Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions

Emilcy Hernández-Leal a, b, Néstor Darío Duque-Méndez a, *, Cristian Cechinel c

a Universidad Nacional de Colombia, Colombia
b Universidad de Medellín, Colombia
c Universidade Federal de Santa Catarina, Brasil

HIGHLIGHTS

The first analysis of performance from Elementary and Public High School students in Colombia at a regional level.
A revisited and up-to-date list of challenges for the implementation of large scale Learning Analytics initiatives.
Results of this work may be used to propose policies towards the integration of data coming from different institutional levels.

ARTICLE INFO

Keywords: Educational data
Learning Analytics
Primary education
Secondary education

ABSTRACT

Even though the field of Learning Analytics (LA) has experienced an expressive growth in the last few years. The vast majority of the works found in literature are usually focusing on experimentation of techniques and methods over datasets restricted to a given discipline, course, or institution and are still few works manipulating region and countrywide datasets. This may be since the implementation of LA in national or regional scope and using data from governments and institutions poses many challenges that may threaten the success of such initiatives, including the same availability of data. The present article describes the experience of LA in Latin America using governmental data from Elementary and Middle Schools of the State of Norte de Santander - Colombia. This study is focusing on students' performance. Data from 2013 to 2018 was collected, containing information related to 1) students’ enrollment in school disciplines provided by Regional Education Secretary, 2) students qualifications provided by educational institutions, and 3) students qualifications provided by the national agency for education evaluation. The methodology followed includes a process of cleaning and integration of the data, subsequently a descriptive and visualization analysis is made and some educational data mining techniques are used (decision trees and clustering) for the modeling and extraction of some educational patterns. A total of eight patterns of interest are extracted. In addition to the decision trees, a feature ranking analysis was performed using xgboost and to facilitate the visual representation of the clusters, t-SNE and self-organized maps (SOM) were applied as result projection techniques. Finally, this paper compares the main challenges mentioned by the literature according to the Colombian experience and proposes an up-to-date list of challenges and solutions that can be used as a baseline for future works in this area and aligned with the Latin American context and reality.

1. Introduction

Education in Latin American countries has shortcomings and is far from being able to reach the levels of developed countries (Ducoint, 2019). The 2018 World Development Report indicates that there are three dimensions to the learning crisis: the first highlights how unsatisfactory learning outcomes are with low levels, high inequality and slow progress, the second refers to immediate causes, highlighting how schools are failing students, poorly prepared students, unmotivated teachers, deficiencies in school management and in school supplies. Finally, the third-dimension deals with the root causes, in this case, the educational system is the one that is failing schools with limited management capacity, technical and political difficulties (World Bank, 2018). Initiatives promoted by UNESCO such as Media and Information Literacy.
(MIL) places great value on data and information to work optimally in organizations. In the midst of this context, educational institutions recognize the importance of making use of the data through analysis that allow them to understand what is happening in the educational process. There are still large gaps to achieve considerable growth in the number of Learning Analytics (LA) implementations, mainly due to a lack of guidance on how to coordinate the interaction between policy formulation and implementation (Broos et al., 2020). This phenomenon occurs in higher education (Hilliger et al., 2020), in which a change has begun to be seen, but in K12 education the lack of reporting initiatives is even greater. The lack of initiatives is the origin of the greatest problems of the Latin American educational systems with alarming figures of illiteracy and causes such as extreme poverty (Duscoing, 2019).

In Colombia, educational policies are formulated at the national and regional level. At the regional level, the development plans are based on the formulation of educational public policies. In the department of Norte de Santander —object of this study—, it was concluded, from the analysis of the development plan of 2016, that the information for decision-making in the educational sector becomes outdated and the diagnostic analyses are short and shallow. Consequently, there is no data that allows inferring the needs of educational institutions and this limits the action of the governmental department (Aguilar Barreto et al., 2018). To this extent, there is a call for increasing the use of data to generate knowledge, but there are still few studies on educational data in Colombia, mainly at levels other than higher education.

Different methods are currently being used to analyze educational data in the field of LA, such as: Structural Equation Modelling (SEM), Data Mining (DM), and Social Network Analysis (SNA). Learning Analytics and SNA can complement other research methods to analyze knowledge construction in online interactions. SNA characterizes the information infrastructure that supports the construction of knowledge in social contexts, and its combination with LA can be an alternative to other traditional approaches (Gunawardena et al., 2016). On the other hand, the SEM has been used in studies of analysis of student satisfaction in online courses (Kucuk and Richardson, 2019) as well as their level of participation (Kog, 2017). At last DM has a close relationship with the LA and works with the prediction of school dropout, academic performance, and interactions on virtual platforms, among others (Romero and Ventura, 2013).

Learning Analytics tools to support data analysis in educational institutions have gained strength mainly in higher education (Avella et al., 2016; Kasemsap, 2016; Leitner et al., 2017; Tsai and Gasevic, 2017; Viberg et al., 2018). Most of the studies have been conducted by researchers from Computing and ICT (Information Communication and Technologies) departments (Slater et al., 2016). Moreover, reported works are usually concentrated in particular courses, academic programs, and virtual learning platforms (Firat, 2016; Martin and Whitmer, 2016; Sahar et al., 2016), and very few deal with regional or national data (Macarini et al., 2019a).

Applications in the LA field differ in many ways, as they may be focused on different stakeholders (students, teachers and administrators), or use different techniques (DM, SNA, visualization, statistics, machine learning and artificial intelligence) (Hoppe, 2017). LA also tackles different problems associated to the field (Lawn, 2013), such as: school performance, dropout, coverage, educational quality, assessment methods, learning styles (Conde and Hernández-Garcia, 2013; Ferguson, 2012; Hu et al., 2017; Mangoaras and Giannakos, 2016). Moreover, LA deals with workspaces in the management and integration of different information systems and data sources of a variety of types, scales and granularity levels (Osila et al., 2015). A parallel and complementary field of LA is the so-called Educational Data Mining (EDM) (Dutt et al., 2017; Romero and Ventura, 2007) that seeks to evaluate and extract meaningful patterns from the data produced and stored inside educational contexts in order to better understand problems, and generate strategies at a pedagogical, curricular and institutional level (Romero and Ventura, 2010, 2013; Scheuer and McLaren, 2012).

The present work describes the results of an effort focused on unveiling educational patterns of students of a Colombian region. For that, academic performance and other students’ information coming from different databases were integrated and used in a series of visualization and mining experiments. Precisely, data of 6400 students —extracted from three different sources—from the Norte de Santander Department were collected, preprocessed, cleaned, mined, and analyzed. Initiatives like this one, which uses regionwide (or countrywide) educational data that transcends a particular group or course to a much broader population are scarce (Ruiz-Calleja et al., 2019), especially in Latin America (Cechinel et al., 2020). To the best of our knowledge, the only work in Latin America that used a large data set from secondary education to conduct a nationwide LA initiative is the work of Macarini et al. (2019a). This work intends to unveil educational patterns found during the experiments, but also to contrast the difficulties and challenges faced during the execution of the project with previous literature. The present work proposes to answer the following research questions:

RQ1 – Which patterns and findings emerge from exploratory analysis and data mining techniques in primary and secondary Colombian educational institutions?

RQ2 – Does the analysis of data of different scales and levels of granularity (institutional + regional + national) show different knowledge about the performance of students in the Colombian educational system?

RQ3 - Which are the challenges in the design and execution of a regional level Learning Analytics experience in Colombia and how do they differ from challenges encountered in the state of the art in Latin America?

