Learning Analytics focused on student behavior. Case study: dropout in distance learning

Jose Aguilar

Departamento de Computación, Universidad de Los Andes, Mérida, Venezuela.
Prometeo researcher, Universidad Técnica Particular de Loja, Ecuador
aguilar@ula.ve

and

Guido Riofrio, Eduardo Encalada

Universidad Técnica Particular de Loja, DCCE, Loja, Ecuador
geriofrio@utpl.edu.ec, aeencalada@utpl.edu.ec

Abstract

Normally, Learning Analytics (LA) can be focused on the analysis of the learning process or the student behavior. In this paper is analyzed the use of LA in the context of distance learning universities, particularly focuses on the students' behavior. We propose to use a new concept, called "Autonomic Cycle of Learning Analysis Tasks", which defines a set of tasks of LA, whose common objective is to achieve an improvement in the process under study. In this paper, we develop the "Autonomic Cycle of LA Tasks" to analyze the dropout in distance learning institutions. We use a business intelligence methodology in order to develop the "Autonomic Cycle of LA Tasks" for the analysis of the dropout in distance learning. The Autonomic Cycle identifies factors that influence the decision of a student to abandon their studies, predicts the potentially susceptible students to abandon their university studies, and define a motivational pattern for these students.

Keywords: Learning Analytics, Distance learning Universities, Data Mining, Autonomic Cycle.

1 Introduction

In the literature, there are a lot of LA tasks. LA normally is based on the collection and analysis of data about students and their learning contexts, for purposes of understanding and optimizing his/her learning process and the environments in which it occurs. For that, LA uses statistical techniques, machine learning approaches, data visualization techniques, among others.

Traditionally, the LA tasks can be grouped in two groups: LA tasks to generate indicators about the learning process, and LA tasks to understand the student behaviour. In this paper, we are interested in the LA tasks focused on student behaviour. Normally, they are used to discover learning styles, to determine how the performance of each student varies according to the learning patterns of the courses, to identify students' problems and needs, among other things. In this paper, we are interested in analysing the reasons for students dropping out. One of the current problems is the correct utilization of sets of LA tasks, in order to reach complex goals. The integration of LA tasks allows solving complex problems, which have so far been impossible to study, by the amount of knowledge required for resolution.

Additionally, some papers have determined the dropout can be caused by professional, academic, health, family or personal reasons. In this work, we like analyse the dropout in the context of distance learning institutions. The analysis of the dropout is very complex, and require a lot of knowledge in order to solve it. In this paper, we propose the paradigm of the "autonomous cycle of tasks of LA", in order to define a strategy that allow the integration of tasks of LA to understand this complex problem, and to generate the knowledge required to make decisions that minimize the dropout.

In that sense, in this paper we introduce the concept of the autonomous cycle of tasks of LA tasks, to analyse the dropout in the context of distance learning institutions. According to the autonomous cycle of LA tasks for the dropout problem in the context of distance learning institutions, we require the next tasks: a classification task of the

students as Deserters and non-Deserters, a predictive model to determine susceptible students to abandon their university studies, a descriptive model to identify the factors that influence the decision of a student to drop up, and a task to build a motivational pattern for the susceptible students to abandon their university studies.

Also, we use a methodology developed in [1], which allows us to guide the specification of the LA tasks. In that sense, the methodology characterizes the multidimensional data model as well as the tasks of data mining and OLAP (On-Line Analytical Processing) operations required. In this paper, the method is adapted to our case as follows. It starts with the characterization of target situations, which in our case is to analyse the dropout in the context of distance learning institutions. Based on it, we define the autonomous cycle of tasks of LA tasks, that in our case is composed of the set of the following LA tasks: the classification of the students as Deserters and non-Deserters, the construction of the descriptive and predictive models of the students dropping out, and the construction of the motivational pattern for the potential Deserter students. Then the data model required for such tasks is defined, and the extraction and processing of data from operational databases are performed to build the data view on which the process is applied. Finally, are defined the OLAP operations and the data mining tasks related to: i) Classify the students as desserts or not, ii) Identify the factors that influence a student's decision to abandon his studies, and iii) Predict the potential students to leave their studies and, iv) Build the motivational pattern for the potential deserters. In particular, the data source used by the distance education institution comes from its Virtual Learning Environment (VLE).

The rest of the paper is organized as follows: section 2 presents the theoretical aspects of our proposal. Section 3 presents our proposition of autonomic cycle to analyse the dropout problem. Section 4 characterizes the process of BI in our autonomic cycle of LA tasks, and Section 5 analyses the results obtained with it. Finally, the Section 6 presents the conclusions.

2 Theoretical Aspects

2.1 Learning Analytics

LA can be defined as the use of data produced by the students during their learning processes in order to build models, patterns, etc., required to analyse and improve these processes [2], [3]. Some of the possible results are predicting the performance of the students, guiding the learning process, recommending learning resources, among others.

The roots of LA are in several fields, particularly: Data mining, Business Intelligence, Web Mining, Recommender systems, among others. LA encompasses a set of educational technologies, algorithms, models, techniques, methods, and best practices, in order to analyse the learning trajectory of a student [2], [4].

Normally, the LA tasks must be focused on the learning process or on the students' behaviour [4]. In this paper we are interested in the second type. Some of the LA tasks focused on the students' behaviour are [2], [3], [4]:

- Discover learning styles: in this case the idea is to discover the learning styles of a student, of a group of students (for example, during a course). The idea is (re)build the student profile based on what the student does, score of the student, etc. For this specific activity, SaCI must use a learning style model like the Felder-Silverman model, which determine the tools, the learning strategies, the evaluation methods, of a style of learning.
- Determine how the performance of each student varies according to the learning patterns of the courses. For that, the LA task must analyse the student's participation in the course, among other things.
- Identify students' problems and needs: in this case are used the LA tasks in order to discover the subjects not covered in assessments, identify topics that students need more attention, which questions students fail more, etc.

2.2. Methodology for BI projects

In this paper, we use a methodology proposed in [1] for BI projects, but adapted to our case. The methodology defines the objective situations, the indicators to see if these are achieved, and the tasks of data/semantic mining to calculate these indicators.

Normally, the objective situations describe knowledge that generally with the operational data is impossible to obtain. Objective situations are determined by indicators, which can be obtained from data/semantic mining tasks or OLAP operations. The methodology consists of the following steps [1]:

2.2.1 Step 1: Defining objective situations

In this case are defined the main questions that should answer the BI project. Normally, the objective situations are strategic questions that the organization should respond. These target situations justify the BI project, and define the needs and business opportunities identified. The objective situations are very important because they define the indicators looking to get. Thus, the main objective at this stage is to define the objective situations, describing different types of states in an organization: strengths, weaknesses, opportunities and threats (SWOT). Another important aspect in this phase is to identify potential indicators, which describe each objective situation to study.

