sustainability

MDPI

Article

A Business Intelligence Framework for Analyzing Educational Data

William Villegas-Ch [1,*], Xavier Palacios-Pacheco [2] and Sergio Luján-Mora [3]

[1] Escuela de Ingeniería en Tecnologías de la Información, FICA, Universidad de Las Américas, 170125 Quito, Ecuador

[2] Departamento de Sistemas, Universidad Internacional del Ecuador, 170411 Quito, Ecuador; xpalacio@uide.edu.ec

[3] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, 03690 Alicante, Spain; sergio.lujan@ua.es

* Correspondence: william.villegas@udla.edu.ec; Tel.: +593-098-136-4068

Received: 28 May 2020; Accepted: 13 July 2020; Published: 17 July 2020

Abstract: Currently, universities are being forced to change the paradigms of education, where knowledge is mainly based on the experience of the teacher. This change includes the development of quality education focused on students' learning. These factors have forced universities to look for a solution that allows them to extract data from different information systems and convert them into the knowledge necessary to make decisions that improve learning outcomes. The information systems administered by the universities store a large volume of data on the socioeconomic and academic variables of the students. In the university field, these data are generally not used to generate knowledge about their students, unlike in the business field, where the data are intensively analyzed in business intelligence to gain a competitive advantage. These success stories in the business field can be replicated by universities through an analysis of educational data. This document presents a method that combines models and techniques of data mining within an architecture of business intelligence to make decisions about variables that can influence the development of learning. In order to test the proposed method, a case study is presented, in which students are identified and classified according to the data they generate in the different information systems of a university.

Keywords: business intelligence (BI); educational data mining (EDM); learning management systems (LMS); learning analytics

1. Introduction

Currently, the use of information and communication technologies (ICTs) is included in all the activities of society. Universities are not far behind, and include ICTs in most of their processes. These processes integrate the administrative management on which the existence of the universities depends or use them as support for academic management [1]. The most extended use of ICTs for academic management is the learning management system (LMS) [2] that supports online interaction between teachers and students. However, there are scenarios in which specific support by ICTs is needed to solve common problems centered on learning. These scenarios allow ICTs to apply new models and educational methods in student learning. A guide to this may be the personalization that companies have achieved with their customers through data analysis models that allow managers, executives and analysts to discover trends and improve the services and products they offer to their customers.

Personalized service can be introduced to educational environments where the process is similar to that applied at the business level, but the objective in education is to improve the methods or activities that generate learning in students [3]. Learning environments are based primarily on a range of interactive and delivery services. Personalized learning recommendation systems can provide learning recommendations to students based on their needs [4,5]. Companies use data analysis architectures whose results help them make decisions about their business. These architectures are called business intelligence (BI); their ability to extract data from different sources, process them and turn them into knowledge is a solution that can be also included in the educational management of a university [6].

As a precedent, it is important to consider that several universities use a BI platform with an administrative or operational focus, which helps them make decisions in the financial management of the institution [7]. In the same way, previous works [8,9] have performed an analysis of desertion rates considering models and statistical tools with the use of economic and academic variables, segmenting the analysis into whether students have enrolled or not in the next semester. This formula is perfectly valid; however, it leaves aside the causes that determined why students abandon their studies. By contrast, our proposal is differentiated by its ability to analyze the data of students' academic activities and focus on the learning problems they present. This analysis helps make decisions in educational management and the improvement of the learning methods established by teachers [10].

In this work, three research questions are proposed that help to align the concepts and processes in their design; in addition, they seek to establish the current situation of the environment where this work is carried out:

- Is the execution of a BI architecture viable for the analysis of academic data?
- What are the difficulties that can be found in the implementation of a BI architecture?
- In addition, do the results coincide with an educational reality that collaborates with learning?

To answer each of these questions, this work includes the description of a BI framework that bases its design on a detailed review of the previous works, the Unified Modeling Language (UML) diagram and a complete method for applying academic data mining. This work extracts data from various academic sources, processes them and allows us to identify, through data mining algorithms, the strengths and weaknesses of each student. Once the results are obtained, knowledge is generated about the learning process of each student, allowing appropriate decisions to be made in order to improve the way the student learns.

This article is organized as follows: Section 2 reviews the existing work relating to the purpose of this study; Section 3 describes the components and processes of the proposed framework; Section 4 applies the method to a case study, to test the feasibility of the method; and Section 5 presents the conclusions.

2. Related Work

The literature review presented follows the guidelines published in the methodology for systematic literature review proposed by Kitchenham et al. [11] and by Petersen et al. [12]. Kitchenham et al. describe how the results of a literature review in software engineering should be planned, executed and presented; Petersen et al. provide a guide on how to conduct a rigorous review of the literature and follow a systematic procedure. For our literature review, the works were grouped according to the type of tool, model, paradigm or discussion they employ in their own analysis of educational data. For this type of classification, it was necessary to know the status of scientific work in learning environments that include the use of BI techniques that improve education. The objective of this literature review is to try to learn how they do it, and what methods and techniques they use. The search string "business intelligence AND education" was selected, and only documents published during the last 5 years were considered.

The searches were made based on the information provided in the title, abstract and keywords of the works. From the selected works, a detailed reading of the introduction and conclusions was performed, in order to filter out the non-related publications.

The inclusion and exclusion criteria used for the selection of documents were the following:

1. Journal article: the first selection criterion is that the papers considered are journal articles, and that they are written in English.
2. Out of context: this criterion applies to works that do not directly relate to the research topic. These documents met the search criteria because the terms were mentioned in the text, but their contributions did not relate to the subject of this study.
3. Duplicated: this criterion excluded all the works whose contribution was already included in the selection of another database.

Figure 1 represents the flowchart of the bibliography selection process; the first phase collects the articles from the online databases. The string terms used to search for articles in online databases, such as Springer Link, Web of Science, ACM Digital Library, IEEE Digital Library (Xplore) and Scopus, are found in Table 1. In the selection process, each of the articles was analyzed according to the guidelines that must be met for the design of a BI. In the next stage, we reviewed the works that included data mining applications. This filter was applied because a BI platform integrates data mining algorithms that generate knowledge of the analyzed data. These articles then passed to the classification stage and, finally, were integrated as valid literature for the study. Works that did not meet the conditions defined in the selection were automatically excluded from the process.