The paper is organized as follows. Section 2 presents the state of the art in Latin America and describes the functioning of the Colombian educational system. Then, section 3 presents materials and methods. After, section 4 describes the experiments and results, section 5 the discussion of the results and challenges encountered. Finally, section 6 concludes the paper and proposes future work.

2. Literature review

This section describes the state of the art of learning analytics and EDM in Latin America and gives a context of the Colombian Educational System.

2.1. State of the art in Latin America

Du et al. (2019) analyzed more than 900 papers in the field of LA and described that most of the works are supported by traditional tools such as statistics and data visualization. Moreover, the main topics covered were focused on modeling the performance and behavior of students as well as giving feedback to the different stakeholders in place. This feedback was given to both teachers and students, as well as at the directive level. Depending on the object of the study, cases such as the detection of behavior patterns given to teachers, or the recommendation of learning objects for students are some examples. The authors also raised a significant lack of studies using data from K-12 education and pointed to the lack of data availability as one of the major causes. According to the authors, it is common that K-12 environments do not have enough resources or staff to systematically store and maintain data required for the implementation of LA initiatives.

Even though Latin America already has a number of initiatives in the field of LA (dos Santos et al., 2017), they are also mainly concentrated on the use of data coming from the tertiary level (Cechinel et al., 2020; Hilliger et al., 2020). Works that include data from primary and secondary level are very few in Latin America, as well as the works including the integration of different granularities of data sources coming from
different institutions. This situation is corroborated by a number of regional authors. Moreno Cadavid and Pineda Corcho (2018) reported an increase in the number of works in Latin America starting from 2014 and highlighted that the goals of the papers are concentrated on learning management systems data and surveys in the context of the university courses and specific areas like programming and mathematics. Moreover, Cala Wilches and Grisales-Palacio (2019) analyzed the works exclusively developed in Colombia and found that the vast majority of the documents analyzed came from engineering faculties.

So far, and to the best of our knowledge, there is only one work developed with large-governmental data from secondary schools in Latin America: the work of Macarini et al. (2019a), where the authors developed a system prototype to follow the academic trajectory in Uruguay using K-12 data. The authors conducted experiments using clustering algorithms and association rules. The challenges found during that project were grouped into eight different dimensions: 1) nuances of the education system, 2) ethical and legal requirements, 3) access to data, 4) inconsistencies and integration the database, 5) time restriction, 6) scope versus agency needs, 7) selection of algorithms and tools, and 8) transfer of the results obtained. Among the educational findings of the initiative, one can mention the following ones: first, the behavior of the approved and reproved students was analyzed, taking into account when they were below a certain threshold and considering grades in specific subjects. Second, the search for the most problematic subjects found in mathematics and Spanish (maternal language) the two main focus of school failure.

Still covering works using data at national level, there is the paper of Ruiz-Calleja et al. (2019), where the authors analyzed six case studies carried out in Uruguay and Estonia. In Uruguay: Impact of a new device, English teaching practice and use of an Adaptive Math Platform. In Estonia: ICT students drop-out, Digital Mirror and School performance indicators. From the review of these cases, the authors concluded that in Estonia the studies promote educational innovation and the competitiveness of educational institutions.

The state of the art in Latin America shows the large concentration of works dealing with data from tertiary education and corroborates the importance and relevance of carrying out initiatives towards the development of large-scale LA projects using data from different educational levels.

2.2. Colombian education system

Colombian education system is organized in five different levels (see Figure 1). The present work is concentrated on the Elementary School, which is structured in two cycles: primary (1st to 5th grades) and secondary (6th to 9th grades), and on Middle School (10th and 11th grades).

These different levels are coordinated locally, regionally and nationally by three different actors (see Figure 2): The Ministry of National Education, the State/District Secretariats, and the Educational Institutions. This section briefly explains the role each actor plays and which kind of information they store.

The Ministry of National Education is the department that leads education in Colombia and that defines the educational policies and purposes of the country. Education policies aim to increase the number of enrolled students at all levels and regions. For that, several crucial challenges must be faced such as closing gaps in terms of participation and educational quality. Disadvantaged children coming from low-income families have educational inequalities since they do not begin school at the proper age or they normally attend lower-quality educational institutions. For the population living in poverty the school life expectancy is only six years, while the ones with the best economic conditions reach 12 years. In higher education only 9% of the poorest enrols, while in wealthier people, this rate rises to 53% (Jiménez Angel et al., 2013).

At the national level, there is a state company linked to the Ministry of National Education, the Colombian Institute for the Evaluation of Education (ICFES), that assesses education at all levels. These exams are called SABER (i.e., "to know" in Spanish). For the primary and intermediate levels, there are four SABER exams (at 3, 5, 9 and 11 grades) with the objective of reviewing the knowledge of the students and their competencies in different topics (logic reasoning, language, natural sciences, math, among others) (ICFES, 2019).

According to SABER evaluations, the socioeconomic status and the educational background of the parents have a strong effect on the achievements of Colombian students (Delgado Barrera, 2014). The results also show that Colombian students continue presenting low-performances. In the reading and writing exams of 2014 it was found that 49% of the students in the third grade, 67% in the fifth grade and 73% in the ninth grade were not complying with the minimum standards. Besides, in SABER 11 it was found that by 2013 27% of the students were not well-prepared to enter into higher education. This causes severe implications to the continuity and success of their studies during higher education (Organización para la Cooperación y el Desarrollo Económicos (OCDE), 2016).

The State/District secretariats are the governmental departments responsible to formulate, supervise and coordinate sectorial educational policies at a regional level. The Secretariats are also responsible to control the supply of the educational services in the states and foster research focused on curriculum formulation and teaching methods. Moreover, they are also in charge of managing the information systems and generating datasets at the academic environments from Early Childhood up to the Middle School. These departments are responsible for maintaining information related to student's enrollment, and also for the infrastructure of the databases.

Educational Institutions (EI) are the schools that provide educational services and are in charge of all teaching processes. EIs are responsible for education planning, assuring quality of the services, and students' enrollments. At last, EIs are responsible for storing data at a lower detail level such as the records of student's grades in the subjects.

At the institutional level there is an aspect that is important to highlight in this study and it is the subject of evaluation and promotion. With the issuance of Decree 1290 of 2009, the national government in
Colombia granted the faculty to EIs to define the Institutional System for Student Assessment (SIEE in Spanish), this being a task that requires study, reflection, analysis, negotiations and agreements between the entire educational community. Evaluation is not an isolated task of the training process; therefore, it must be linked and coherent (conceptually, pedagogically and didactically) with the entire educational proposal that the Educational Institution has defined. The evaluation should be aligned with the mission, purposes, model or pedagogical approach. Such activity implies that at the time of designing the SIEE, it must be articulated with the Institutional Educational Plan (PEI in Spanish), not only because of its incorporation into it, but also because of the correspondence that must exist between the teaching approach and the evaluation approach. The evaluation criteria are the rules to verify if a student reached the expected level of performance in a learning area. Promotion criteria are the rules by which students are promoted to the next school grade, these may be different for each grade or educational level, for example, the promotion criteria for 1st grade may be different from those of the other grades of primary basic education, and likewise, the secondary promotion criteria may be different from those of primary. This clarification is made given that several of the findings in terms of educational patterns are associated with the passing/failing of the school year.

3. Materials and methods

Data mining is a method of knowledge discovery in databases—KDD—, it follows a series of steps common to all analysis processes, and independent of the field of study or data type studied. According to the above and in view of the lack of concrete methodologies

![Diagram](image-url)
for the application of LA, in this work we used a methodology that covers the general and iterative steps of the KDD process. The methodology approaches interactive data mining (Hübscher et al., 2007), which is defined as an interaction between a computer and a researcher where both collaborate mutually to find the connections between the records and the student's behavior. In the methodology, there were seven general steps, some of which had to be carried more than once (Figure 3).