2.2.2 Step 2. Data Model of the BI project

In this phase is prepared the data to use in order to calculate the indicators. Normally, at this stage the data warehouse that contains both, historical and current data, is conceptually designed. For that, we use techniques and strategies from the "data sciences" domain [5], [6]. Data warehouses are based on multidimensional data models, in order to take into account the main data to analyse the objective situations [7]. Traditional relational database cannot normally handle objective situations. Data warehouses require the extraction, transformation, processing, integration and analysis of data. At this stage are defined the extraction, transformation, processing and integration tasks. Some of the steps in this stage are:

- Analyse operational data of the organization, that is, in this step the data sources are identified.
- Design the multidimensional database, which defines the logical and physical data model.
- Design the ETL process, which will be used to recover data from the data sources.
- Run the ETL process with the operational data of the organization, in order to build the operational minable view stored in the data warehouse.

Multidimensional models represent an extension of the relational model, and typically are based on a star schema [7]. This model is used to calculate performance indicators.

2.2.3. Step 3. The knowledge extraction (indicators)

At this stage are calculated the different models and statistical metrics for the interpretation and analysis of the objective situations, using the respective indicators. Typically, in this phase is extracted the hidden knowledge in the data warehouse, by using data/semantic mining tasks or OLAP operations, in the form of indicators.

The OLAP engine is a query builder to explore and analyse the information in the multidimensional data warehouse [7]. OLAP tools provide analysis to find trends in the data, but do not discover hidden information (patterns, etc.). These tasks require more powerful tools such as semantic/data mining techniques.

Data mining is fundamentally about processing data, in order to identify patterns and trends, which can be used to decide or judge [8], [9]. Semantic mining is responsible for extracting semantic knowledge from different semantic sources, such as web pages, annotated graphs, and ontologies, among others. The semantic mining is divided into three groups [10]: semantic data mining, web mining and ontological mining. To implement the data mining or semantic tasks, there are several methodologies. The main aspects to consider is that the tasks of data mining/semantic should be well defined, and for each one an important step is the preparation of data, which is composed of two aspects: the definition of the conceptual view of the data, and the construction of the operational view of the data. If these views are incorrect, the results of data mining /semantic are also incorrect.

2.3. Related Work

In the last years, we have seen that new areas of study are emerging in the educational domain, for example the Educational Data Mining (EDM) and LA [11]. In general, there are a lot of papers in these domains, focused on course level analysis, predictive models, semantic content analysis, sentimental analysis, and social network analysis, in the general educational context [1], [12], [13], [14], [15], [16], [17], [18], [22], [26], [33].

Specifically, in the domain of dropout analysis, there are several works based on data mining or LA approaches. For example, in [27] is purposed the expansion of the analysis of the early dropout identification, in order to improve the prediction, through the use of a longitudinal database, rather than national survey data or district data. The predictive model is based on the logistical regression and four independent variables (age, attendance, gender, and test score), in order to predict the high school dropouts. They identify as dependent variable, the completion status and the career, and that the two variables more significant predictors are the age and gender. In [28], they study the prediction of dropouts through data mining approaches in an online program. The sample is composed of 189 students of the online Information Technologies Certificate Program, during 2007-2009. They collected data

from online questionnaires about 10 variables, which were gender, age, prior knowledge, educational level, occupation, readiness, previous online experience, the focus of control, self-efficacy, and the dropout status as the class label (dropout/not). To classify dropout students, they used the data mining approaches: K-Nearest Neighbour, Decision Tree, Naive Bayes and Neural Network. In their study, the best approach was K-Nearest Neighbour. Additionally, they use a Genetic Algorithm to find the most important factors in predicting the dropouts, which were the efficacy, the online learning readiness, and the previous online experience. In [29], they analyse the causes of the first year students' dropout rates in higher education institutions, using data from the engineering program at Latvia University of Agriculture. They evaluated the next variables: gender; secondary school grades and the finance source (government-financed or self-finance). The results show that the main reasons for dropping out are the low motivation to study engineering and the students' low secondary school knowledge. [30] analyses the dropout levels of Public Secondary Schools in Kericho District of Kenya, for the period 2004- 2007. They collected data from 64 public secondary schools. Statistical analysis was done, to obtain means, frequencies, T-tests, among others, to establish internal efficiency levels. The study determined that dropout levels were higher in day schools compared to boarding schools, single stream schools compared to more than one stream schools, and mixed schools compared to single sex schools. The study also found that dropout rates increased with increasing levels of education. This information was used to make decisions about school size, school regime and school type, to improve the indicator about dropout rate. In [31], they analyse information on students from the higher education system, to define the key processes in order to enhance the efficiency of studying. They use the next data mining approaches to predict the student dropout: logistic regression, decision trees and neural networks. The models were built according to the SEMMA methodology. In addition, they define a model for strategic planning to improve the efficiency of studying. [19] applies data mining methodologies on educational data, to limit student dropout in university-level distance learning. They argued that the dropout can be caused by professional, academic, health, family or personal reasons, and varies depending on the education system adopted by the distance learning institution, as well as the subject of studies. Finally, in [33] is described the results of an educational data mining case study in a smart classroom, which predicts the Electrical Engineering (EE) students dropout after the first semester of their studies, as well as identifying success-factors specific to the EE program. They determine that decision trees give useful results, with accuracies between 75 and 80%. Besides, they analyse the misclassifications, in order to show a few ways of further prediction improvement, without having to collect additional data about the students. Some future challenges in LA, defined in [3], [4], to analyse the dropout problem, are:

- Establish a bridge between LA and the learning sciences (cognition, metacognition and pedagogy). In order to optimize a learning process, it is required a good understanding of how learning takes place, how it can be supported, among other things.
- Exploit a wide range of data around learning environments, including not only the data from VLE or LMS, but also from informal or blended learning environment, behaviour of the students in the Internet, academic information, etc. Additionally, must be included mobile data, biometric data, mood data, etc.
- Must be transparent, which can be used to refine the analytics.
- Must provide knowledge with pedagogical and ethical integrity. It must provide learning indicators that genuinely promote meaningful learning.

The previous works give an idea of the variety of research in LA to analyse the dropout problem. They show how the knowledge generated can be used to analyse the dropout problem, and they describe specific aspects of this problem: prediction, the main factors that cause the dropout, among others. But in general, they do not propose an integral approach to analyse this problem, in order to monitor, analyse and make the decisions in order to reduce the dropout rate in a given context. This paper proposes an integral approach to analyse the dropout problem in higher education using LA tasks, based on the concept of autonomic cycle.

3 Autonomic Cycle to analyse dropout problem in the distance learning universities

3.1 Conceptualization of an autonomic cycle of LA tasks

An autonomous cycle of LA tasks is composed of a set of tasks of LA, in order to improve the learning process [32]. This set of LA tasks have different roles: to observe the learning process, to analyse it, to make the decisions to improve it. In this way, there is an interaction and synergy between the tasks, in order to generate the knowledge required, with the goal of improving the learning process. These tasks have sense together, and need to work together to reach the improvement goal. The autonomous cycle is a closed loop of tasks of LA, which supervises permanently the learning process.