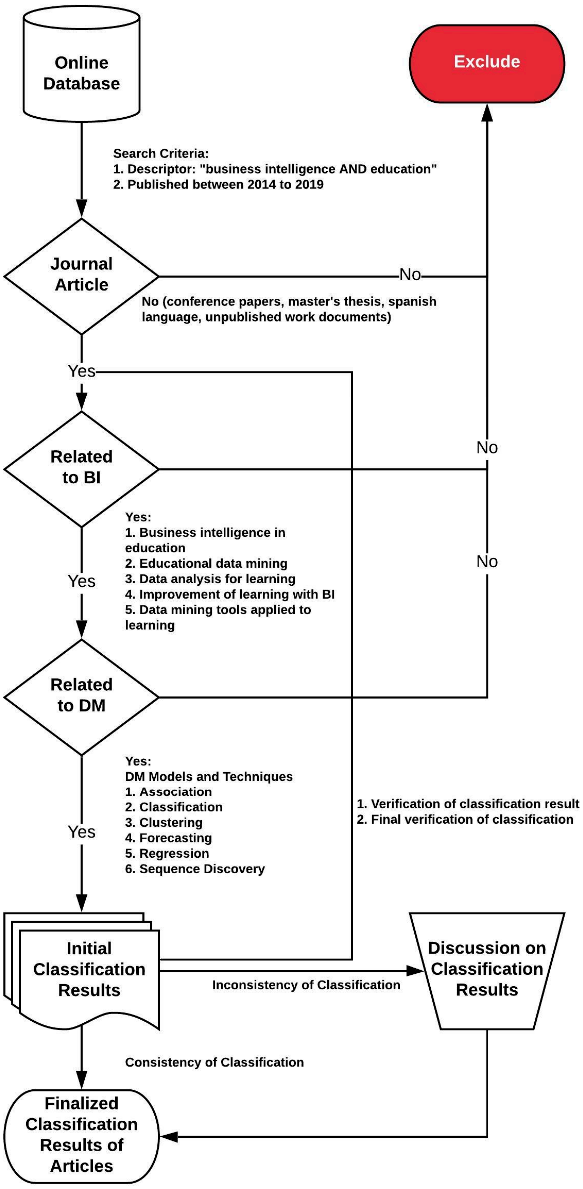
Online Database
Search Criteria:
1. Descriptor: "business intelligence AND education"
2. Published between 2014 to 2019
Exclude
Journal Article
No
No (conference papers, master's thesis, spanish language, unpublished work documents)
Yes
Related to BI
No
Yes:
1. Business intelligence in education
2. Educational data mining
3. Data analysis for learning
4. Improvement of learning with BI
5. Data mining tools applied to learning
Yes
No
Related to DM
Yes:
DM Models and Techniques
1. Association
2. Classification
3. Clustering
4. Forecasting
5. Regression
6. Sequence Discovery
Yes
1. Verification of classification result
2. Final verification of classification
Initial Classification Results
Inconsistency of Classification
Discussion on Classification Results
Consistency of Classification
Finalized Classification Results of Articles

Figure 1. Flow diagram for the selection of the bibliography.

Table 1. Bibliographic search string terms.

Source	Terms of Search	Scope of the Search
Springer Link	Query = "business+intelligence"+education	Only articles in English; published between 2014–2019
Web of Science	TS = ("business intelligence" AND education)	Refined by: years of publication: (2019 Or 2018 Or 2017 Or 2016 Or 2015 Or 2014); and types of documents: (Article)
ACM Digital Library	(+"business intelligence" +education)	Only articles in English; published between 2014–2019
IEEE Digital Library (Xplore)	(business intelligence and education)	Only articles in English; published between 2014–2019
Scopus	TITLE-ABS-KEY ("business intelligence" AND education)	Only articles in English; published between 2014–2019

The works were classified according to the type, contribution and scope of the research. The articles were classified according to the type of research based on the processes proposed in [11] and [13], prioritizing articles in which the proposed solution to a problem is innovative or a significant extension of an existing technique. Obtaining the results of the review began with the location of the primary studies, then moved on to the extraction of the data and, finally, the categorization and the resulting scheme.

In the first search conducted of the five academic sources, 2138 documents related to the terms used in the search string were obtained. These works were filtered using several additional parameters to limit the area of action of the topics. The first filter applied was based on requiring that the search string be present in the title, as well as in the keywords. A brief review of the abstracts of the most promising articles related to the search was made, considering 169 documents. From these articles, an exhaustive review of the introduction and conclusions was performed in order to identify its guidelines and the relationship between BI and learning in universities. After this process, 20 articles were selected; these works were taken as a basis for the development of certain areas of this study.

Table 2 presents the articles considered as the basis of this study; they are classified according to the type of study. The 'Type of study' column depends on the proposal of the work; the different types of study than have been identified are the works that present an application, a model or a case study. The 'Degree of incidence' column helps to classify the works that are directly related to the study. A high level indicates that the work are closely related to the application of an integrated BI model in education. The average level indicates that the work is related to the theme when deploying a BI model applied to a specific part of an educational environment. The low level contains the works that address the issue of BI in education, but in its development, and which do not contribute a model, architecture or framework that contributes to the analysis of educational data. The 'Article' column contains the references of each of the works. The selected works are divided into two groups. The first group manages a lot of information on specific cases where BI platforms are used to determine improvement in academic management; this group focuses on the management of universities without accessing the academic management of students where they can detect how they learn. The other group focuses on the analysis of the data generated in certain courses, but if the analysis changes the environment, the method does not adapt to it. This is the starting point for the implementation of the BI framework, since the designed method is adapted to all possible variables that are considered in an academic environment. One important contribution of our BI framework is its versatility, and how it can be used to solve different problems that arise in their environment. The BI framework presented in this paper is founded on the gaps and needs identified during our literature review.

Table 2. Classification of articles according to the type of study, by incidence.

Type of Study	Degree of Incidence	Article
Application	High	[14–17]
Model	High	[10,18–21]
Study	High	[3,22]
Application	Medium	[6,23,24]
Model	Medium	[25]
Study	Medium	[26]
Application	Low	[27,28]
Model	Low	[29]
Study	Low	[30]

3. Method

For the development of the method of a BI framework for learning analytics, the seven guidelines the design science research in information systems (IS) were considered [31]. These guidelines have been the basis for research in IS over the last few years. The research used a method that is an alternative, or complement, to the natural science approach that is dominant in IS research. In design science research, the researcher creates and evaluates ICT artifacts designed to solve the identified organizational problems [32].

Taking the design science approach as a guide, the BI method applied to education was developed. The method of the integrated BI framework for learning analytics was composed of six stages: the selection of data sources, the selected data, the preprocessed data, the transformed data, the data mining and the discovery of knowledge. Figure 2 details the stages considered in the implementation of the BI framework [14]. In the following, the different stages of the BI framework are explained in detail.

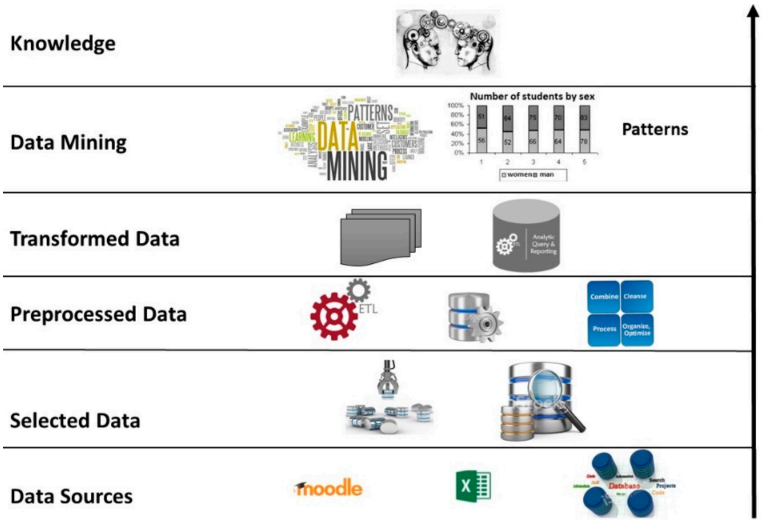

Figure 2. Stages in the implementation of the BI framework.