Step 1 consisted of collecting the data and was explained in detail in section 4. In step 2, considerable work was required, an ETL module was built oriented to the data of the educational domain, for this, three phases were followed: the data understanding phase, the extraction and filtering phase, and the transformation phase. In step 3, data was loaded and stored into an integrated relational model database. This model was built in a PostgreSQL database engine and consisted of five tables, which correspond to student, academic data, socioeconomic data, institution, and center. Steps 4 (dataset extraction) and 5 (application of algorithms) were carried out by hand, the dataset was extracted according to the requirements of the algorithms and the relationship or pattern that was looked for inside the data.

Considering the RQ1, a series of experiments were made, which aimed to answer the question about the role of the integration of scales and granularity levels in the knowledge to be discovered in the data of the educational system. Analysis of results and pattern identification was carried out as the experiments were executed, so it was iterated in the same way. Steps 4, 5, and 6 are expanded in the next section. Step 7, visualizations of results, was performed both for the results of the association and clustering algorithms and for the phase of descriptive analysis of data.

3.1. Data collection description

Figure 4 gives a panoramic overview of the work developed here. The present section focuses on presenting the characteristics of the data sources, their capture and processing, as well as the problems encountered in this process.

3.1.1. Databases description

Data was collected from three different sources and then integrated into a single database. Figure 5 depicts these three sources, their sectors of origin, and some of the information contained inside them.

The first data source—DB1—corresponds to a total of 22,000 records from 6,400 students from four EI of Norte de Santander Department in Colombia (see Table 1). Besides, the students’ final grades in the different subjects (yearly average achieved), the data also contained the final status of the students (approved or reproved) together with its educational level, center and institution. In Colombia, at the end of the school year, students receive the status of approved when they reach the necessary qualifications and competences to attend the next grade and the status of reproved when they do not meet the requirements to reach the next grade; therefore, they must retake the current grade. The data was supplied directly by each EI in .pdf files and for the periods from 2014 to 2018. As shown in Table 1, institutions have very different numbers of students, due to their location. While EI2 and EI4 are institutions with high rural presence, EI1 and EI3 have a greater presence of students from urban areas. It is worth clarifying that there are around five records for each student, one for each school year. Each record corresponds to the set of final grades for each of the subjects, the final status and the data of grade, course, course session, and center. The term register denotes the vector with the values of attributes mentioned for each student per school year.

The second data source—DB2—was supplied by the Secretariat of Education of the Norte de Santander and contains five categories of information: educational institution, student identification, geographic location, socioeconomic and academic status. Table 2 shows attributes per year, records for the four EIs and records for all public EIs in the region.

From this dataset we selected only those records related to the 6,400 students present in DB1.

Finally, the third database—DB3—was downloaded from the open data repository of the ICFES1 and contained the results of the evaluation exams of educational quality that are performed by the students in grades 3, 5, 9 and 11. In this dataset, the results of all the schools in the country were available, but we selected only data from the four EI included in our study and for SABER 11 (see Table 3). That is, the data analyzed was that corresponding to the test SABER presented by students of 11.

DB1 and DB2 were integrated into a single database with a relational model design that was used in the experiments. All data was anonymized in order to guarantee students’ privacy. DB3 was analyzed separately, since for the SABER 3, 5 and 9 exams the data is represented by municipalities.

3.1.2. Problems experienced in processing data to be analyzed

Data collection can be considered the most difficult and complex stage of the present work, mostly due to the fact that data were found in different sources, formats, and were located in different institutions. Contacting each institution required a huge effort from the researchers, especially in the case of the EI that provided DB1. To collect DB2, the Secretariat of Education was contacted through the project that supports

---

1 https://www.icfes.gov.co/web/guest/investigadores-y-estudiantes-posgrad o/acceso-a-bases-de-datos.
the development of this research. From the Secretariat the approach to the EI was made, through a letter of introduction and with the accompaniment of officials from that department. Even though an initial contact with the school directors was possible, it was quite difficult to keep in touch due to their busy agenda. This situation and the time taken to attend the requests led us to work with only four EI.

Several challenges were faced: The State/Regional Secretariats do not have a common and standardized structure for data management and the storage of academic performances. There are cases where the EIs outsource the management of their data to private companies as each school director has autonomy to contract the company which stored and managed their data. This lack of a centralized storage structure prevents them to both store and generate value from the data; therefore, making it almost impossible to make informed decisions based on the data, or building institutional improvement policies based on past and present patterns. Moreover, the system used by EI considered in this study only allows them to generate and to export reports in PDF files. This situation required extra effort in the pre-processing stage, as the PDF files had to be converted to an editable format (CSV).

Access to DB3 was relatively easy as the data was open and available for researchers to use through a web portal.

4. Experiments and results

Experimentation was done in three moments. First, a descriptive analysis was carried out that took the DB1 and DB2. We performed the descriptive and exploratory analysis of the data using the Tableau tool2, which allowed us to have a general overview of the data and to obtain some first hints about student’s behavior. Tableau is a tool that seeks to help see and understand the data through the presentation of dynamic visualizations.

The second moment was the application of decision tree algorithms to find educational patterns. For this, the Orange3 and Weka4 tools were used. The datasets used were DB1, DB2, then a combination of DB1 + DB2 and finally DB3. Finally, in the third moment, we try to find patterns through Clustering algorithms using the RapidMiner5 tool and DB3.

4.1. Descriptive analysis

To start the descriptive analysis, some statistics of the grades contained in the DB1 were generated including the four EI, in terms of the measures of central tendency (mean, median and mode) and position (variance, standard deviation and quartiles) the main subjects of primary (see Table 4) and secondary (see Table 5) were selected. In primary school, 95.26% approved and 4.74% reproved. However, in secondary the reproved rate was higher, reaching 13.61% compared to an approval of 86.39%. It should be clarified that the range of grades goes from 0 to 5 with an average passing grade of 3.

Table 1. Amount of DB1 data per educational institution.

| Educational Institution (EI) | Location | Years       | Students quantity (approx.) | Records quantity (approx.) |
|------------------------------|----------|-------------|----------------------------|---------------------------|
| EI 1 Urban                   |          | 2014-2018   | 2,300                      | 11,500                    |
| EI 2 Rural                   |          | 2014-2018   | 600                        | 3,000                     |
| EI 3 Urban                   |          | 2014-2018   | 3,000                      | 15,000                    |
| EI 4 Rural                   |          | 2014-2018   | 500                        | 2,500                     |
| Total                        |          |             | 6,400                      | 32,000                    |

Table 2. Amount of DB2 data per year.

| Year | Attributes quantity | Records for EI 1, 2, 3 and 4 | Total records quantity for the region |
|------|---------------------|------------------------------|--------------------------------------|
| 2014 | 55                  | 6,161                        | 146,193                              |
| 2015 | 58                  | 5,994                        | 145,196                              |
| 2016 | 55                  | 6,028                        | 143,195                              |
| 2017 | 57                  | 6,027                        | 145,938                              |
| 2018 | 60                  | 6,104                        | 148,527                              |

Table 3. DB3 data by EI (total registers for SABER 11 of 2018).

| EI   | 2014 | 2015 | 2016 | 2017 | 2018 |
|------|------|------|------|------|------|
| EI1  | 120  | 145  | 131  | 121  | 115  |
| EI2  | 30   | 14   | 21   | 21   | 24   |
| EI3  | 160  | 167  | 174  | 135  | 175  |
| EI4  | 23   | 23   | 22   | 22   | 36   |
| TOTAL| 333  | 349  | 348  | 299  | 350  |

2 https://www.tableau.com/.
3 https://orange.biolab.si/.
4 https://www.cs.waikato.ac.nz/ml/weka/.
5 https://rapidminer.com/.
When analyzing the classification of approved/reproved students, it was common for the four EIs that one of the groups with the highest percentage of failure were the males of the sixth grade (Figure 6). A possible explanation for this pattern is the fact that sixth grade is the transition of students from primary to secondary level and that the students ages are between 10 and 12 years, i.e., the period of change from childhood to puberty.