Some of the integrated tasks include the social LA [14], in order to solve complex problems, which have so far been impossible to study, by the amount of knowledge required for the resolution. Particularly, the LA tasks have different roles to analysis the process [32]:

- Observe the learning process: This set of tasks must monitor the learning environment, and must capture the data and information about the behaviour its different components (VLE, etc.). That means, they generate a picture about the current state of the learning environment. Some of these data can be predicted, estimate using other information, extract from the student registers, among other things.
- Analyse the learning process: This set of tasks has the goal to interpret, to understand, to diagnose, among other things, the current learning process. These tasks build knowledge models about the dynamics of the learning process, in order to detect patterns of behaviour, to diagnose specific situations, to determine the causes of a phenomenon, among other things.
- Make the decisions to improve the learning process: These tasks help the process of making decisions, because they generate knowledge useful. This knowledge can be used by the tools/peoples that generate decisions about different aspects of the learning process.

In this way, there is an interaction and synergy between these tasks, in order to generate the knowledge required, with the goal of improving the learning process. In this close loop, it is required a data model that characterizes the data required by the LA tasks. Normally, it is composed by the classical data warehouse generated from the transaction databases of an educational institution. The data model requires specific tasks to prepare the data, which are different according to the source of information and the requirements of the LA tasks. In this way, we need to define different mechanisms to prepare them.

One important remark is that our autonomic cycle can execute data mining tasks and semantic mining tasks (base of the social LA). In this way, it can use information from the organization, but additionally can include other information outside of the organization (for example, from internet). Additionally, it can use different types of knowledge representations: ontologies, cognitive maps, etc. Additionally, it is transparent to the techniques of data processing. In this context, the classical design strategy of data analysis tasks must be redefined.

3.2 Autonomic Cycle of tasks of LA to analyse the dropout problem.

In this section, we present our autonomous cycle of tasks of LA to analyse the dropout problem (see Figure 1).

Figure 1. Our Autonomic cycle model

This autonomic cycle analyses the student desertion, its purpose is to avoid academic desertion by implementing tasks of monitoring, analysis and decision making, in order to get an idea of the causes of the desertion, profiles of deserters students, and patterns to motivate the students. Our Autonomic cycle is composed of four Tasks:

- First LA task: it is called “Classification of the Students as Deserters and non-Deserters”, and it seeks to divide the general student population into two big groups, in order to extract the group which we are interested, the “deserters”. In this case, we use a classical classification technique based on the dropout status that is the class label (dropout/not). We extract of the student registers the information about each student, and our classification model uses this label to group the students in two groups: “deserters”, or “Non-deserters”.
- Second LA task, it is called “Build a Deserter Profile”. With the information of the first task, this task defines patterns of the deserters. For that, this second task seeks to determine the connection between the student condition (career, mode of study, subjects) and its desertion. The frequency which a variable of the student condition appears in his profile is an important information for further decisions. In this task we build a descriptive model, using association rules [25]. Particularly, the *Apriori* algorithm [25]. The associative rules are a method for discovering relations between variables, using some measure. These relationships between variables are defined like rules, and the classical measures are support, which is defined by the proportion of transactions in the database which contains the item-set in the antecedents of the rule, and the confidence, which is the proportion of the transactions that contains the antecedents and consequents of the rule. The *Apriori* algorithm is an algorithm for identifying the frequent individual items in a database. The association rules have been used to describe the profile of a student based on parameters such as grades, attendance, work notes, etc. [9]. All information regarding the student profile, is extracted from the database of the educational institution, and the profiles are only generated for students classified as deserters in the first task of the autonomic cycle.
- Third LA task, called “predict a potential student deserter”. This task predicts the potential students to dropout, in order to apply the strategies defined in the next task in them. We build a predictive model in this task, by using the profiles created in the second task, in order to determine the most vulnerable population for desertion, and to motivate them. In this task phase, the predictive model is based on a Bayesian network. We have analysed several techniques (neural networks, a Bayesian network), and how the training time with the neural network was significantly higher compared to Bayesian networks, and the results of precision and recall were similar, the Bayesian network technique was selected. Bayesian networks are based on a probabilistic model, representing the conditional probabilities between variables [20], [23]. With this relation between the variables, can be defined a directed graph. This classification technique has been used in many fields. For example, in the educational context, [20], [24] applies Bayesian networks to infer learning styles.
- The last LA task is called “Create a motivational pattern”. This task builds a pattern of the elements to be used to motivate the potential deserter students. With this information, can be defined strategies, in order to avoid the desertion. This task uses the information generated by the second task (the rules), and generates a specific pattern for each rule, according to the attributes that compose them. For example, if one of the rules is a given career, the motivation pattern must stand out the relevance of the career, the main aspects of the career in the society, among other things. This task is the more complex of the autonomic cycle, because it must mix different semantic mining techniques, to extract information from different sources. For example, to use linked data to search specific information about the attributes of the rules, to use text mining to extract motivational sentences from the retrieved text, among other things.

4 Methodology for BI Projects in a LA Process Focused on the Student Behaviour

In this section, we use the methodology presented in the section 2.b, to describe the implementation of the LA tasks of our autonomic cycle, proposed in the previous section.

4.1 Step 1: Defining objective situations

The distance learning process is based on a virtual environment where teachers and students are in different places, and perhaps times [20], [21]. This process can be directed and adapted to the characteristics of the different students (student-centered), and has different interfaces and methodologies specific online learning. Some online virtual environments usually involve learning management systems (LMS), VLE, etc. In order to identify the behavior of students, we propose the following objective situations:

- Classify the students as deserters or not-deserters
- Identify the factors that influence a student's decision to abandon his studies,
- Predict the potential students to leave their studies.
- Build a motivational pattern to avoid the desertion.

Some features to consider of each student are:

- Degree
- Study Centre
- Gender
- Age
- Geographical area (Canton Province)
- Type of disability
- Size of the study centre (volume of students)
- Progress in the training program

Some important questions that should be able to answer are

- How does the academic performance of student influence the dropouts?
- Are there degrees whose dropout rates involving urgent decisions?
- Is it possible to predict when a student is likely to leave the studies?
- Does the geographic location of the student is a determining factor in the dropout?

In general, the indicators that allow us to characterize the objective situations relating to the dropout are:

- Dropout Rate of the careers
- Study centres with greater dropout rate
- Approval success rate
- Dropout Rate accumulated by level of progress
- Dropout Rate according to age range of students

4.2 Step 2. Data Model of the BI project

The first to define is the multidimensional dimensional model (conceptual view); and then the operational view, for which we develop ETL operations from the VLE environment of the distance learning institution.