Once the sources that are important in the analysis have been identified, the selection of the data is made. The objective is to avoid ambiguity, and for the BI process to guarantee quality results. This process includes a data dictionary, which is effectively a technical description for each of the repositories belonging to the university [33]. In the data dictionary, the fields of the data, their origin, their availability and the person in charge are recorded. With the information gathered about the sources and data dictionary, the variables according to which the BI system will work are identified. In an educational environment, unlike a business one, other types of variables must be integrated that influence the analysis process. For example, the variables considered are academic, psychosocial and financial; this additional information gives flexibility to the system when it is necessary to change the object of study.

The selected data must go through a preprocessing stage; this phase involves a process of extraction, transformation and loading (ETL). The ETL extracts the data that are part of the analysis directly from the different data sources [34]. Depending on the tool, complete tables can be extracted from a database or, if greater granularity is desired, only one column can be extracted, depending on the data entry model. The available data pass to a transformation stage, where techniques are applied that allow the treatment of the null or voiding values, eliminating records or imputing missing data. From this process, adequate raw material is obtained, which means that the data will be complete, clean and consistent. The transformation also allows the form of the data to be modified by creating newly derived attributes, changing the type of attribute or making a partial or total change of a table. Another technique used is the discretization in which a numerical value is converted to an ordered nominal value. For example, qualification records can change from 0 to 100 for 'approved–failed'. This is performed in order to facilitate understanding when presenting the results to data analysts. There are data that need a more elaborate process for their transformation; in these circumstances, nested queries and stored procedures are applied in the databases; this process depends on the complexity of the information that will be added to the system. Finally, in this phase, the same ETL is used to load the data that has been transformed and validated to a repository, whose characteristic is that it consolidates all the data from different sources that are accurate and of high quality. The load, as well as the extraction, depends on the tool that is used for the ETL [35].

The data that have gone through preprocessing need to be stored in a repository that allows easy exploitation and ensures high availability. To comply with these requirements, a data warehouse (DW) was designed considering the star model [36]. The design of the model consists of a fact table that contains the data necessary for the analysis and is fed by several dimension tables [13]. The BI architecture developed allows data to be consolidated, segmented or reorganized as necessary, in order to process a wide variety of questions relevant to the user's area of interest [37]. This feature was obtained by creating cubes of on-line analytical processing (OLAP) that generate management modules, which make the cubes extensible and customized [38]. In the last phase of the system, management tools are developed, allowing reports to be considered that can easily be viewed in a browser or as Excel files, depending on the needs. As a method for knowledge discovery, data mining algorithms such as decision trees, cluster, etc. are applied. These algorithms allow correlations between data, patterns and important facts to be identified and knowledge maps to be created [39].

3.1. Visual Modeling of Business Intelligence

In our approach, the design of the BI architecture was defined through various Unified Modeling Language (UML) diagrams. The objective was to clarify each stage of the framework design so that it is replicable in any educational institution. Figure 3 details the activity diagram, where the first step is the collection of information about students from the different repositories, and the loading of the data. The clean, quality data are stored in the DW; the DW is an optimized multidimensional database that facilitates the creation of OLAP cubes. The cubes consist of making a disposition of the data in vectors to perform a quick analysis of the data. If the data exist, they pass to the processing stage through the ETL, which is responsible for extracting the information.

The next step is the evaluation of the variables, where it is verified that the cubes contain enough information about the evaluated variable. For example, there could be dimensions that are not complete or that require additional data; to these dimensions it is necessary to add all the information that affects the event. Similarly, when there are no measures in the analysis dimension, it is necessary to go back to the beginning of the process to add these variables. If the process continues and is not corrected, the result cannot be quantified; therefore, the decision-making is compromised.

If the variables considered in the design of the cubes contain the information required for the analysis, the mining algorithms are applied to the data; this application allows the classification of the students according to the patterns identified by the process. The application of data mining algorithms is usually performed within the same BI application thorugh the use of internal tools within the process. If, for any reason, the process of data mining has to be executed independently, it is necessary to consider tools that integrate the design of the proposed architecture. Finally, the BI application has to present the data in an appropriate way to the user, and in exact times. The decision-making systems present the results in dashboards; their integration is acheived through connectors to the DW that contains the quality data; another option is that the data presentation applications are directly integrated with the tool of data mining.

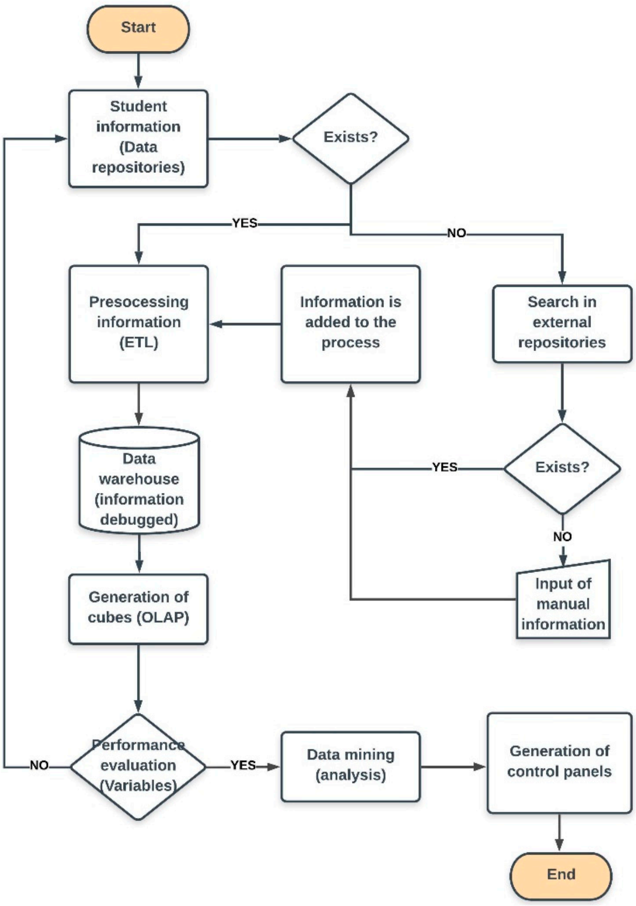
Start
Student information (Data repositories)
Exists?
YES
NO
Presocessing information (ETL)
Information is added to the process
Search in external repositories
Data warehouse (information debugged)
YES
Exists?
NO
Input of manual information
Generation of cubes (OLAP)
NO
Performance evaluation (Variables)
YES
Data mining (analysis)
Generation of control panels
End

Figure 3. Flow diagram of a BI applied to education.