Educational Pattern 1 (EP1): The grade of transition from primary to secondary level (6th grade) concentrates the highest percentage of reproved students, most of them males.

Another finding, to be highlighted here, is the distribution of the students in the socioeconomic classes and their relation with the state of approved/reproved. In Colombia, families are stratiﬁed in levels ranging from 0 to 6 according to a series of economic conditions evaluated by the National Administrative Department of Statistics (DANE). Levels 0 to 2 correspond to the neediest and low-income families; level 3 is the middle level and from level 4 to 6 are the families with better economic conditions. For the data studied, the students of the four EIs are distributed from social level 0 to 3, this can be justified because the EIs part of this study are all public free schools and the families of higher social level usually enroll their children in private institutions. Figure 7 shows that there is a higher number of reproved among students in social level 1.

Educational Pattern 2 (EP2): Most of the students from public schools of the region of Norte Santander Colombia come from socioeconomic level 1 families. The proportion of students from levels 4, 5 and 6 in public schools is very low; also, few come from level 3.

Educational Pattern 3 (EP3): There is a high number of reproved students who come from families of socioeconomic level 1; however, in terms of percentage, reproved has a similar behavior for all social levels.

Finally, after analyzing the approved/reproved status in relation to the final average grade in each of the subjects, we found an important behavior related to the final average grades in mathematics and language. As shown in Figure 8, students with an average grade fewer than 3 ((on a scale of 0-5)) in mathematics and language, that is, the ones who reprove on these subjects, also tend to reprove the year (marked in red).
This behavior is regular in all grades and for both genders (represented on the x-axis in Figure 8). It is important to mention the average for approval is 3 or higher (on a scale of 0–5), but this may slightly vary in each EI.

After checking the distribution of the data (they followed a normal distribution), a t-test was applied to evaluate whether there was a difference in the means of the performances of both genders in both disciplines (language and mathematics) for EI1. Considering a level of
significance of 5%, both tests indicated that there is a statistically
difference between the means of the performances of both genders and in
both disciplines. The t-test was also applied without splitting the sample
per discipline, and a statistically significant difference between the per-
formances of both genders was also found. In all tested scenarios, the
female gender presented better performances. To complement this
analysis, we also performed an analysis of variance ANOVA of two factors
with unbalanced data and we found statistically significant differences
between the grades of both genders and for both disciplines (separately
and in the aggregate). These analyzes were corroborated both by the
corresponding statistic and by p-value. It should be clarified that the test
shows the significant difference between the qualifications of the two
 genders, however, we do not affirm that the observation, as a pattern, can
already be considered significant, for this a more in-depth analysis is
required, this study concentrated on a exploratory analysis.

Educational Pattern 4 (EP4): Students failure is mostly concen-
trated on the mathematics and language subjects. Students who fail
these subjects tend to reprove the year, in almost all grades and for
both genders.

4.2. Finding educational patterns through decision trees

In this group of experiments, decision trees were used to analyze the
approved/reproved status aiming to understand how different attributes
of the three DB influence students’ performance, as well as to know the
gain of knowledge that can be obtained by adding to the experiments the
data of each DB. The dynamic of the experiments was conducted using
the following dataset configurations: 1) DB1, 2) DB2, 3) DB1 + DB2, and
4) DB3.

4.2.1. Decision trees using DB1 only

Decision trees were used to better understand the behavior of the
different variables contained in the three databases. By using decision
trees, one is able to observe the generated rules and the characteristics of
the patterns. A set of experiments was carried out using several config-
urations of DB1 (divided by EI and the three educational levels). A total
of 12 decision trees were generated using J48 and K-10 cross-validation.

| EI   | Tree | Dataset | Root/Secondary node | % instances correctly Approved | % instances correctly Reproved | F1 Score | F1 Score Naive Bayes |
|------|------|---------|---------------------|--------------------------------|--------------------------------|----------|----------------------|
| EI 1 | 1    | Primary | Language/none       | 100%                           | 87.1%                          | 0.995    | 0.918                |
|      | 2    | Secondary | Language/Biology  | 99%                            | 81.2%                          | 0.954    | 0.871                |
|      | 3    | Middle   | Maths/none          | 99.8%                          | 60.3%                          | 0.975    | 0.874                |
| EI 2 | 4    | Primary   | Language/Maths     | 95.5%                          | 88.4%                          | 0.986    | 0.908                |
|      | 5    | Secondary | Maths/Arts         | 95.2%                          | 96.1%                          | 0.943    | 0.861                |
|      | 6    | Middle    | Physics/Biology    | 97.1%                          | 80.3%                          | 0.969    | 0.857                |
| EI 3 | 7    | Primary   | Language/Maths     | 97.4%                          | 86.5%                          | 0.961    | 0.842                |
|      | 8    | Secondary | Language/Biology   | 94.3%                          | 90.7%                          | 0.954    | 0.871                |
|      | 9    | Middle    | Maths/Language     | 97.3%                          | 80%                            | 0.963    | 0.915                |
| EI 4 | 10   | Primary   | Language/Maths     | 99.4%                          | 82.2%                          | 0.979    | 0.888                |
|      | 11   | Secondary | Language/Maths     | 95.6%                          | 90.1%                          | 0.957    | 0.900                |
|      | 12   | Middle    | Natural sciences/Maths | 97.3%                          | 86.7%                          | 0.962    | 0.931                |
represent the test result, and leaf nodes represent the label of the class (Sharma and Kumar, 2016). Table 7 summarizes the experiments carried out with DB1 and the results in terms of instances correctly classified for each condition.

Specifically, the process was to divide the DB1 into three groups: the first group with grades from first to fifth, the second group with grades from sixth to ninth, and in the third group with grades tenth and eleventh. After this division, the data of the qualifications for each of the subjects were taken together with the status approved/reproved as a class attribute. Later, the algorithms were applied, making a cross validation. In general, a percentage of instances correctly classified for the class "approved" is identified between 94 and 100%, while for the class "reproved" between 60 and 96%. For the failed class, a percentage of 60 is found for one of the EIs and for one of the levels, for all the others it is above 80%; however, the difference is clearly perceived with respect to the percentages reached by the approved class. The F1 Score metric is also shown for each of the trees and is compared by means of this same metric with the Naive Bayes classifier as a reference point, finding in all cases a better performance of the decision tree.

Figure 9 shows an example of the trees obtained; in this case, it corresponds to tree number 10 in Table 7. This has a percentage of correctly classified instances of 99.4% for the approved class and 82.2% for the class reproved. The root of the tree is the "Language" discipline, being the most representative attribute for the sample analyzed (primary, 1st to 5th of IE4). The next attribute is "Mathematics", achieving that 95% of the analyzed grades fail when they obtain a grade lower than or equal to 2.9 in Language and 2.8 in mathematics. Another of the subjects that has an influence is biology, which, like language and mathematics, are part of the fundamental subjects at the primary educational level.