4.2.1 Architecture

Figure 2 describes the general architecture, which describe the process to generate the data-warehouse. The data sources are the academic database of the University (e.g. UTPL (Universidad Técnica Particular de Loja)) and the database of the National Institute of Statistics. According to the indicators that allow us to characterize the objective situations, is defined the conceptual data model with the detailed academic record necessary to calculate these indicators. In the next step is extracted the information from the databases using query operations. This data can be improved during a transformation step, and finally are loaded in the multidimensional model. In this model are executed the OLAP operations or data mining tasks, in order to calculate the indicators.

Figure 2: Architecture to build the data-warehouse

4.2.2 Multidimensional model (star schema)

Figure 3 shows the dimensional design for our case study, which focuses on the history of the studies carried out for each student, characterizing each study completed based on several factors (dimensions). Figure 3 is the conceptual model (conceptual view). Thus, we define a star model in which the fact table contains indicators calculated for the objective situation, and pointers to different dimension tables, which are:

- Student: basic information about the student qualifications, etc.
- Study: contains information that describes the different courses taught at the institution.
- Program: information about the curriculum.
- Mesh: contains the plan of courses that the student has followed.
- Period: indicates the last academic term followed by the student.
- Centre: educational institution where the student regularly attends its studies.
- Canton: where the university is located
- Province: Canton to which it belongs.

In addition, Figure 4 shows the hierarchy of dimensions in our multidimensional model. We can see the different dimensions, and the aggregation of each one. For example, an educational centre is in a canton, which belong to a province.

Figure 3: Multidimensional Model (star schema)

Figure 4: Multidimensional Model (with hierarchy of dimensions)

4.2.3 ETL Operations

At this stage, we define the specific operations of (E)xtraction, (T)ransformación and (L)oads carried out on the operational database, required to build the operational view to be used in the following phase. These operations will feed with data to the multidimensional model.

To extract was used as source the academic system of the UTPL, specifically, information about enrolment and academic record for each student who has completed undergraduate studies. From there, we have generated a view of initial data of 32 attributes, from different tables in the operational databases. Some of these tables were:

- University centres: with information about the different university centres.
- Student: it contains information about the students.
- Academic program: information about the academic program
- Educational component (subject): information about the courses taken.
- Academic Result: contains information about the results obtained by the students.
- Academic period: information about the teaching periods 2,515,488 records were extracted, which are analysed, purged and summarized to feed the multidimensional scheme.

Additionally, public information provided by the INEC (*Instituto Nacional de Estadísticas y Censos*) served to have information about the territorial division and population in cantons and provinces.

The transformation involved the summarization of the data to get a single record for each study completed by a student, complemented with measures required in the fact table:

- F_ULT_ACTIVIDAD: estimated date of the last activity record
- AGE: age of the student's.
- T_UNID_PRA: number of credits or courses required to complete the studies.
- T_UNID_APROB: number of credits achieved by the student.
- PRC_AVANCE: level of student progress.
- NRO_ASG_REGISTRADAS: number of courses registered in the record.
- NRO_ASG_ACREDITADAS: number of courses passed by the student.
- NRO_ASG_CURSADAS: number of courses taken in regular course.
- NRO_ASG_APROBADAS: number of courses passed in regular course.
- NRO_INTENTOS_APROB: Total attempts to approve courses in regular periods.
- PROMEDIO_APROBACION: average of approved courses
- PROMEDIO_GENERAL: Average grades for courses passed or failed.
- PROMEDIO_HISTORICO_APROBACION: approval historical average in the career for the courses taken by the student.
- PROMEDIO_HISTORICO_GENERAL: historical average approval and disapproval in the career for the courses taken by the student.
- STATUS_CARRERA: current status of each completed study (completed, in progress, abandoned).

Figure 5 shows an example of a SQL operation applied, in order to determine the number of courses required to complete the studies (T_UNID_PRA). In this query, *datos* are the source of data, *variaciones* are the *catalog* of careers, *pra_id* is the id of careers, *ent_id* is the student id, *coe_id* is the id of the course, *tipo* is the type of career and *etr_nombre* is the approval status of the course. Similar queries were developed to fill out the data warehouse.

```
CREATE VIEW estad_carreras_asg
FROM
SELECT c.rownum, c.pra_id, c.t_asg, c.cantidad
FROM (
    SELECT
        IF (@fila = b.pra_id, @rownum :=@rownum + 1, @rownum := 1) AS rownum,
        @fila := b.pra_id AS fila, b.*
    FROM (
        SELECT a.pra_id, a.t_asg, COUNT(*) AS cantidad
        FROM (
            SELECT d.ent_id, d.pra_id, COUNT(DISTINCT d.coe_id) AS t_asg
            FROM datos d, variaciones v
            WHERE d.pra_id = v.pra_id AND v.tipo = "A" AND d.etr_nombre <> "REPROBADO"
            GROUP BY d.ent_id, d.pra_id
        ) a
        GROUP BY 1,2
        ORDER BY a.pra_id, a.t_asg DESC
    ) b, (SELECT @rownum:=0) x, (SELECT @fila := '') f
) c
WHERE c.rownum <= 6;

-- Estadística de número de asignaturas requeridas para culminar estudios
CREATE VIEW estad_asg_requeridas
FROM
SELECT eca.pra_id, MIN(eca.t_asg) AS t_unid_pra
FROM (
    SELECT ca.pra_id, MAX(ca.cantidad) AS cant_max
    FROM estad_carreras_asg ca
    GROUP BY ca.pra_id
) a, estad_carreras_asg eca
```

```
WHERE a.pra_id = eca.pra_id AND a.cant_max = eca.cantidad
GROUP BY eca.pra_id;
```

Figure 5. SQL Operations for the T_UNID_PRA indicator

To load the data to the multidimensional scheme, first were loaded the data of the facts, and then the data associated with its dimensions. Finally, were calculated the indicators in the facts table. Figure 6 shows the process of calculating and updating the approval average and general average (PROMEDIO_APROBACION, PROMEDIO_GENERAL) of each study completed. That process is similar for the rest of data to load.

```
CREATE VIEW tmp_notas_componente_carrera
FROM
SELECT x.*, t.nota
FROM (
    SELECT DISTINCT d.ent_id, d.pra_id, d.coe_id,
        IF (d.ETR_CODIGO in ('HOMOLOGADO','APROBADA','VALIDADO','REVALIDADO'),
            'ACREDITADO','NOACREDITADO') AS "PROMO"
    FROM datos d
    ORDER BY 1,3
) x, tmp_notas_componente_carrera t
WHERE x.ent_id = t.ent_id AND x.coe_id = t.coe_id AND x.PROMO = t.PROMO
ORDER BY x.ent_id, x.pra_id, x.coe_id;

CREATE VIEW estad_promedios_notas
FROM
    SELECT t.ent_id, t.pra_id,
        round(ifnull(avg(IF(PROMO='ACREDITADO',t.nota,NULL)),0),2) AS "PROMEDIOAPROBACION",
        round(avg(t.nota),2) AS "PROMEDIIOGENERAL"
    FROM tmp_notas_componente_carrera t
    GROUP BY 1,2;

UPDATE c_h_estudios_cursados h
SET h.promedio_aprobacion = (SELECT e.promedioaprobacion
    FROM estad_promedios_notas e
    WHERE e.ent_id = h.ent_id AND e.pra_id = h.pra_id)
h.promedio_general = (SELECT e.promediogeneral
    FROM estad_promedios_notas e
    WHERE e.ent_id = h.ent_id AND e.pra_id = h.pra_id);
```

Figure 6. SQL Operations to load averages

4.3 Step 3. Knowledge extraction (indicators)

At this stage, we characterize all the process of calculation of indicators, required to determine the two objective situations. Some of them are based on OLAP operations, or data mining tasks. The data mining tasks are specified in the following section.