Figure 4 represents the management process the tool uses as a SIPOC diagram (an acronym of supplier > inputs > process > outputs > customers). In the first phase, it determines the data on which the process is based; for this work, the data generated by the students and the data of the academic and financial systems of the students were considered. The entries are the data generated by the students and recorded on various platforms, such as an LMS, academic systems that store grades, student attendance.

The information systems are supported by the entries of the financial information, as well as the socioeconomic information of the students. The BI needs all possible data that provide information on the performance of students, including information that comes from the high school, and from this, it can make projections and know the state of knowledge of each student. The process phase includes the entire knowledge discovery in databases (KDD) process, which begins with the data collection where all the data considered in the entries are integrated [40].

The selected data are processed in the ETL for later stages, submitting them to the data mining algorithms and discovering the knowledge of the object of study. The output phase manages two transitory points; the data processed by the ETL and stored in the DW. In the first point, the output presents clean, quality data that is used in the data analysis.

The second point, and most important, analyzes the data, allowing for the identification of patterns, the classification of individuals or the making of projections. This phase allows us to aquire knowledge about the students, the way they learn and the activities that help them generate learning. The final stage is made up of the customers that are the students, teachers and administrators in our BI framework. The consumption of data is achieved through a dashboard, where the information is easy to consume.

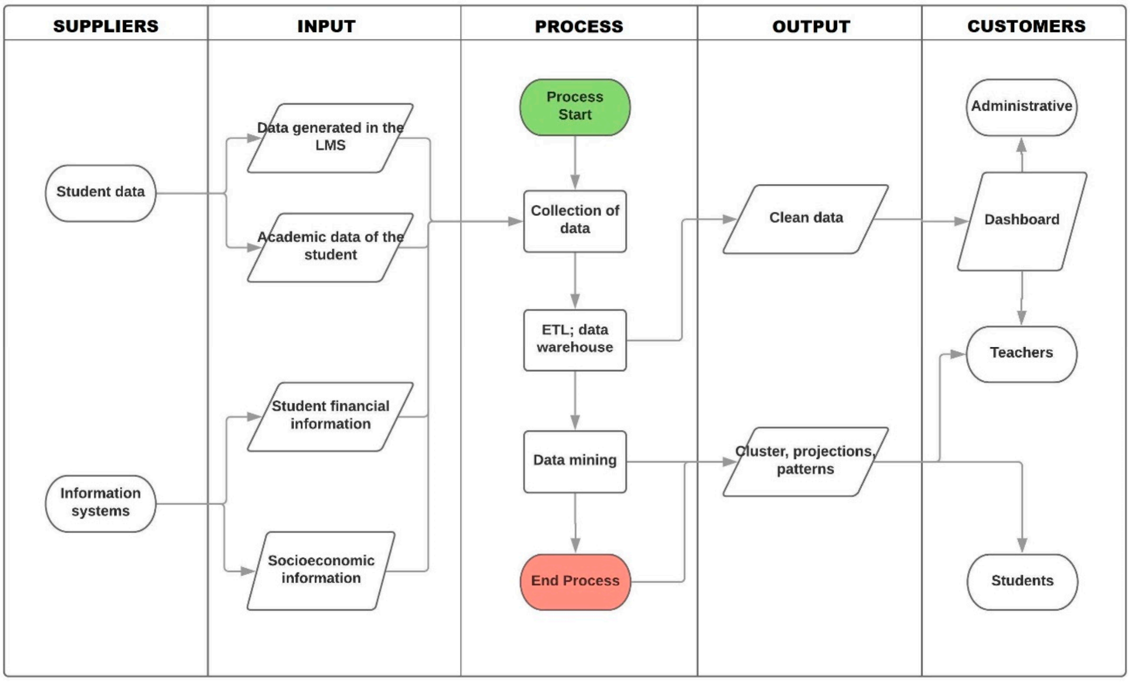

Figure 4. SIPOC diagram for the management of processes in a BI architecture.

3.2. Selection of Data Sources

In the analysis of learning outcomes, data scientists integrate several additional sources, such as structured and unstructured databases. Traditionally, knowledge about aspects of student learning was obtained from information stored in databases. With the inclusion of new study variables in learning, as well as new tools, it has been necessary to obtain more information about student behavior. This behavior does not necessarily have to be registered as part of a traditional database model. For example, one source that can help define student trends is social networks, whose data can be analyzed by techniques such as sentiment analysis.

At present, the greater the volume of data on an object of study, the greater the knowledge that can be obtained using methods and techniques that allow the analysis of learning. However, it must be remembered that any analysis process should be evaluated based on the quality of results, as well as the time at which they are presented, so that the resulting decisions can be made appropriately. To guarantee the speed and quality of the analysis process, it is important to establish a preliminary phase in the KDD process. The selection of the sources is presented as a preliminary phase, in which the data scientist analyzes what each of the sources is, their importance, and whether they already exist or need to be generated.

The available data sources are evaluated by different criteria. The criteria are variables that measure whether the sources are relevant, efficient, impactful or sustainable. What is sought is the identification of the degree of importance that each source has for the analysis process carried out by the BI.

3.3. Selection of Data

For the selection of data, it is important to know the existing data, their relationships and the characteristics of each field, so the development of a data dictionary is essential. A data dictionary is a centralized repository of information about the data, such as their meaning, relationship to other data, origin, use and format. The data dictionary contains the logical characteristics of the students that will be used in the BI framework, including names, descriptions, identifiers, content and organization. With these details, the analysis can make use of data from different repositories and assess their relationships with lower processing costs and heightened effectiveness. In the process of data selection, the records of a database are integrated into the analysis, depending on the problem that needs to be resolved, to the point that all the records can be included. However, it is better to avoid adding all records, so as not to overload the process and thereby reach the desired level of depth in the analysis at appropriate times. To facilitate the data selection process, it is common to work with questions that help identify the different fields that will comprise the variables, for example:

- Is there information available on the systems that can be used?
- Does this information help the analysis?
- Of all the available types of information, which one interests us?
- Are the details of all the available information interesting, or just the details of the information that we need?

Once the data have been selected, a method for ETL is defined. This process depends on the availability of databases; sometimes, it is possible that ETL processes are not able to access certain repositories. In these particular cases, it is necessary to collaborate with the analysts of the database in order to create queries and preprocess the data, and that these are exported and loaded into the DW.

3.4. Preprocessing

The data found in the different repositories may not all be clean, or may contain errors. A lack of cleaning and treatment in the data reduces the quality of the analysis by generating useless rules in the data mining stage. The problems found in the data repositories considered in this study include incomplete data that are missing attribute values, inconsistent data and even discrepancies between the data [41]. The preprocessing of data is achieved with filters that can be applied in a supervised and unsupervised manner. In both cases, there is the option of cleaning the attribute or instance; this depends on the type of data that the analysis needs. The advantage of prior data preparation is that a smaller data set is generated, improving the efficiency of the data analysis process, especially at the mining application stage.