These results allowed one to extract the following two educational patterns:

**Educational Pattern 5 (EP5):** Language discipline is the attribute that most contributes to determine the final status of the students in Primary Level. This is also true for Secondary Level with one exception, where Mathematics is the most important attribute.

**Educational Pattern 6 (EP6):** Exact sciences (Math and Physics) are the disciplines that most contribute to determine the final status of the students in Middle Level with one exception, where Natural Sciences is the most important one.

### 4.2.2. Decision trees using DB2 only

A second step was focused on using the DB2 (students’ enrollments and socioeconomic data) to generate decision trees to the different educational levels for each EI. With this group of experiments, it was possible to observe that the use of DB2 did not help to describe and understand the reproved category. With DB2 it was only possible to determine rules for the approved category as the performances of the models are too low to the reproved one.

Therefore, we decided to make a feature selection over the variables to identify, through a set of data that have certain attributes, those that have more weight when determining whether the data is of one class or another (Li et al., 2017; Cai et al., 2018). This was intended to determine the DB2 attributes, which were the ones that have the greatest influence on the approved/reproved class classification.

However, after performing an automatic feature selection, the following attributes were the ones that most influenced a student’s final status: social level, shift (morning/afternoon), geographical area (urban/rural), gender, and academic status of the previous year. Moreover, other attributes related to whether the school is located in a region of conflict did not stand out as a relevant information to the problem in hand. Then, two new educational patterns are determined:

**Educational Pattern 7 (EP7):** The following attributes are highly correlated to the final status of the students: social level, shift (morning/afternoon), geographical area (urban/rural), gender, and academic status of the previous year.

**Educational Pattern 8 (EP8):** The location of the school (region of conflict or not) is not a relevant attribute to determine the performance of the students.

An additional feature ranking analysis using xgboost was performed to ratify previous patterns and findings. The results of this analysis corroborate the following most important attributes which influence students' success: social level, gender, zone of residence, academic situation in the previous year and shift (course_session), with an F score for above 100. Figure 10 shows the graph with the results of the selection with xgboost. Moreover, as it can be seen from the figure, two new attributes appear as important in comparison with the ones pointed out in the previous sections, which are: methodology and new student. Methodology corresponds to the type of the adopted teaching method.
(traditional education, ethno-education, rural education and adult education); and a new student corresponds to whether the student is new to the institution or if he/she already was enrolled at the same institution in previous years.

4.2.3. Decision trees using DB1 and DB2

In the third moment, the dataset was used integrating DB1 and DB2. In other words, a union of the attributes of the two databases was carried out for each student. Then, for this group of experiments, the scheme by institution was also followed. With this, the same number of experiments outlined in subsection 6.2.1 were carried out.

In this case, it is highlighted that the trees and rules continued to be directly related to the qualifications of the disciplines, presenting similar results of the experiments using only DB1. This was in hand in hand with the results obtained in the second group of experiments, in which it was determined that DB2 could not describe the reproved class. Therefore, no new pattern stands out from this group of experiments.

4.2.4. Decision trees using DB3

We tried to integrate DB1, DB2 and DB3 and perform the tests following the same proposed scheme in the previous two subsections. However, DB3 does not allow us to do this process for all grade levels, as SABER has only results of exams for grades (school year) 3, 5, 9 and 11. Moreover, for grades (school year) 3, 5 and 9, results of SABER are only available to a given region and not for each EI of the cities that belong to the region. Therefore, the level of granularity of DB3 makes it incompatible for integration with DB1 and DB2.

SABER 11 presents the results by institution, but the integration with DB1 could not be done because the student’s identification was not available. Thus, the experiment that was proposed for this case was the comparison of the rules (patterns) obtained with DB1 versus the possible rules to obtain with DB3. Once again there is a difficulty, the information corresponds to two different types, for the case of the results of SABER, there is an overall performance, which was tried to adjust to compare with the condition of approved/reproved.

One of the findings in this experiment, for EI4, establishes that the rule of failure of 11th grade students for the institution is given by the subject natural sciences, while in SABER, a low performance is directly related to the result in mathematics. With the previous finding, it is possible to think that when generating policies at the institutional level, it would be decided to strengthen the training in natural sciences. However, if the policy is constructed taking into account the data of the national level, it would be oriented to strengthen mathematics.

4.3. Finding patterns through clustering

Clustering is a data analysis technique that allows dynamically grouping by calculating centroids and distributing records according to a distance measure, placing them next to the closest centroid (Hossain et al., 2019). We performed a cluster analysis over DB3 (SABER 11 exam in 2018) using the four institutions included in the dataset all together and using each institution separately. The attributes were the performances of the students in the disciplines. Figure 11 shows the results of the clusters for all institutions together (Figure 11 a - left side) and for educational institution 4 (EI4) alone (Figure 11 b - right side). The “x” axis shows the disciplines evaluated in SABER 11 and the “y” axis corresponds to the score, which can range from 0 to 100 for each discipline. The best formation of clusters was for k = 3, this cluster number was defined by elbow method, representing the three groups of students with low, medium and high performances in the disciplines.

As it can be seen in Figure 11a (left side), the three clusters are very separated and the grades of the different disciplines within each cluster tend to be between the same range. This means to say that students with a given type of performance in one given discipline tend to have the same type of performance in all disciplines (with very small variations). When we look at the clusters of an EI4 alone (Figure 11b - right side) the pattern

![Figure 10. Feature ranking using xgboost. Source: own elaboration.](image1)

![Figure 11. Clusters in DB3 for the 4 EI all together (a) and Clusters in DB3 for EI4 (b). Source: own elaboration.](image2)
is similar, with the exception of the performances in the English discipline. For this discipline, students generally have lower performances in all three clusters. Even students with high performances in the other disciplines, tend to have lower performances in English.

To facilitate the interpretation of the clusters, t-SNE (t-distributed stochastic neighbor embedding) graphs were used as a projection technique, thus allowing to visualize the complete distribution of the elements inside each cluster. Figure 12 shows the projection of the clusters for the four EI all together (left side of the figure - 12a), and for EI4 (right side of the figure - 12b). In both cases, the data included the performance of all the disciplines evaluated by the SABER11 test. As it can be seen from Figure 12a, students are clearly distributed in three clusters. On the other hand, for the particular case of EI4 (Figure 12b), the clusters are not well delimited, which can be associated with the fact that for this institution the centroids of the clusters for some disciplines are quite close (e.g., English).

Given the above, a zoom was made to review the behavior of the clusters, taking into account the performances in the discipline of foreign language (English). As it can be seen from Figure 13a, the clusters of this specific discipline for EI all together (left side of Figure 13a) are more mixed. At the same time, it is noticeable the clusters for the EI4 (right side of Figure 13b) are much more mixed.

We also performed a cluster analysis using data from EI4 stored in DB1 (same level and year - level 11, year 2018) to contrast with the previous results. Figure 15 presents the centroids for the three clusters in each discipline. It is important to highlight that the disciplines from SABER are slightly different from the disciplines of the schools as SABER organizes its exams in 5 main areas (Reading, Math, Natural Sciences, Social Sciences and English). Even though it is difficult to compare directly the results (as the disciplines are different), it is possible to see from Figure 15 that the centroids of the clusters overlap for some disciplines. This is a pattern totally different from the previous analysis. This difference in the patterns between these two analyses is quite disturbing as, in principle, both data (from Figure 11b and 15) are representing the same population.