4.3.1 OLAP operations required for the calculation of indicators.

Table 1 shows the OLAP operations required to calculate some of our indicators.

Table 1 OLAP operations required

Indicator	OLAP operation	Description
Careers dropout rate	Roll-up	This is the relation between the total of dropout students, with respect to the total of enrolled students, grouped by their fields.
University support centers with greater dropout rate	Roll-up	This is the relation between the total of dropout students, with respect to the total of enrolled students in each university center, sorted in descending order.
Dropout rate by age range of students	Roll-up	This is the relation between the total of dropout students versus the total of enrolled students, by each range of age
Cumulative dropout rate by level of progress	Roll-up	This is the dropout rate for each level of study (the first level is 0%, and 100% when the students finish their careers).

Figure 7 shows the roll-up operation for the "Careers dropout rate" indicator, and Table 2 the result limited to 10 careers with the highest rate of dropout.

```
SELECT t.tit_nombre AS "Titulacion", round((y.t_est_des / x.t_est),2) AS "IndiceDesercion"
FROM c_d_titulaciones t,
  (SELECT h.tit_id, count(distinct h.ent_id) AS t_est
   FROM c_h_estudios_cursados h
   GROUP BY h.tit_id) x,
  (SELECT h.tit_id, count(distinct h.ent_id) AS t_est_des
   FROM c_h_estudios_cursados h
   WHERE h.status_carrera = 'ABANDONADO'
   GROUP BY h.tit_id) y
WHERE t.tit_id = x.tit_id AND t.tit_id = y.tit_id
ORDER BY 2 DESC;
```

Figure 7. SQL operation for "Careers dropout rate"

Table 2 Careers with high dropout rates

Careers	Desertion Index
CIENCIAS DE LA EDUCACIÓN MENCIÓN CIENCIAS HUMANAS Y RELIGIOSAS	0.63
INFORMÁTICA	0.54
CIENCIAS DE LA EDUCACIÓN MENCIÓN QUÍMICA Y BIOLOGÍA	0.51
CIENCIAS DE LA EDUCACIÓN MENCIÓN LENGUA Y LITERATURA	0.48
INGENIERÍA EN ADMINISTRACIÓN DE EMPRESAS TURÍSTICAS Y HOTELERAS	0.48
CIENCIAS DE LA EDUCACIÓN MENCIÓN FÍSICA Y MATEMÁTICA	0.47
COMUNICACIÓN SOCIAL	0.47
ECONOMÍA	0.47
CIENCIAS DE LA EDUCACIÓN MENCIÓN INGLÉS	0.47
ADMINISTRACIÓN DE EMPRESAS	0.45

Also, OLAP operations were defined from the cube, to generate the minable view for data mining tasks, which includes the following variables:

- UltimoAño: Last year of study in which the student activity is recorded.
- TasaEfectividadAprobacion: % students passing academic components.
- AlcanceHistoricoAprobacion: Percentage ratio of the average student approval with respect to the historic average approval in the career.
- AlcanceHistoricoGeneral: Percentage ratio of the overall average student on the general historical average in the career.

Figure 8 shows an example of an OLAP operation applied on the cube to generate the minable view.

5 The knowledge extracted using our Autonomic Cycle

In this section we present the models to analyse the dropout problem, using our Autonomic Cycle of tasks of LA. Additionally, we analyse the quality of the models and results. In order to instance the Autonomic Cycle of tasks of LA that must be applied, the context of the educational institution at which distance education is imparted was analysed.

For the implementation of the predictive model, we have used Java applications, which invoke the Weka libraries. For the descriptive model, the "arules" and "igraph" packages were used. These packages belong to R software. For the implementation of the classification model and the construction of the motivational pattern, we have developed a Java application. For the development of the LA tasks, we are used a combination of the methodologies proposed in [5], [6] called MIDANO.

5.1. Classify Students as Deserters and non-Deserters

In this case, we have developed an application, which read the register from the data warehouse, and according to the "STATUS_CARRERA" value, it stores this register in the "deserter" or "not-deserter" classes. "STATUS_CARRERA" is a class label, because it defines the status of each completed study (completed, in progress, abandoned).

In this way, this first LA task creates two groups from the set of the student registers (“deserter” or “not-deserter”) using the value of the “STATUS_CARRERA” attribute. This LA task provides the source of information for the rests of LA tasks of our autonomic cycle, because it defines the target group that will be analysed for the next tasks. This group must be described, characterized, in order to discover useful knowledge about the deserter students. The classifier algorithm is very simple, it compares the value of the “STATUS_CARRERA” to determine where will be stored the students. We do not require the evaluation of performance of this algorithm, because the classifier model considers only the value of this attribute for the classification (the model is not complex).

5.2. Identify the factors that influence a student's decision to abandon his/her studies

Table 3. Some rules and metrics obtained

To understand the graph of the rules obtained, we carry out the next comprehensive analysis of the results:

The main conclusion of this task is that the descriptive model based on association rules allowed to characterize adequately defectors students and establish their profile. In general, it is not possible to define a unique profile that describes all dropped out students. We have selected models with low levels of support (ranging between 0.5% and 2 %), but with confidence levels higher than 85%, i.e. applicable rules to specific groups of students but with an effectivity of success very high. The results are very revealing. First, it has been possible to identify indicators that are crucial in the characterization of dropped out students, these are:

- The subject in which each student registered her/his worst performance (ComponentePeorRendimiento)
- The effectiveness of approval, that is, the existing ratio of the number of subjects passed and the total number of attempts of passing. (TasaEfectividadAprobación)
- The average scores of each student (PromedioGeneral , PromedioAprobacion)
- The percentage difference of individual student performance compared to the average historical performance and major (AlcanceHistoricoGeneral , AlcanceHistoricoAprobacion)
- The career (Titulacion) with problems

Then, the obtained rules have allowed to reveal important information in order to decision making. Here are some cases:

- When the lowest student performance is associated with subjects such as: MATEMÁTICAS I, CIENCIA PENAL, DERECHO ROMANO I, COMPUTACIÓN BÁSICA, LÓGICA MATEMÁTICA, among others, it involves desertion in almost all cases (over 97 %). That is, the failure of those subjects leads the student to question his/her continuity in the career chosen.
- In the case of “CIENCIAS DE LA EDUCACIÓN MENCIÓN CIENCIAS HUMANAS Y RELIGIOSAS” major in Loja, the dropout rate is 96%. It corroborates the facts described in Table 2.
- For the “Ciencias de la Educación Mención Educación Básica” career, if the student's overall average is less than 3, it involves drop out in 91 % of cases.
- Students of “CIENCIAS DE LA EDUCACIÓN MENCIÓN INGLÉS” career, who have failed to pass any subject at the beginning, they will end up leaving in 86 % of cases.
- Students whose worst performance is in the subject of "REALIDAD NACIONAL Y AMBIENTAL" , which effectiveness of approval is between 19 % and 29 %, whose average approval total is ranging from 28 to 32 over 40, leave school at 91% of the cases .