The tasks and classifications assumed in preprocessing are responsible for obtaining data from various sources of information, solving representation and coding problems by integrating data from different tables to create homogeneous information. The data preprocessing consolidates the data in an appropriate way to extract the information that is produced through the data summary or the aggregation operation. The data reduction filters relevant data for the application of the mining models through the selection of features, the selection of instances and discretization.

3.5. Transformation

Data transformation consists of creating new attributes from the original attributes. Transformation also includes global transformations, which involve the exchange of rows with columns, where the data is transformed through a certain type of selection. The selection of data can opt for a vertical selection, which acts directly on the attributes of the object of analysis. Horizontal selection then acts on the data instances and, finally, a sample of the population can be used to perform the transformation. These instances are followed by two possible actions [34]. In the first action, several interviews are applied to detect the origin of the data. In the interview, the administrator of the database is responsible for disclosing the age of the records that are online. The second action allows us to know the structure of the tables and the fields that appear in the entity-relationship diagram of each database included in the investigation.

3.6. Data Warehouse Design

Due to the large amount of data handled in the university environment, the design and construction of a DW is necessary [42]. The integration of the DW with the system provides agility in processing and high availability of the data. For the design of the DW, the Kimball methodology was adopted, which is based on the dimensional life cycle of a business.

The solution of 'DW and BI' is extremely complex when implemented through the Kimball methodology; this complexity can be simplified by improving the processing of BI [43]. The tasks of this 'life-cycle' methodology are shown in Figure 5, where two characteristics can be observed. First, the central role of the task of defining requirements must be highlighted. The requirements for measuring an academic event form the initial support of the subsequent tasks. The measurement has an influence on the project plan. Second, three routes or paths that focus on three different areas are shown:

- Technology (upper path): this has implications relating to specific software.
- The dimensional model (middle path): its design and implementation integrate the ETL subsystem to store the data in the DW.
- BI applications (bottom path): this route includes tasks in which business applications are designed and developed for end users.

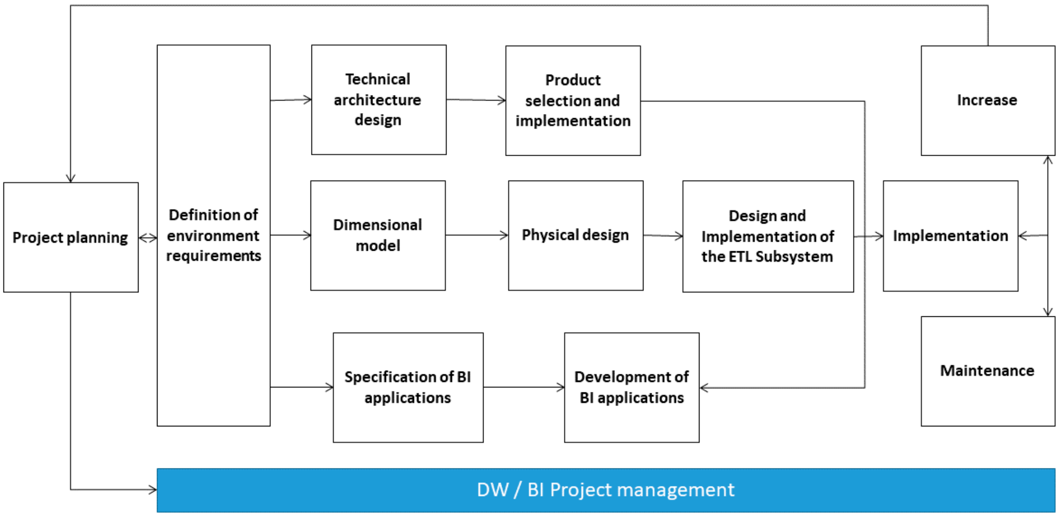
Project planning
Definition of environment requirements
Technical architecture design
Product selection and implementation
Dimensional model
Physical design
Design and Implementation of the ETL Subsystem
Specification of BI applications
Development of BI applications
Increase
Implementation
Maintenance
DW / BI Project management

Figure 5. Tasks of the Kimball methodology. Source: [44].

The analysis of the requirements for the design of DW was achieved through interviews with administrative, academic and technical personnel. The objective was to learn as much as possible about a university environment, its competitors and the student population. Part of the preparation process was to find out who really should be interviewed, which involved carefully examining the organigrams of the organization.

The four groups of interviewees considered crucial for the collection of the data necessary for the analysis of requirements were: the people responsible for making the strategic decisions at an administrative and academic level; personal systems, i.e., the people familiar with the types of the administrators and intermediate academics responsible for exploring strategic alternatives and applying decisions; the people in charge of solving computer problems and managing the data that exists; and, finally, the people that need to be interviewed for political reasons.

Finally, the most critical business requirements or processes were prioritized, for which a dimensional model was created through a dynamic and highly iterative process. The design process began with a high-level dimensional model obtained from the prioritized processes of the matrix [45].

For the iterative process, four steps were considered:

1. The selection of the business process: in this step, the area to be modeled was selected; this is a management decision, and depends mainly on the analysis of requirements and analytical issues.
2. The selection of dimensions: the dimension tables identified were those whose attributes are possible headings in the reports, pivot tables, cubes, or any form of visualization, one-dimensional or multidimensional.

3. The identification of measurements and fact tables: the measurements are attributes of the tables selected to be analyzed; they summarize or group the data using the cut criteria known as dimensions.
4. The linking of measurements to the level of granularity: each of the records of the DW contains a measurement expressed in numbers, such as the number of students or tasks, time and notes, etc.

3.7. Data Mining

The data submitted to be preprocessed were available for the application of data-mining algorithms. In the data-mining application, we looked for patterns that are relevant in a certain segment [46]. The choice of algorithm for a data analysis task can be considered a challenge, since each algorithm generates a different result [47]. However, this does not mean that an analysis cannot use more than one algorithm; the important thing is to determine the most effective entries and then apply a certain algorithm to predict a particular result based on existing data [48]. The choice of data-mining algorithms was made according to Table 3, where several examples of tasks and algorithms that can meet these needs are presented. The data mining algorithms for the construction of the model were selected based on the study [39] that offered an ideal process for this phase.

Table 3. Data mining algorithms applied to different tasks. Source: [49].

Tasks	Algorithms
Predict a discrete attribute. Identify students from a list of subjects as students with good or bad perspectives. Calculate the probability that a student will fail a subject in the next six months. Classify the evolution of students and explore related factors.	Decision tree algorithm Naive Bayes algorithm Cluster algorithm Neural network algorithm
Predict a continuous attribute. Predict the grades of a student next year. Predict the students that will visit the LMS based on provided historical and seasonal trends. Generate a risk score from past data.	Decision tree algorithm Time series algorithm Linear regression algorithm
Predict a sequence. Perform a clickstream analysis of a university website. Capture and analyze sequences of activities during student visits, to formulate recommended practices in common activities.	Algorithm of sequence clusters
Search for groups of common elements in transactions. Suggest additional activities to students. Analyze the data from a student survey to discover which activities were correlated in order to program future activities.	Association algorithm Decision tree algorithm
Search groups of similar items. Create groups of students with risk profiles based on attributes such as demographics and behaviors. Analyze users through search patterns. Identify activities with similar usage characteristics.	Cluster algorithm Algorithm of sequence clusters

3.8. Interpretation and Evaluation

Once the entire architecture of a KDD process has been deployed, the model goes to the validation phase, verifying that the results are valid and sufficiently satisfactory. In the interpretation stage, the data are analyzed according to the patterns and models that were found in them. Patterns are local structures that make statements about a space restricted by variables; for example, a common anomaly is the detection of a student's absences from a course [50]. Models are global structures that make claims about any point in the measurement space; for example, based on the behavior of a group of students in certain subjects, a projection of their effectiveness in future activities can be made. In the presentation of results, the visualization techniques are important because the final models or the descriptions in text format can be difficult to interpret for the final users.