Taking the above into account, it was decided to expand the analysis by applying self-organized maps (SOM) to review behavior of the clusters, both for performance in general and for some disciplines in particular. It was found that the trend is maintained and if particular cases such as Math (Figure 14 (a)) and Conduct (Figure 14 (b)) are reviewed, it is clear that in the case of Math there are two strong groups and a third cluster that it mimics the others, coinciding with what is presented in Figure 15, in which the centroids of clusters 1 and 2 for this discipline almost overlap. For Conduct, the superposition of the centroids for the three clusters presented in Figure 15 is ratified in Figure 14 (b), with a uniformity in performance that leaves no room for a particular distinction or concentration.

These results confirm that analyses with data from different levels may lead to different results. This may be due to the dynamics of educational environments, but also to the fact that performances are not recorded in the same way in the datasets. These results also reinforce the need to pursue the integration of the datasets in all levels, so that a better understanding of the educational patterns can be achieved.

It is revealed that in the external test, the performance is lower than in the evaluation carried out within the EI. To check the statistical significance, after evaluating the distribution of the data (they follow a normal distribution), the t-test was used to evaluate the means of the samples for the four main subjects (language, mathematics, natural sciences, and English) with a level of significance of 5%, it was found that for the four subjects, there is a statistically significant difference in the means of the qualification in the SABER 11 test concerning the qualification obtained in the EI. These analyses were corroborated both by the corresponding statistical and by p-value. It should be clarified that the test shows the significant difference between the results obtained by the students of the analyzed institutions, however, we do not affirm that the observation, as a pattern, can already be considered significant and generalized, for this a more in-depth analysis is required, this study focused on an exploratory analysis.

5. Discussion

This section discusses the main findings of the initiative while answering the research questions proposed in the introductory section.

RQ1 - Which patterns and findings emerge from exploratory analysis and data mining techniques in primary and secondary Colombian educational institutions?

Experiments using DM and visualization techniques helped to uncover several educational patterns that we presented during the paper. For instance, we confirmed that students from high-income families (level 4, 5, and 6) do not enroll in public institutions and that most of the students of public schools studied are from level 1 families.

Moreover, visualization techniques helped to demonstrate that the grade of transition from primary to secondary level (6th grade) concentrates the highest percentage of students who fail (mostly males). This is an important finding that can help to create educational practices and social policies tailored to this particular group of students at-risk. One can venture that the concentration of failures at this grade may be related to the proximity with the age they are entering puberty and to the changes that involve the transition to another kind of education.

The transition from Primary to Secondary represents a great challenge for all educational actors. Children find it extremely difficult to adapt to the norms, structure, teaching style, development of tasks, and other activities required by the Secondary. Teachers struggle to achieve the framing of their classes, advancing in the development of classes, and
Figure 13. Projection of the clusters considering only performances in English and for all institutions (a) and for only EI4 (b).

Figure 14. Clusters with DB1 for Math and Conduct (Level 11, year 2018). Source: own elaboration.

Figure 15. Cluster with DB1 for EI4 (Level 11, year 2018). Source: own elaboration.
fostering children's achievement of the objectives and expectations generated for the course. Parents also mention how this grade step is sometimes experienced by children who present high levels of anxiety and concerns about new demands, adaptation to the teaching team, and the dynamics of day-to-day life in the Secondary (Gaviria Arbeláez, 2016). If in primary fewer subjects are taken (time intensity is lower), in secondary students have to deal with new subjects/disciplines and very often need to change their educational center and shift (morning/afternoon). These are all situations that involve anxiety and that could benefit from strategies such as the Krashen's Theory of Affective Filter (Hui and Lin, 2008; Warr and Downing, 2010) which considers a series of affective variables (motivation, self-confidence, and personality traits) to facilitate the process of teaching and acquiring knowledge.

The exploratory analysis through the use of DM techniques helped to unveil that student's failure is mostly concentrated on mathematics and language (with approximately 50% of failure) and that language is the attribute that most contributes to determine the final status of the students in the disciplines were quite different for the databases. One of the problems encountered to compare the results has to do with the change of variables over time in DB3 (the scales and metrics for measuring the performance of the students changed over the years).

RQ3 - Which are the challenges in the design and execution of a regional level analytic learning experience in Colombia and how do they differ from challenges encountered in the state of the art in Latin America?

Macarini et al. (2019b) have identified eight challenges during the development of a K-12 countrywide LA initiative in Uruguay. Here we go through some of these challenges relating them with our experience in Colombia and expanding them when it is possible. Figure 12 summarizes the challenges, including those proposed by Macarini et al. (2019b) are gray-highlighted. Black and white-highlighted challenges are proposed in this work; indeed, the later are considered subcategories of other previous challenges.

The implementation of a LA experience on a medium scale in Colombia presents several challenges and difficulties that are faced by the stakeholders involved in the initiative (see Figure 16). One of the main challenges arises from receiving the data from different institutions in many different sources and formats. Processing and integrating the data required the involvement of an expert with the knowledge about the educational context. To deeply know the country's educational system is one of the indispensable aspects for the identification of influential factors in an experience like this one. In our experience, it is ratified the challenge C1, which emphasizes the need of an active participation of the governmental actors responsible for the data.

The management of ethical and legal aspects was another sensitive issue that is ratified with this experience (C2). However, unlike the
Uruguayan experience in which the government had structured policy for data privacy and access, in our initiative there was no clear policy for such a scenario. Thus, the strategy for data access was built on-the-fly by the researchers and the entities in charge of the data. For DB3, there was no need for new policies as the data was openly available and already anonymized. The experiments were carried out after assuring the security of the data and signing a confidentiality agreement with the data owners. However, the lack of a concrete policy for data access generated additional efforts which we classified as challenge C3.1. Challenge 3 (C3) reported in the Uruguayan experience is extended since there was a delay in the completion of access requirements, but the policy existed. Therefore, it is suggested that from the national level (Ministry of Education), the Departmental Education Secretariats should focus on the development of an ethical and legal mechanism for data access, security, and privacy. As mentioned before, one of the main challenges consisted of receiving data from different institutions and in many different formats, scales and levels of granularity. The challenge C4, needed an integrated approach as it involved different stakeholders that knew just a small part of the data. Considering that, the work of the researchers also included the construction of a general panorama of the facts in order to achieve the integration of data sources (CA4.1). For this reason, it is also recommended the creation of mechanisms that allow an access policy to data from scales such as the institutional one. Moreover, this type of data can be included in the analyzes since the analysis carried out at the national level fail to cover or identify the reality of the institutions, or of the regions. Regarding Challenge 5, the present work still did not enter the phase of developing a system, and is restricted to the discovery of educational patterns.

In C6.1, the need to transcend the initial proposal and understand the phenomena presented in education systems is raised. This expands C6, which emphasizes the need to face the proposal of the researchers following the requirements of the interested entities. For this work, it is identified the need to properly construct the dataset, tending to achieve an adequate functioning of the algorithms (C7.1). Subsequent to the execution of the analysis, another aspect in which special interest should be paid is the results visualization of it (C8). This is important because the transference of the knowledge extracted from the mining process and the descriptive analysis guarantee a better assimilation of the findings and later use for the creation of plans of improvement or policies. In the Colombian experience, a new challenge is posed (C9) related to the ownership of data and the open data dynamics that are being implemented in the country. This has to do with the fact that many schools have their data managed by private companies, compromising the access and usage of the data for LA purposes. It is therefore recommended that it is essential to have a data protection policy when these are handled by private entities and not directly by educational institutions, since much traceability is lost and opportunities for analysis are wasted.