And so, there are more than 50 association rules, each one describing a specific reality with a very high confidence. Particular realities that can be very useful for managers in each area, due to they reflect the conditions that can lead students to take the decision to drop out the career.

Each rule defines a relationship between variables, which can be used to make decisions, because each one identifies a problematic situation. For example, the career “CIENCIAS DE LA EDUCACIÓN MENCIÓN CIENCIAS HUMANAS Y RELIGIOSAS” in Loja has very important problems to keep the students. It is necessary a depth analysis of this career, in order to determine the reasons of students desertion.

All these rules are new knowledge, not shown so far, which can be the basis to optimize university efforts in order to reduce dropout rates.

5.3. Prediction of potential dropout students

With this LA task, the idea is to create a predictive model that, based on the known characteristics of each student (these attributes were previously mentioned), forecasts potential dropout students. Solving this problem is essential since, by using this model, potential dropout students can be identified, and this information can be useful for implementing programs to assist these students. A system of such nature is a valuable tool to accomplish the mission of the University, which involves helping students who are going through extreme situations in their teaching-learning processes. On these students will be applied the strategies derived from the motivational patterns built by the last LA task.

For the experimental part, it was necessary to establish a processing phase, in which the problem of student dropout in the early years of study could be identified. In this way, our model is focused on students in the first semesters, specifically, those students whose level of progress in their studies is lower than 10%. Additionally, some attributes that do not contribute to the results of the models, or have a high correlation coefficient, were deleted.

After that, we applied Bayesian networks, obtaining satisfactory results. First of all, we tried to create a universal model, that is, a model that can be applied to our entire data universe (150.000 instances). However, due to the particular characteristics of the data, it was impossible to obtain acceptable results with a universal model. For this reason, we decided to create multiple models that can be applied to sets of specific data segmented under certain conditions. Table 4 shows a sample of the results obtained. The experiments have been conducted by using 80% of the information for training and 20% for tests.

Table 4. Random sample of 8 models obtained

N.	Excluded Attributes	Conditions	Instances	Precision	Recall
1	None	2;14,12,2,1,52 4;5,2 6;1,2	2529	0.6965	0.8947
2	5	4;1,2,4 6;2	1637	0.7108	0.8439
3	None	4;5,2,4 6;1,2	4490	0.6971	0.8422
4	5	4;5,1 6;1	1869	0.6708	0.8278
5	5	4;5,1,2 6;2	2094	0.8351	0.7403
6	None	4;5,1,2,4 6;2	2555	0.8500	0.7315
7	None	2;2,52 4;5,2,4 6;1,2	1016	0.8959	0.7279
8	None	4;1,4 6;2	1055	0.8477	0.7193

Table 4 represents a sample of eight models obtained. For the experiments, the attributes used are the defined in the previous section. The second column indicates if some of these attributes are deleted to build the models. The attributes deleted were: "Titulacion", "UltimoAnio", "Centro", and "NroEstudiantesCentro", "ProvinciaNacimiento". Additionally, the third column describes the filters used for several attributes. For example, the fifth model indicates that 2 of the attributes explained in the previous section, have been used as filters in the training: "Provincia" and "Población". These attributes are located in positions 4 and 6, respectively. Thus, this model is specifically applicable to data that have two conditions coded in the algorithm: $\langle\langle 4;5,1,2 \rangle\rangle$ and $\langle\langle 6;2 \rangle\rangle$. Each condition has two parts. The first one is related to the attribute to which the filter is applied, and the second one (separated by commas) is related to the values that must be kept in that attribute. For the previous example, the first condition: $\langle\langle 4;5,1,2 \rangle\rangle$ indicates that a filter must be applied to attribute 4 ("Provincia") that must have the values 5,1,2; which are associated to the values of the provinces of Guayas, Loja and Azuay in Ecuador (see Figure 10). The second condition $\langle\langle 6;2 \rangle\rangle$ indicates that, from the attribute "Género" (6), we have selected the students of the male gender, that is, "Masculino" (2). This process is carried out for the rest of the tests.

Figure 10. Training using Weka

Due to the conditions of the problem, the metrics of Precision and Recall were defined as the appropriate metrics in this context. In addition, considering that the rates of student dropout can be reduced by implementing programs to assist and motivate most of the students who are prone to dropping out (even though this means the inclusion of students without intentions of leaving them behind), we consider that the Recall metric must be the most important since it allows us to decrease the student dropout rate by measuring the number of examples correctly classified as positive from the total number of positive examples on the database. Consequently, those models with Recall metrics higher than 0.70 and Precision metrics higher than 0.65 have been selected.

The algorithm developed has generated numerous models, from which 110 models meet the conditions of Precision and Recall previously indicated, and cover approximately 85% of the student population under study. Similarly with the experienced with descriptive model, it is not possible to predict the potential deserter students. The main conclusion of this LA task is that it is not possible to achieve a universal model to predict the dropout in every possible scenario, so various models have been required, referring to a particular context (specific values of attributes). In that sense, they were randomly selected some of the models with acceptable indices of precision and recall (greater than 65 % and 70 %, respectively), and with them, tests were applied using samples of historical data extracted from an academic database, checking that indeed the predictions of desertion are correct, between 70% and 85%, with a level of not predicted cases that did not exceed the 30%.

Figure 11 shows a test scenario made with historical data applied to the Faculty of Economics, the dropout rate amounts to 47 % (see Table 2). A sample of 40 students that was validated by one of the models obtained, was taken. At this stage, 19 cases were deserters and 21 were not deserters. The predictive model was applied, and as a result, 17 dropout predictions were obtained which 14 were successful (true positives) and 3 were wrong (false positives). Thus, an accuracy of 82.35 % (14/17) and a recall of 73.68 % (14/19) was obtained. The main metric is recall, that is the metric that must be improved in our model because it determines the number of students classified as a drop out (they need to be supervised). Recall determines the set of students to motivate, in order to avoid their drop out. We need to improve this model in order to reach 100% in this metric.

Figure 11. Test scenario for the ECONOMY career.

