML_ExercisesSeconds	Time spent in exercises
31	ML_CorrectExercises	Correct exercises

Valle [17]. Dominguez [18] proposes an adaptation of the General and Academic Procrastination Scale proposed by Busko. Table 1 presents extracts of these scales, and its adaptation to Spanish, only two items are presented for each scale.

2.3 Related Work

The collection and preprocessing of data are an important part of the data mining process.

Educational data have unique characteristics that must be considered when preprocessing them. Research have been carried out to propose diverse methods for the treatment of educational data.

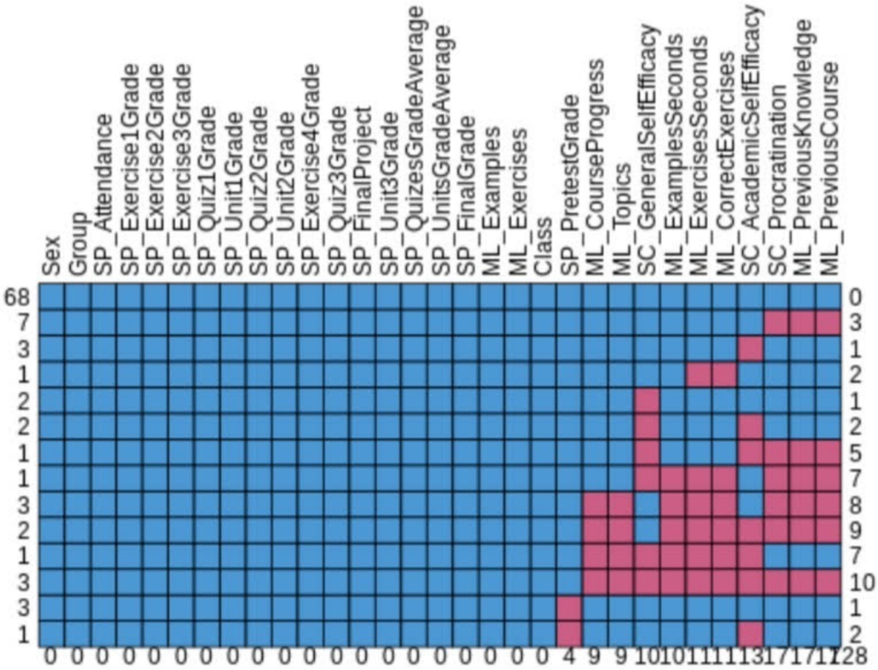


Fig. 3. Patterns of missing data in the dataset.

Romero, Ventura, and Romero [2] provides a survey of research related to the preprocessing of educational data. They state data pre-processing is one of the most important but less studied tasks in educational data mining research and shows the main tasks and issues in the pre-processing of educational data. They pay attention to types of data.

Romero, Ventura and García [3] expose the need to transform data from relational databases to transactional databases to have them in a format that can be read by data analysis tools and machine learning algorithms. To access this data stored in relational databases, it is necessary to use database queries, such as in SQL.

Subsequently, with the data obtained, a transactional summary table can be created. They also describe the process of collecting and preprocessing data from a course taught through the Moodle platform, which stores the data in a relational database (MySQL and PostgreSQL), to subsequently apply machine learning algorithms.

The stages they present are data selection, creation of summary tables, data discretization, and data transformation. They also present statistical data analysis tools. Chango and colleagues [5] describe the preprocessing and integration of data from different sources.