In conclusion, stands out the importance of data integration at all levels, better policies for access and manipulation of information. In this case, the integration did not reach the most desired level, but it can be inferred that the stronger the integration of systems is, the more knowledge can be achieved. At all three levels, a public policy needs to be formulated that focuses on the creation of spaces and mechanisms for the application of regional and in the longer-term national learning analyses.

6. Conclusions and future work

This work is a first attempt towards a regional initiative of learning analytics and educational data mining, and tries to show some knowledge that can be extracted from the educational systems in K12 and the challenges we faced to provide these preliminary results in the regional level in Colombia. There is the need for more in-depth studies that help to better understand the value that LA can add in Colombia. One of the main factors that influenced the need to reduce the scope of the study was the fact that there was no standardized rule for reaching the approved/reproved condition, because each EI can count on its particular criterion. However, one achievement was to transcend the experience from an institutional to a regional scope.

Looking at the challenges found in this experience, which started from the absence of clear policies of sharing and using the data, it was also confirmed that the preprocessing and cleaning of data is one of the most arduous stages in EDM for reasons such as the difference in scales, the granularity, and decentralization of the educational data. The integration of data was another major challenge and scope of the process, in this sense one of the contributions given to educational institutions and government entities is to define a student identifier that can be used to track their trajectory within and across institutions. This is an important step to take given that the identity document in Colombia changes over the years and there are additional problems with differences in the digitization of the name and surname.

The progress achieved in this project created a working space with interest for government entities in charge of data and the generation of educational policies of the region, who could perceive that advantages of LA are highlighted not only at the university level, but also at previous levels (basic and middle education). However, this also poses challenges because the information systems are decentralized at this educational level in Colombia, and additional with difficulties of access. Continuing advancing in open data use and sharing policies is a fundamental aspect in the generation of a methodology and strategy for LA. Findings from this experience and related experiences can begin to further motivate stakeholders and make their contribution to these research initiatives more active.

In the case of data that is treated at the regional and institutional level, it must start with the adoption of policies for the use and sharing of data, which can be supported by the Ministry of Education that already has progress in the matter. It is proposed as a strategy that educational institutions and their principals be trained in the importance of protecting and storing the data that is being generated inside their facilities and as part of the development of educational processes. This task is not easy given that elementary and middle education institutions do not have their own information systems, but instead contract these services with third parties and in many cases the rectors do not have the technical knowledge to negotiate on issues such as for example the delivery of the databases. Regarding this aspect, an initiative in the analyzed region has already begun to be implemented, it is an academic information system, which allows institutions to integrate most of their processes and store their data, such as student registration, enrollment, evaluations, among others. This initiative of the Departmental Education Secretariat is in the pilot test stage and the idea is to unify under this software all the public EIs of the department. A future work of this study is to analyze the data collected by this software, currently functional in 26 institutions.

In this experience, it was possible to use data mining through experiments with decision trees to find in a more automated way the relationships and patterns inside the data. One of the reasons that influenced the need to reduce the scope of the study was the fact that there was no standardized rule for reaching the approved/reproved condition, as each EI may formulate/apply different criteria to determine when a student should be approved/reproved.

In other hand, Educational visualization data falls under the scope of the field of LA dashboards and have been frequently mentioned as a fundamental step to help in the understanding of learning processes and educational scenarios (Viberg et al., 2018). For this project, and after the integration of the databases, we started the development of a system to provide data visualizations in the three educational levels. Next steps of the development will involve visualizations similar to the ones proposed by Macarini et al. (2019a).

Finally, as future work, in addition to those already mentioned, the leadership of an initiative to broaden the comparison of the challenges
encountered in Latin America for the implementation of a LA strategy is suggested. The proposal needs the help of communities that have been consolidated and work on the issue, and making known the importance of carrying out experiences not only in the university environment but also in primary and secondary education. Also, in order to evaluate the coherence between the results in the institutional and national performance, it is intended to determine whether students have the same behavior in the institutional and in the national exams.

Acknowledgements

This work was supported by Minisciences -The training program “High Level Human Capital for the Department of Norte de Santander”- Call 753 of Colciencias, and partially supported by CNPq (Brazilian National Council for Scientific and Technological Development) [Editorial Universal, proc.404369/2016-2] [DT-2 Productivity in Technological Development and Innovative Extension scholarship, proc. 315445/2018-1 and Colciencias - The research program "Reconstruction of the Social Fabric in Post-Conflict Zones in Colombia" SIGP Code: 57579] with the research project “Teaching strengthening from Informational Media Literacy and CTeI, as a didactic-pedagogical strategy and support for the recovery of trust in the social fabric affected by the conflict” SIGP code 58950 Funded with the Colombia Cientifica call, [Contract No FP44842-213-2018].