5.4. Creation of a motivational pattern to avoid the desertion

In this last LA task, we seek to build diverse motivational patterns for the different potential deserter students identified in the previous phase. At experimental level, we relax the conditions used in the second task, at the level of the support and trust required by the rules to be used in this task (in this case will be 0.7% for support and 77% for trust), in order to get more rules to construct motivational patterns. This allowed us to obtain a greater number of rules than the 18 rules obtained in the second task, almost 350 rules, to build the motivational patterns (see Figure 12). This set of rules covers approximately the 50% of the universe of deserter students. We have reduced a lot of the support and trust to cover an important part of the population of deserter students.

Figure 12. Support and confidence/trust for the generated rules for the last LA task

That is, given that the support represents the number of records covered by the rules, we have relaxed the value of the support (see Figure 12). Now, we have obtained 345 rules, which cover more than 40% of the total data. With this set of rules, we build the motivational pattern. In this first version of this LA task, our motivational pattern is composed of the next attributes:

- The importance of the career for the society
- Potential learning resources for the topics with problems
- Relationship of the career with the gender, or the region

Based on these attributes, we propose a set of information about each one, to be used by the university, in order to motivate the students to continue their career. Particularly, there is a semantic enrichment of the patterns, in this case, using the information from the Internet, and the attributes of the rules of the previous tasks as the criteria of searching. For this search, we have used the tool proposed in [34], which allows natural language queries on the Internet.

The university uses this information, to define different mechanisms to motivate the interest of the students. The table 5 shows an example of the information recovered from the Internet using [34], and provided to the university. With this information, the university can define a set of strategies, in order to motivate the students.

Table 5. Attributes extracted from the rules, and information recovered by them.

Attributes	Information	Strategies
Gender=FEMENINO, Topics=ANTROPOLOGÍA	https://es.wikipedia.org/wiki/Antropologia_feminista “La antropología feminista y la categoría género”, Marta Lamas, Nueva Antropología. Revista de Ciencias Sociales, 1086	Links to review, about the role of the womens in anthropology Reading for the discussion
Topics =COMPUTACIÓN BÁSICA	http://www.idea.edu.pe/huaral/curso01/03.shtml http://www.maestrasjardineras.com.ar/informatica1.html http://www.istvidanueva.edu.ec/bibliotecavirtual/phocadownload/Libros/completo%20manual%20de%20informatica.pdf	Websites with basic computer courses and Basic computer books, in order to be used by the students as support to their learning processes.

Career=CIENCIAS DE LA EDUCACIÓN BÁSICA,	DE LA MENCIÓN	http://www.cursosycarreras.com.ec/ciencias-de-la-educacion-mencion-educacion-basica-el-oro-machala-umet-FO-144026 http://www.ula.ve/estudios/carreras/nucleos/nucleo-trujillo/educacion-mencion-basica-integral	Websites with information about the career, its importance, its relevance in the society, to be discussed with the students
---	---------------	--	---

This last LA task, must be improved in the next works, with the "linked data" paradigm, to search specific information in Internet, based on the idea of the "data navigation" allowed with this technique, which can be executed using the attributes determined by our rules. Additionally, the graph mining techniques can be used, in order to exploit the connectivist learning paradigm, which allows linking students with other students, with sources of learning, in order to promote the collaboration and the integration of learning communities. All these aspects must motivate the students to continue their studies in an agreeable environment with other partners.

6 Conclusions

In this work, the dropout problem is studied in distance education institutions, with exploiting organizational data. For that, we have proposed an integral approach to attack this problem, based on LA tasks, which we are calling autonomic cycle, considering the different aspects to be considered during the analysis of this problem: to observe the target students, to analyse the causes, to predict potential deserters, and to implement motivational strategies for them. Our autonomic cycle allows a permanent supervision of the academic environment, to avoid the desertion of students.

In particular, the LA tasks are defined in order to build models to predict potential students with the problem and the factors that affect them, or patterns which can be used to motivate the students. Specifically, the LA tasks analyse the behaviour of students. To specify the LA tasks we have used a BI methodology, which was adapted to allow building the minable view required by the LA tasks, and the introduction of the concept of autonomic cycle of LA tasks. The different models built, of description, classification, prediction and the motivational pattern, allow to analyse in depth the dropout problem, because they can provide information to attack two of the main reasons for students dropping out: lack of educational support, student with special needs (They require special attention), etc.

This is the first approach that combines BI and LA paradigm, and which integrate a set of LA tasks in a autonomic cycle, in order to permanently analyse the dropout problem in an educational institution. We use these paradigms simultaneously, to define data models and data mining tasks for the specific situations detected as target goals. In [1], [27], [28], [29], [30], [31], [33] define LA tasks to understand the different factors on student's dropout incidents. In our paper, we analyse this problem, and concluded that we cannot build a universal model of desertion, and it is required a model of desertion for each context. In addition, we define an additional model to predict the desertion of the students. This result is very important, because defines the set of student to be supervised, motivated, etc., in order to reduce the rate of desertion. In this way, our approach enriches the previous results with this capability of prediction.

With respect to the knowledge generated, our LA tasks generate different knowledge models with different types of information to be exploited. One model can predict the desertion rate of the student, in different situations. Particularly, the main knowledge generated is the set of students which maybe drop. The classification model defines the target information to be used to extract knowledge. The descriptive model defines a set of relationships between variables, which must be analysed in order to understand the desertion phenomenon. For example, there are careers in certain locations with a high level of desertion, it is necessary analysis these careers in these locations, to improve the educational conditions. This is the type of analysis to carry out with each rule generated. Finally, with the motivational pattern we can extract information from the Internet to semantically enrich the strategies to attack the desertion in an academic context.

Future efforts should be devoted to analyse other potential aspects linked to dropout in the last task, such as the environment, which will eventually have to apply semantic mining techniques in the tasks of LA. Also, we must specify the precise actions in the process of learning, to improve student performance, or adapt it to their needs and requirements (centered on the student). In this work, we have obtained very interesting preliminary results, which motivate these future work, for converting such "knowledge as a service" to optimize teaching and learning processes. Specifically, three fundamental aspects remain to be carried out:

- We should expand the operational minable view (add more data) to improve the identified scenarios and therefore the predictive models. Thus, we would have a broader data for the training phase, which should lead to better recall and precision values for the models.

- We should start using these models in a management system of open and distance learning, in order to identify potential cases of desertion among students, especially among those enrolled in the first cycles. They help managers give a personalized follow-up and individualized specific measures for a student or a group of them. The effectiveness and real benefit of these predictive models can be seen when the dropout rates are reduced in each career.
- We should improve the implementation of the last task, with sophisticated semantic mining techniques, as web mining, graph mining or linked data, in order to exploit the information available in the Internet, to enrich the motivational strategies.

Acknowledgements

Dr. Aguilar has been partially supported by the Prometeo Project of the Ministry of Higher Education, Science, Technology and Innovation (SENESCYT) of the Republic of Ecuador.