4. Case Study

As an example of the application of our method of BI for education, this case study presents an analysis of the desertion of students at a university. The study began with the analysis of a data set generated from the information concerning the registration process and academic performance record of 3,207 students who entered the distance learning program of a university in Ecuador. The data were evaluated for the period between March 2014 and September 2018; the analysis includes six cohorts because the academic process in this university takes place in semesters, which means that all students who attended up to the sixth semester were considered.

The databases examined were those of the academic and financial management that included a certain amount of data from the LMS [51]. These data were requested from the admissions department and from the academic management of the registry of the higher education institution participating in this study. The inclusion criteria were limited to the information available, and were as follows: age, sex, college grade, subjects taken, subjects passed and status in program [52]. Once the process of quality control and the validation of the information corresponding to the study variables was carried out, 2030 (63.3%) records were included. The variable of analysis in the study was student desertion (dependent variable), defined as a condition of those students enrolled in the first three years of their course, or the initial six semesters. The independent variables that explain the desertion were a set of variables called 'pre-university': the college grade corresponding to the numerical value of the general average of a student's secondary studies, the number of subjects approved; the number of enrollments in the defined periods, the subjects taken (between one and 24, coded according to the average number of subjects taken), sex, and the age of the students between 18 and 26 years old.

The problem addressed by the study concerns the detection of the causes of university desertion; previous works have considered [53,54] desertion to constitute the failure of a student in a consecutive period. For example, in the case of the LMS data, the BI required access to all the records of the activities carried out by the professors and the students that are usually stored in MySQL. All the data obtained from the different sources went through a processing and transformation phase so that clean data could be obtained, and these were loaded into a data repository generated in SQL. From here, the data-mining algorithms were applied according to the analysis carried out in the method, where the operation and the expected results are indicated.

4.1. Data Mining

The selection of the data was based on the study variables; these variables were chosen according to the Bean model [55]. This model identifies the academic, psychosocial and environmental variables that lead to desertion syndrome. As Figure 6 details, each of the variables is constituted by a field that serves as a guide for the extraction of data from the different repositories belonging to the educational institution. For the academic variables, the performance and the integration of the student were considered; the common repositories that store this information are the academic records systems. Psychosocial variables included objectives, alignment, peer interaction and interaction with teachers. The variables that manage the interaction with peers in the Bean model were supplanted by interaction with the LMS. Environmental variables were parameterized according to funding, external social relations, transfer opportunities and interaction with teachers who, as in the previous point, change the parameter through interaction with the LMS and the use of ICT.

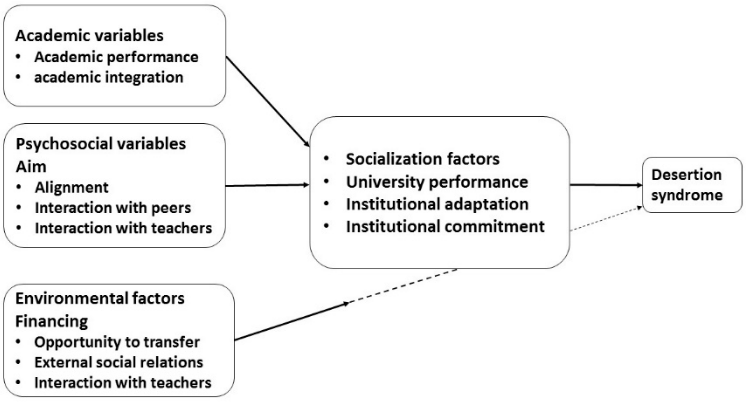

Figure 6. Educational desertion syndrome. Source: [12].

The percentage value of the general average of secondary studies is 70 %, which is registered in two categories: the number of enrollments in the university, and the number of repeated subjects, defined as the numerical value corresponding to the number of times a student takes a subject. Another variable is the recorded age at university entry, from 18 years [56]. The population considered were studying via the modality of distance learning; specifically, within the administration and communication program. Figure 7 shows the evolution of the number of students from 2014 to 2018; the graph shows data for two semesters per year, except for 2014. The data for the period September 2014–January 2015 was not considered because the cohort of that period corresponds to the year 2014. Another consideration is that the information corresponds only to active students. However, it serves as a reference for the verification phase of the results.

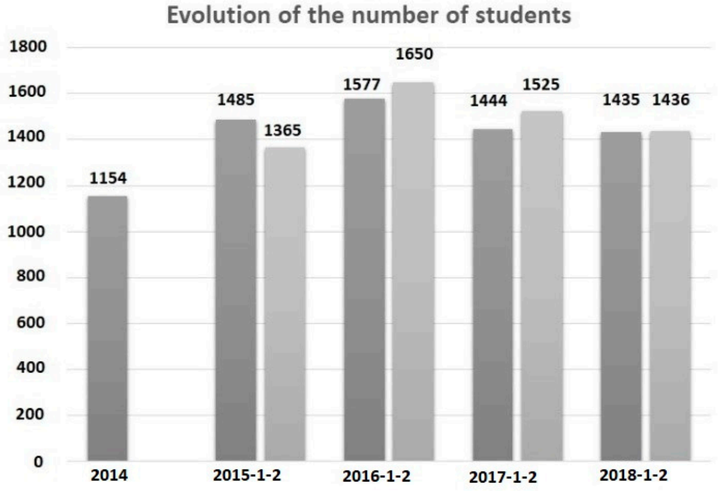

Figure 7. Student registration by period from the years 2014 to 2018.

4.2. *Identification of Necessities*

The university emphasizes the evaluation of learning, the retention of students and the continuous improvement of educational quality. These factors are the main problems to which institutions seek a solution based on the analysis of the educational data held by the systems responsible for learning management.

The student retention rate is given by the following equation:

$$SRR = \frac{NSEA}{TNSA} \quad (1)$$

SRR: student retention rate.
NSEA: number of students enrolled during the regular academic period in which the evaluation of the institution takes place, who were admitted two years before.
TNSA: total number of students admitted to academic programs in institutes of higher education two years before.

The retention rate is a result of good practices in the training process. According to the national regulations of Ecuador and its bodies, it is expected that 80% of students will continue their studies in higher education institutes [57].