References

Aguilar Barreto, A.J., Rodríguez Manans, G.A., Aguilar, C.P., 2018. Management of educational public policies: a feature in the North of Santander (Colombia). Espacios 39 (30), 5–19. http://www.revistaespacios.com/vol/a13/vol30/18395005.html.
Avela, J.T., Kebritchi, M., Nunn, S.G., Kanai, T., 2016. Learning analytics methods, benefits, and challenges in higher education: a systematic literature review. Online Learn. 20 (2), 13–29.
Broos, T., Hilliger, I., Pérez-Sanagustin, M., Hitn, N., Millemp, M., Pésant-Cebra, P., Solano-Quinde, L., Sigüenza-Guzman, L., Zuniga-Prieto, M., Verbret, K., De Laet, T., 2020. Coordinating learning analytics policymaking and implementation at scale. Br. J. Educ. Technol. 51 (4), 928–954.
Cai, J., Luo, J., Wang, S., Yang, S., 2018. Feature selection in machine learning: a new perspective. Neurocomputing 300, 70–79.
Calz Wilches, O.E., Grisales-Palacio, V.H., 2019. Learning Analytics in Colombia: Una revisión a la literatura y análisis del esfuerzo de investigación local. II Conferencia Latinoamericana de Analíticas de Aprendizaje – LALA- 1-11. http://ceres.org/V ol-2425/paper07.pdf.
Cechinel, C., Ochoa, X., Lemos dos Santos, H., Carvalho Nunes, J.B., Rodés, V., Marques Queiroga, E., 2020. Mapping learning analytics initiatives in Latin America. Br. J. Educ. Technol. 1–23.
Chica Gómez, S., Galvis Gutiérrez, D., Ramírez Hassan, A., 2012. Determinantes del rendimiento académico en Colombia. Pruebas ICFES - Saber 110, 2009 Rev. Univ. IAPFT 46 (160), 48–72.
Conde, M.A., Hernández-García, A., 2013. A promised land for educational decision-making?: present and future of learning analytics. In: TEEM '13 Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multicultural, pp. 239–243.
Delgado Barrera, M., 2014. La educación básica y media en Colombia: retos en equidad y calidad. https://www.repository.fedesarrollo.org.co/handle/11445/190.
dos Santoyo, B.R.C., Cechinel, C., Núñez, A., 2017. An Initial Review of Learning Analytics in Latin America. 2017 Twelfth Latin American Conference on Learning Technologies (LACLO), 1–9.
Dulva, C.J., Talmon, R., Gear, C.W., Colman, R.R., Kevrekidin, I.G., 2015. Data-Driven reduction for multiclass stochastic dynamical systems. Appl. Dyn. Syst.
Du, X., Yang, J., Shelton, B.E., Hung, J.L., Zhang, M., 2019. A Systematic Meta-Review and Analysis of Learning Analytics Research. Behaviour and Information Technology.
Ducoing, P., 2019. La educación secundaria en el mundo : el mundo de la educación secundaria (Colombia, Brasil y Argentina) (Primera edición). Universidad Nacional Autónoma de México, Instituto de Investigaciones sobre la Universidad y la Educación. http://132.248.192.241:8080/jspui/handle/ISU/UNAM/438.
Dutt, A., Ismail, M.A., Herawan, T., 2017. A systematic review on educational data mining. IEEE Access 1–1.
Ferguson, R., 2012. Learning analytics: drivers, developments and challenges. Int. J. Technol. Enhanc. Learn. (ITEFL) 5/6.
Firat, M., 2016. Determining the effects of LMS learning behaviors on academic achievement in a learning analytic perspective. J. Inf. Technol. Educ. 15, 75–87.
Gaviria Arbeláez, M.T., 2016. La Transición de la educación primaria a la educación secundaria, un asunto por entender y atender desde la cotidianidad escolar. Corporación Universitaria Lasallista.
Guzmán-Paredes, C., Flórez, N., Gómez, D., Sánchez, D., 2016. Analyzing social construction of knowledge online by employing interaction analysis, learning analytics, and social network analysis. In: Quarterly Review of Distance Education, 17, pp. 35–60. Issue 3. https://search.proquest.com/openview/c03f7525a61b7f532e61be3f58197690/1?pq-origsite=phd
Hilliger, I., Ortiz-Orjas, M., Pesant-Cebra, P., Scheihing, E., Tsai, Y.S., Muñoz-Merino, P.J., Bros, T., Whitelock-Wainwright, A., Pérez-Sanagustin, M., 2020. Identifying needs for learning analytics adoption in Latin American universities: a mixed-methods approach. Internet High Educ. 45, 100726.
Hoppe, U., 2017. Computational methods for the analysis of learning and knowledge. In: Lang, C., Siemens, G., Wise, A., Gasevic, D. (Eds.), Handbook of Learning Analytics, First, Society for Learning Analytics, pp. 23–33.
Hossain, Z., Akhtar, N., Ahmad, R.B., Rahman, M., 2019. A dynamic K-means clustering algorithm for data mining. Indones. J. Electr. Eng. Comput. Sci. 13 (2), 521–526.
Hu, X., Cheng, C.W.L., Ding, W., Wu, M., 2017. A systematic review of studies on predicting student learning outcomes using learning analytics. In: LAK ’17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 528–529.
Hübscher, R., Puntambekar, S., Nye, A.H., 2007. Domain Specific Interactive Data Mining. Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling, 81–90. http://dlid.redesign.wceruw.org/pub lications/Hubsescher_2007_UM.pdf.
Hui, G., Lin, C., 2008. Pedagogies proving Krashen’s theory of affective filter. J. Eng. Educ. 104, 113–131. https://eric.ed.gov/?id=ED503681.
Instituto Colombiano para la Ia, & Evaluación de la Educación Icfes, Portal Icfes. http://www.icfes.gov.co/web/iast/funciones/icfes.
Jiménez Ángel, F., Espinosa Restrepo, J.R., Parra Heredia, J.D., García Villegas, M., 2013. Separados y desiguales: Educación y clases sociales en Colombia. Centro de Estudios de Derecho, Justicia y Sociedad. https://www.dejusticia.org/publicatio n/separdos-y-desiguales-educacion-y-classes-sociales-en-colombia/.
Kaesman, K., 2016. The role of learning analytics in Global higher education. In: Developing Effective Educational Experiences through Learning Analytics, pp. 282–307.
Koc, M., 2017. Learning analytics of student participation and achievement in online distance education: a structural equation modeling. Educ. Sci. Theor. Pract. 17 (6), 1893–1910.
Kucuk, S., Richardson, J.C., 2019. A structural equation model of predictors of online learners’ engagement and satisfaction. Online Learn. 23 (2), 196–216. https://eric.ed.gov/?id=EJ1218390.

Lawn, M., 2013. The Rise of Data in Education Systems: Collection, Visualization and Use. Symposium Books.

Leitner, P., Khalil, M., Ehner, M., 2017. Learning analytics in higher education—a literature review. In: Petia-Ayala, A. (Ed.), Learning Analytics: Fundamentals, Applications, and Trends, pp. 1–23.

Li, Y., Li, T., Liu, H., 2017. Recent advances in feature selection and its applications. Knowl. Inf. Syst. 53 (3), 551–577.

Macarini, L.A., Lemos dos Santos, H., Cochinel, C., Ochoa, X., Rodés, V., Pérez Casas, A., Díaz, P., 2019a. Towards the implementation of a countrywide K-12 learning analytics initiative in Uruguay. Interact. Learn. Environ. 1–25.

Macarini, L.A., Ochoa, X., Cochinel, C., Rodés, V., Dos Santos, H.L., Alonso, G.E., Díaz, P., 2019b. Challenges on Implementing Learning Analytics over Countrywide K-12 Data. LAK 2019. In: 9th International Conference on Learning Analytics and Knowledge, pp. 441–445.

Mangaroska, K., Giannakos, M.N., 2018. Learning analytics for learning design: a systematic literature review of analytics-driven design to enhance learning. September 1 IEEE Trans. Learn. Technol.

Martin, F., Whitmer, J.C., 2016. Applying learning analytics to investigate timed release in online learning. Technol. Knowl. Learn. 21 (1), 59–74.

Moreno Cadavid, J., Pineda Corcho, A., 2018. A systematic literature review in Learning Analytics (1). In: Anais Dos Workshops Do Congresso Brasileiro de Informática Na Educação, 7, p. 429.

Organización para la Cooperación y el Desarrollo Económicos (OCDE), 2016. Educación en Colombia.

Romero, C., Ventura, S., 2007. Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl. 33 (1), 135–146.

Romero, Cristobal, Ventura, S., 2013. Data mining in education. Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 3 (1), 12–27.

Romero, Cristobal, Ventura, S., 2010. Educational data mining: a review of the state of the art (6). In: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40, pp. 601–618.

Ruiz-Calheira, A., Garcia, S., Tammets, K., Aguerrebere, C., Ley, T., 2019. Scaling Learning Analytics up to the National Level: the Experience from Estonia and Uruguay. Learning Analytics for Latin America LALA, pp. 1–10. http://ceur-ws.org/Vol-2425/paper01.pdf

Sahar, Y., Seifedine, K., Sicilia, M.-A., 2016. A Framework for Learning Analytics in Moodle for Assessing Course Outcomes. Global Engineering Education Conference (EDUCON).

Scheuer, O., McLaren, B.M., 2012. Educational data mining. In: Encyclopedia of the Sciences of Learning. Springer US, pp. 1075–1079.

Sharma, H., Kumar, S., 2016. A survey on decision tree algorithms of classification in data mining. Int. J. Sci. Res. 5 (4), 2094–2097. www.ijsr.net.

Slater, S., Joksimovic, S., Kovanovic, V., Baker, R.S., Gasevic, D., 2016. Tools for educational data mining: a review. J. Educ. Behav. Stat.

Tsai, Y.-S., Gasevic, D., 2017. Learning analytics in education: literature review and case examples from vocational education. In: LAK ’17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 233–242.

Viberg, O., Hatakka, M., Balter, O., Mavroudi, A., 2018. The current landscape of learning analytics in higher education. Comput. Hum. Behav. 89, 98–110.

Warr, P., Downing, J., 2010. Learning strategies, learning anxiety and knowledge acquisition. Br. J. Psychol. 91 (3), 311–333.

World Bank, 2018. World Development Report 2018: Learning to Realize Education’s Promise.