Tasks such as anonymization, normalization of attributes, discretization and data transformation are included. Hanna [6] describes a study about self-efficacy based on

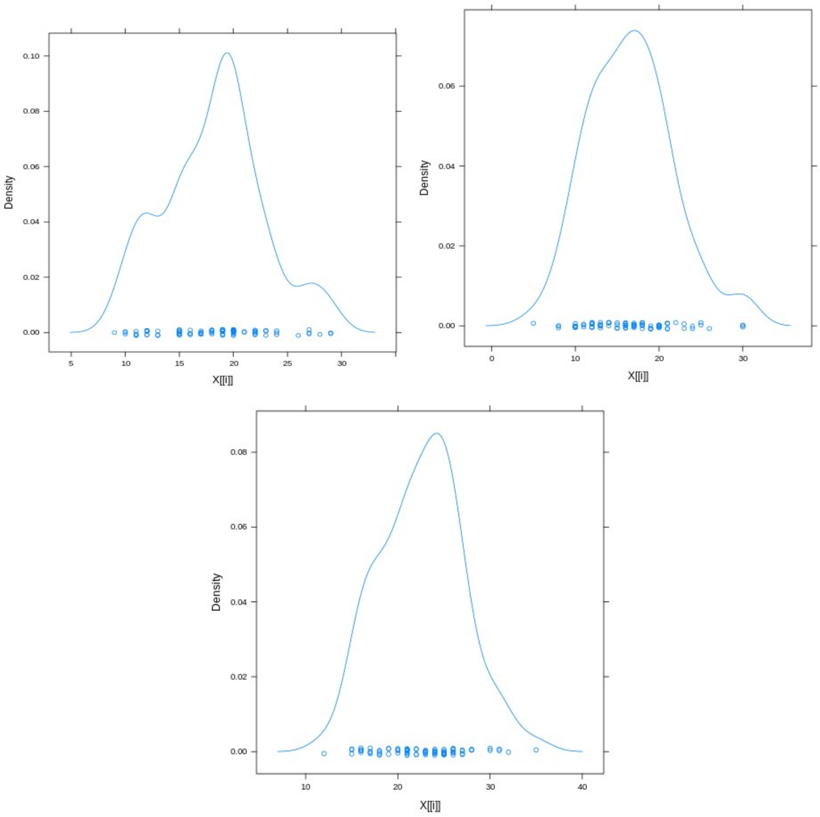


Fig. 4. Distribution of the scales of general self-efficacy (left), academic self-efficacy (center) and procrastination (right).

data mining to include this construct in the student model of an intelligent tutoring system. They state the feature and data selection is a key aspect of the educational data mining to understand the learning process.

3 Data Gathering

We conducted a study to gather data to understand the relationship between self-efficacy and learning. We analyze the academic performance of a group of undergraduate students group enrolled in a university course. Participants are 98 students, 25 female and 73 males, between 19 and 30 years old.

The study consists in three parts. Firstly, we obtained the data about the academic performance during the term, recorded by the professor. These data consist of attendances and grades in quizzes, exercises, and projects.

Then, we introduced MateLog to the students. They were asked to study mathematical logic with this ITS. The interaction of students with MateLog was