4.3. Data Preprocessing

Data cleaning was performed predominantly through the ETL, which has certain advantages, such as the ability to perform an action on an entire table or column according to the need. In this phase, it was perceived that certain columns with important information on the students were scattered in the data sources and contained values that had to be modified. These values were fields with 'NULL' that needed to be replaced with '0' values. This process was achieved with the use of data conversion actions. The data belonged to a single table in the DW, so a combination of the columns in the assigned table within the repository were created and loaded.

The tables, fields and identifiers were designed to cover the largest number of variables, since one of the objectives of the work was to present solutions to the greatest number of critical areas of the university. One common source of data duplication is the excessive number of databases that such an institution manages; the result is that the same student sometimes presents different information [41]. As part of the data preprocessing, both the repositories and the selection of data were reviewed exhaustively.

The ETL chosen for this work was Microsoft's SQL integration service, which possesses several useful tools for processing that facilitate the extraction of data from a source, the transformation and the load. On the other hand, there were data fields that needed to be filtered using another method, such as in the case of missing data. Solutions for the problem of missing data include manually filling in the missing values and filling them in automatically with the word 'unknown'. Data duplication is a major problem in data mining because it often causes time to be lost and it can make the verification of a result difficult. The common problems found in cases of data duplication include several periods being assigned to the same student. In this situation, the ETL detects these cases and displays them with a notification [58]. Figure 8 presents a capture of the ETL that is responsible for the treatment of the data that are deposited in the DW, where blank, null and repeated cells, etc. are eliminated. The clean data is then combined, and the destination that is the DW is chosen.

4.4. Algorithm for Data Analysis

The software available for the application of data mining is varied. The choice depends on the type of results that are expected to be obtained, the profile of the analysts interpreting the results and the economic capabilities of the institution, since there are open source tools available. In this study, greater importance has been given to the results obtained when applying the different algorithms using the WEKA software [59].

This work seeks to establish the independent variables determining student desertion, for which the algorithm J48 (which belongs to the decision tree group) was applied. The J48 algorithm is an implementation of the algorithm C4.5, one of the most commonly used data-mining algorithms. It is a refinement of the model generated with OneR; it presents a moderate improvement in benefits, and may achieve a probability of success slightly higher than the previous classifier [1].

For the connection between the repository and WEKA, an SQL query is generated considering all the fields that need to be analyzed, and it is pasted in the 'Query' section in the interface. Once the WEKA query is executed, it performs an attribute comparison. If necessary, the fields and attributes can be modified to navigate between the information added to the tool. For example, attribute comparisons can be made between students and the selection of activities. WEKA starts the analysis and identifies the patterns and presents them in a matrix that compares all the selected attributes.

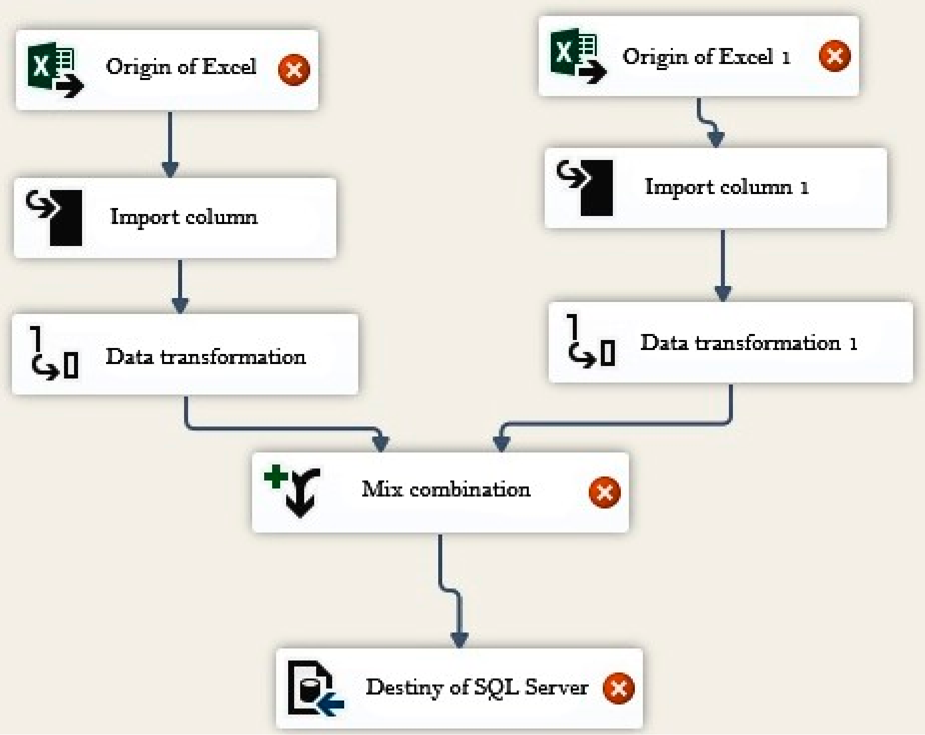

Figure 8. Scheme of extraction, transformation and data loading.

For the analysis, a tree of size 44 with 25 levels was considered; the reliability value was 92.7885%. Table 4 presents the information concerning the data from the generated tree, where the error values and the instances for the analysis are established.

Table 4. Stratified cross-validation.

Correctly Classified Instances.	193	92.7885%
Incorrectly Classified Instances	15	7.2115%
Kappa statistic	0.7628	
Mean absolute error	0.0683	
Root mean squared error	0.2079	
Relative absolute error	30.4551%	
Root relative squared error	625.723%	
Total Number of Instances	208	

An application of the data-mining algorithm does not always include all the variables of the database; the process is improved by selecting the tables that provide information to the hypothesis of the analysis. In order to identify the variables that directly affect the graduation rate (dominant factors), an analysis process was established using, as in the case of desertion, the J48 algorithm. For validation, the data provided by WEKA was verified manually; its efficiency was proven to be 92.7%.

The evaluation of the results acquired from the data-mining algorithms was based on the evaluation of experimental results. To create and evaluate the classification model, we measured its efficiency; that is, the possibility that the classifier correctly classifies numerous samples from the test data set. For the evaluation of the classification precision, the cross-validation method was used. The study and evaluation process were repeated k times, always using a subset as the test set. The estimation of the prediction precision by cross-validation is a random number that depends on the distribution of samples to subsets.

Various metrics were used in this investigation to evaluate the performance of the algorithms. The classifier performance implies precision, error rate, recall and F-Measure. Table 5 presents the detailed precision data by class.