References

- [1] G. Riofrio, E. Encalada, D. Guamán, J. Aguilar, “Aplicación de un general Learning Analytics framework para analizar la deserción estudiantil en carreras a distancia”, *Proc. Latin American Computing Conference (CLEI)*, pp. 567-576, 2015.
- [2] M. Chatti, A. Dyckhoff, U. Schroeder and H. Thüs. "A reference model for learning analytics". *International Journal of Technology Enhanced Learning (IJTEL)*, vol. 4, pp. 318-331, 2012.
- [3] M. Sánchez, J. Aguilar, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán, L. Chamba-Eras, “Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service”, *New Advances in Information Systems and Technologies*, Springer, pp. 993-1002, 2016.
- [4] R. Ferguson, “Learning analytics: drivers, developments and challenges”. *International Journal of Technology Enhanced Learning*, vol. 4, pp. 304–317, 2012.
- [5] C. Rangel, F. Pacheco, J. Aguilar, M. Cerrada, J. Altamiranda, “Methodology for detecting the feasibility of using data mining in an organization”, *Proc. Latin American Computing Conference (CLEI)*, pp. 502-513, 2013.
- [6] F. Pacheco, C. Rangel, J. Aguilar, M. Cerrada, J. Altamiranda, “Methodological framework for data processing based on the Data Science paradigm”, *Proc. Latin American Computing Conference (CLEI)*, pp. 1-12, 2014.
- [7] A. Bara, I. Botha, V. Diaconita, I. Lungu, A. Velicanu, and M. Velicanu. "A model for Business Intelligence Systems", *Development, Informatica Economică*, vol. 13, pp. 99-108, 2009.
- [8] S. Parack, Z. Zahid, and F. Merchant, “Application of data mining in educational databases for predicting academic trends and patterns,” *Proc. 2012 IEEE International Conference on in Technology Enhanced Education (ICTEE)*, 2012, pp. 1–4.
- [9] S. Anupamar, M. Vijayalakshmi, "Efficiency of Data Mining Techniques in Edifying Sector", *International Journal of Advance Foundation and Research in Computer* , vol. 1, no. 8, pp 56-62, 2014.
- [10] C. Rangel, J. Altamiranda, J. Aguilar, "Semantic mining in clusters from signaling pathways networks", *Proc. Latin American Computing Conference (CLEI)*, pp. 297-306, 2015.
- [11] R. Baker, P. Inventado, “Educational data mining and learning analytics,” in *Learning Analytics*, Springer, 2014, pp. 61–75.
- [12] G. Wolfgang; D. Hendrik. "Translating Learning into Numbers: Toward a Generic Framework for Learning Analytics", *Educational Technology and Society*, vol. 15, pp. 42–57, 2012.
- [13] G. Siemens, “Learning analytics: envisioning a research discipline and a domain of practice,” *Proc. 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 4–8.
- [14] S. Buckingham, R. Ferguson. “Social Learning Analytics”. *Educational Technology & Society*, vol. 15, no. 3, pp. 3-26, 2012.

- [15] T. Ranbaduge, "Use of Data Mining Methodologies in Evaluating Educational Data", *International Journal of Scientific and Research Publications*, vol. 3, no. 11, 2013.
- [16] J. Ari, B. White, "Learning Analytics From Research to Practice", Springer, Media New York 2014.
- [17] S. Iglesias-Pradas, C. Ruiz-de-Azcarate, A. Agudo-Peregrina, "Assessing the suitability of student interactions from Moodle data logs as predictors of cross-curricular competencies", *Computers in Human Behavior*, vol. 47, pp. 81–89, 2015.
- [18] P. Munoz-Merino, J. Ruiperez-Valiente, C. Alario-Hoyos, M. Perez-Sanagustin, C. Delgado, "Precise Effectiveness Strategy for analyzing the effectiveness of students with educational resources and activities in MOOCs", *Computers in Human Behavior*, vol. 47, pp. 108–118, 2015.
- [19] C. Conati, A. Gertner, K. Vanlehn, "Using Bayesian networks to manage uncertainty in student modeling," *User Model. User-adapt. Interact.*, vol. 12, no. 4, pp. 371–417, 2002.
- [20] M. Jopling, "1:1 online tuition: a review of the literature from a pedagogical perspective," *Journal of Computer Assisted Learning*, vol. 28, pp 310–321, 2012.
- [21] O. Kozar, "The use of synchronous online tools in private English language teaching in Russia", *Distance Education*, vol. 33, pp. 415-420, 2012.
- [22] T. Nielsen, F. Jensen, "Bayesian networks and decision graphs". Springer, 2009.
- [23] P. García, A. Amandi, S. Schiaffino, M. Campo, "Evaluating Bayesian networks' precision for detecting students' learning styles," *Comput. Educ.*, vol. 49, no. 3, pp. 794–808, 2007.
- [24] C. Borgelt and R. Kruse, "Induction of association rules: Apriori implementation," *Proc of Compstat: Computational Statistics*, 2002, pp. 395–400.
- [25] S. Kotsiantis, C. Pierrakeas, P. Pintelas, "Efficiency of Machine Learning Techniques in Predicting Students' Performance in Distance Learning Systems", Technical Report TR 02-03, Division of Computational Mathematics and Informatics, Dept. of Mathematics, University of Patras, 2002.
- [26] B. Franklin, S. Trouard, "An Analysis of Dropout Predictors within a State High School Graduation Panel", *Schooling*, vol. 5, no. 1, pp. 1-8, 2014.
- [27] E. Yukselturk, S. Ozekes, Y. Türel, "Predicting dropout student: an application of data mining methods in an online education program", *European Journal of Open, Distance and e-Learning*, vol. 17, no. 1, pp. 118-133, 2014.
- [28] L. Paura, I. Arhipova, "Cause Analysis of Students' Dropout Rate in Higher Education Study Program", *Procedia: Social and Behavioral Sciences*, vol. 109, pp. 1282-1286, 2014.
- [29] A. Sang, P. Koros, J. Bosire, "An Analysis on Dropout Levels of Public Secondary Schools in Kericho District in Relation to Selected School Characteristics, *International Education Studies*; vol. 6, no. 7; pp. 247 – 259, 2013.
- [30] M. Jadrić, Ž. Garača, M. Čukušić "Student dropout analysis with application of data mining methods", *Management*, vol. 15, no. 1, pp. 31-46, 2010.
- [31] G. Dekker, M. Pechenizkiy, J. Vleeshouwers, "Predicting Students Drop Out: A Case Study", *Educational Data Mining*, pp. 41-50, 2009.
- [32] J Aguilar, P. O. Buendia, K. Moreno, D. Mosquera, "Autonomous cycle of data analysis tasks for learning processes", In *Technologies and Innovation* (R. Valencia-García, et al., Eds.), *Communications in Computer and Information Science*, Vol. 658, Springer, pp. 187-202, 2016.
- [33] J Aguilar, P. Valdiviezo, J. Cordero, G. Riofrio, E. Encalada "A General Framework for Learning Analytic in a Smart Classroom", In *Technologies and Innovation* (R. Valencia-García, et al., Eds.), *Communications in Computer and Information Science*, Vol. 658, Springer, pp. 214-225, 2016.