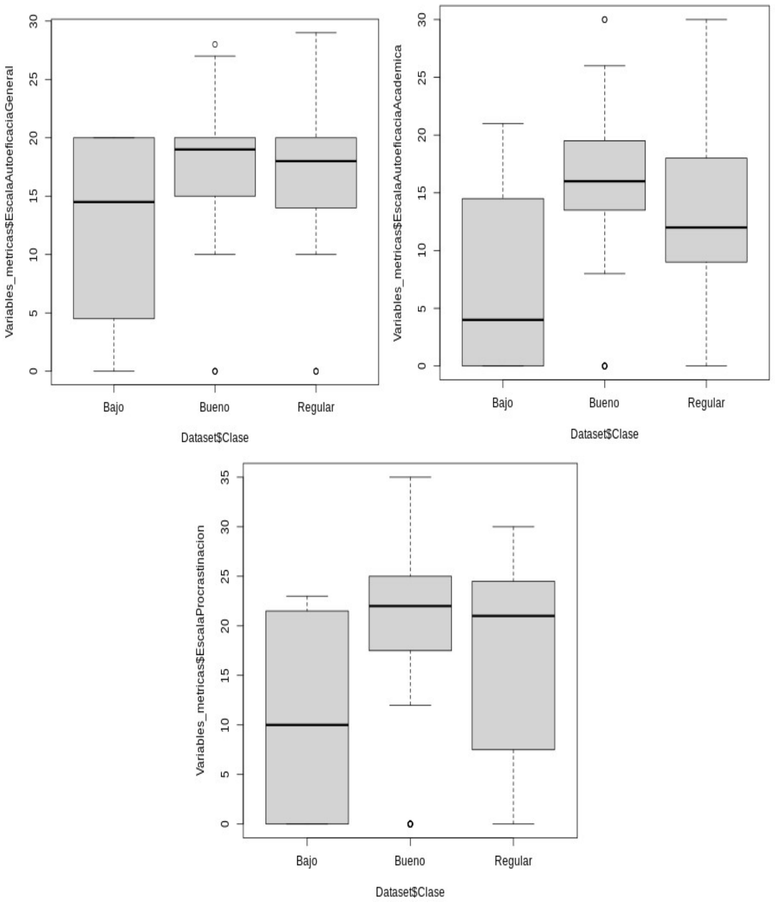


Fig. 5. Identification of outliers in the scales of general self-efficacy (left), academic self-efficacy (center) and procrastination (right).

recorded, these data consist of studied lessons and examples, correct and incorrect exercises. In the third part students filled in three scales about self-efficacy, academic self-efficacy, and procrastination. The scales are presented in section 2 (Table 1). We obtained a dataset with 98 observations. The description of attributes is presented in Table 2.

4 Data Visualization and Preprocessing

After the integration of the three sources of data, we got a dataset with 29 attributes and 98 examples.

The data were analyzed for identification of missing values. We find missing values in age and sex; these were filled looking in other databases. The data were analyzed to find missing data patterns. The result is shown in Fig. 3. There are 69 complete examples and 29 examples with missing data. There are 14 patterns with 1, 2, 3, 4, 6, 7, 8, 9 and 10 missing data.

Irrelevant attributes were eliminated, for example, the school grade since all participants were in the same grade. We build diverse versions of the dataset. In the case of scales answers, we have a version with all the individual answers to each item and other version with the result of the scale.

We analyze data to whether know data distribution. In Fig. 4 the distribution of general self-efficacy, academic self-efficacy, and general procrastination. As can be seen, they have a normal distribution. There are not classes in the dataset, therefore we annotated every example with based on the course grade (*low*, *regular*, *good*), the general self-efficacy (*low*, *medium*, *high*), academic self-efficacy (*low*, *medium*, *high*), and procrastination (*low*, *medium*, *high*).

In this way we have four versions of the dataset with different objective variables. To identify outliers, we used the box and whiskers chart. In Fig. 5 the comparison of scales variables versus performance is presented. As can be observed there are few outliers. After annotating, we have four unbalanced versions of the dataset, for example, for *performance* we have 8 *low* examples, 13 *regular* examples, and 57 *good* examples. Therefore, we must balance classes before to apply the classification algorithms.

5 Conclusions and Future Work

Educational data can be used to improve our understanding of learning and students, and which in turn allows having better educational technologies and a better education. These objectives require further advances in artificial intelligence and in human learning theories. Educational data mining is an emerging discipline that can be useful towards these aims due its potential to support the development of fields related to education.

In this paper, we present a study to gather data to model self-efficacy, to improve learning. The knowledge will be applied in the designing of MateLog, an ITS to teach mathematical logic. The ITS is building by means of applying several educational data mining techniques. The future steps include applying classification with machine learning algorithms to build prediction models, and we want to do clustering to find patterns and relationships in these data.

Acknowledgments. This research was supported by projects TecNM 13869.22-P and TecNM 14923.22-P.

References

1. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J.: Practical machine learning tools and techniques. Data Mining, vol. 2, no. 4 (2005)
2. Romero, C., Ventura, S.: Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, pp. e1355 (2020) doi: 10.1002/widm.1355

3. Woolf, B. P.: Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning (2010)
4. Lavasani, M. G., Mirhosseini, F. S., Hejazi, E., Davoodi, M.: The effect of self-regulation learning strategies training on the academic motivation and self-efficacy. Procedia-Social and Behavioral Sciences, vol. 29, pp. 627–632 (2011) doi: 10.1016/j.sbspro.2011.11.285
5. Koedinger, K. R., Brunskill, E., Baker, R. S., McLaughlin, E. A., Stamper, J.: New potentials for data-driven intelligent tutoring system development and optimization. AI Magazine, vol. 34, no. 3, pp. 27–41 (2013) doi: 10.1609/aimag.v34i3.2484
6. International Educational Data Mining Society: Educational Data Mining. http://www.educationaldatamining.org/
7. Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., Warschauer, M.: Mining big data in education: Affordances and challenges. Review of Research in Education, vol. 44, no. 1, pp. 130–160 (2020) doi: 10.3102/0091732X20903304
8. Merceron, A., Blikstein, P., Siemens, G.: Learning analytics: From big data to meaningful data. Journal of Learning Analytics, vol. 2, no. 3, pp. 4–8 (2016) doi: 10.18608/jla.2015.23.2
9. Bandura, A.: Guide for constructing self-efficacy scales. Self-efficacy beliefs of adolescents, vol. 5, no. 1, pp. 307–337 (2006) doi: 10.1017/CBO9781107415324.004
10. Panadero, E.: A review of self-regulated learning: Six models and four directions for research. Frontiers in psychology, vol. 8, p. 422 (2017) doi: 10.3389/fpsyg.2017.00422
11. Usher, E. L.: Self-efficacy for self-regulated learning. Encyclopedia of the Sciences of Learning (2012) doi: https: 10.1007/978-1-4419-1428-6_835
12. Onwuegbuzie, A. J: Academic procrastination and statistics anxiety. Assessment & Evaluation in Higher Education, vol. 29, no. 1, pp. 3–19 (2004) doi: 10.1080/0260293042000160384
13. Quant, D. M., Sánchez, A.: Procrastinación, procrastinación académica: Concepto e implicaciones. Revista vanguardia psicológica clínica teórica y práctica, vol. 3, no. 1, pp. 45–59 (2012)
14. Stainton, M., Lay, C. H., Flett, G. L.: Trait procrastinators and behavior/trait-specific cognitions. Journal of Social Behavior and Personality, vol. 15, no. 5, pp. 297–312 (2015)
15. Clariana-Muntada, M., i Pros, R. C., Badia-Martín, M. D. M., Gotzens-Busquets, C.: La influencia del género en variables de la personalidad que condicionan el aprendizaje: inteligencia emocional y procrastinación académica. Revista electrónica interuniversitaria de formación del profesorado, vol. 14, no. 3, pp. 87–96 (2011)
16. Bueno-Pacheco, A., Lima-Castro, S., Peña-Contreras, E., Cedillo-Quizhpe, C., Aguilar-Sizer, M.: Adaptación al español de la escala de autoeficacia general para su uso en el contexto ecuatoriano. Revista Iberoamericana de Diagnóstico y Evaluación-e Avaliação Psicológica, vol. 3, no. 48 pp. 5–17 (2018)
17. Del Valle, M., Díaz, A., Pérez, M. V., Vergara, J.: Análisis factorial confirmatorio escala autoeficacia percibida en situaciones académicas (EAPESA) en universitarios chilenos. Revista Iberoamericana de Diagnóstico y Evaluación-e Avaliação Psicológica, vol. 4, no. 41, pp. 97–106 (2018)
18. Dominguez, S., Villegas, G., Centeno, S.: Academic procrastination: validation of a scale in a sample of students from a private university. vol. 20, no. 2, pp. 293–304 (2014)