Table 5. Detailed Accuracy by Class.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.98	0	1	0.98	0.99	0.985	0.99	0.987	Graduate in Career
	0.94	0.03	0.94	0.94	0.94	0.91	0.952	0.88	Graduated
	0.96	0.03	0.941	0.96	0.95	0.925	0.961	0.905	Retired
Weighted Avg.	0.96	0.02	0.96	0.96	0.96	0.94	0.968	0.924	

TP Rate: true positive rate, are the instances correctly classified as a given class. FP Rate: false positive rate, are the instances falsely classified as a given class. Precision: proportion of instances that are truly of a class divided by the total number of instances classified as that class. Recall: proportion of instances classified as a given class divided by the actual total in that class, equivalent to the TP rate. F-Measure: a combined measurement of precision and recall calculated as 2*Precision*Recall/(Precision+Recall). MCC: used in machine learning as a measure of the quality of binary classifications (two classes). It takes into account true and false positives and negatives, and is generally considered to be a balanced measure that can be used even if the classes are of very different sizes. ROC: measurement of the receiver operating characteristic area; it is one of the most important WEKA output values. They provide an idea of how classifiers are working in general. PRC: Precision Recovery Area; the precision recovery graph is more informative than the ROC graph when evaluating binary classifiers on unbalanced data sets.

The measurement of the ROC area is one of the most important values of the WEKA output. The optimal classifier has ROC area values close to 1, with 0.5 being comparable to random guessing. Unbalanced data sets in which a disproportionately large number of instances belong to a certain class can lead to high accuracy rates, although the classifier is not necessarily excellent.

4.5. Analysis of Results

For the identification of the independent variables, several tests were performed; the objective was to determine the percentage of incidence of each attempt with the dependent variable. With this information, the variables that contribute least to the study could be discarded. Another variable considered for the validation of the model was the confusion matrix.

The matrix visualizes the distribution of the records in terms of real classes and predicted classes. This indicates the quality of the current model. Table 6 presents the data of the confusion matrix where, according to the horizontal reading of the table, there are 1,139 items classified in the 'NO' class:

- Of these, 1056 were correctly classified in the 'NO' class;
- 83 of the elements were incorrectly classified in the 'YES' class.
- There are 872 elements in the 'YES' class:
- Of the sample, 29 items are classified as 'NO'.
- 843 of these elements are correctly classified in the 'YES' class.

Table 6. Matrix of confusion.

A	B	C	← Classified as
1056	83	0	| a = NO
29	843	0	| b = YES
0	0	13	| c = Graduate in Career

The analysis considers, as independent variables, whether the student remains in the institution or not. Table 7 presents a part of the tree and the levels whereby the interpretations of the results are centered on the students' status in each school. This status can describe active, passive, graduate and retired students. The levels are chosen for a student that is passive and withdrawn, which directly affects their risk of desertion.

Table 7. Classifier model (full training set).

```
J48 Pruned Tree
------------------
CurrentStatusInSchool = passive
| Enrollment <= 4: NO (848.0/2.0)
| Enrollment > 4
| | Enrollment <= 8: NO (45.0/8.0)
| | Enrollment > 8: YES (7.0)
CurrentStatusInSchool = Graduate: Graduate in Career (13.0)
CurrentStatusInSchool = Graduated: NO (5.0)
CurrentStatusInSchool = Retired
| Enrollment <= 5
| | Work = YES: NO (24.0/3.0)
| | Work = YES: NO (0.0)
| | Work = NO: YES (3.0/1.0)
| Enrollment > 5: YES (15.0/2.0)
CurrentStatusInSchool = Change of School: NO (2.0)
Number of leaves:        25
Size of the tree:        44
```

The passive state is considered the first level of the tree; the second level takes into account the student's enrollment. When the number of enrollments to the course is less than or equal to four, there are 848 cases in which the student left; in the following case, if the interactions are less than or equal to five, the result indicates that 24 students withdrew. In the third case, students who are enrolled in the fifth semester only have 15 students who dropped out.

The objective was to identify those students who do not work and to assess their academic performance. As a result, it was found that 37% of students work and pass all courses without any problems, and the other 63% experience some complication. The analysis conducted shows that 63% of students need at least 19 hours of LMS use to become familiar with their management.

5. Conclusions

This paper presented a method for the implementation of a BI framework that covers the treatment, analysis and presentation of data that can respond to any problem generated in an educational institution. Our framework offers a practical means for universities to begin their journey towards the implementation of BI by crystallizing their most important processes and defining the necessary information to support these processes.

In developing this work, several clarifications were obtained about the use of data mining in the databases that handle students' socioeconomic and academic information. It should be emphasized that, using ICT, the current momentum in education has allowed the creation of modalities that allow students to manage the convergence between work and educational activities. For this reason, the use of ICT should be supported as a mandatory component in virtual education or distance learning.

When applying the EDM models, as shown in the case study examined in this document, we can affirm that technical improvements could be made in terms of their implementation. One improvement that is being worked on is the optimization of infrastructure resources. This allows us to improve the response times in each process; thus, the results of the analysis would appear in real time.

With the analysis of the data, an overview of the behavior of a group of students is formed; the levels of depth depend on the independent variables and their percentage of influence. Some of the results obtained may be obvious; for example, if a student fulfills all the tasks and devotes an appropriate amount of time to self-directed work, he will pass the course without any problem. In the opposite case, it is important to know the causes that lead a student to fail a course, with the aim of recommending a group of activities that can be adjusted to the needs of each student.

As future work, the integration of a system of recommendations of activities in the BI is proposed, in order to create an autonomous system that is capable of making decisions. The process would cover the phase of the processing and analysis of data through the extraction of data. The results obtained would then be sent to an expert system. The expert system would evaluate the factors that influence desertion, academic effectiveness, or any other type of eventuality that may be the object of study. With this knowledge, the system could recommend a group of activities that meet the needs of students in such a way that learning is active and personalized, creating a platform like the one indicated within a university, with the capacity to provide academic, financial, administrative management and marketing solutions. In addition to the possibility that this management is carried out by an expert system, it would undoubtedly help in decision-making and enable the automation of learning processes at a high level of quality.

Author Contributions: W.V.-C. contributed to the following: the conception and design of the study, acquisition of data, analysis, and interpretation of data, drafting the article and approval of the submitted version. The authors X.P.-P. Moreover, W.V.-C. Contributed to the study by design, conception, interpretation of data, and critical revision. S.L.-M. made the following contributions to the study: analysis and interpretation of data, approval of the submitted version. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferguson, R. Learning analytics: Drivers, developments and challenges. *Int. J. Technol. Enhanc. Learn.* **2013**, *4*, 304. [CrossRef]
2. Comendador, B.E.V.; Rabago, L.W.; Tanguilig, B.T. An educational model based on Knowledge Discovery in Databases (KDD) to predict learner's behavior using classification techniques. In Proceedings of the IEEE International Conference on Signal Processing, Communications and Computing, Conference Proceedings, Hong Kong, China, 5–8 August 2016; pp. 1–6.
3. Apraxine, D.; Stylianou, E. Business Intelligence in a Higher Educational. In Proceedings of the 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, Greece, 25–28 April 2017; pp. 1735–1746.
4. Jeong, H.; Choi, C.; Song, Y. Expert Systems with Applications Personalized Learning Course Planner with E-learning DSS using user profile. *Expert Syst. Appl.* **2012**, *39*, 2567–2577. [CrossRef]
5. Villegas-Ch, W.; Palacios-Pacheco, X.; Buenaño-Fernandez, D.; Luján-Mora, S. Comprehensive learning system based on the analysis of data and the recommendation of activities in a distance education environment. *Int. J. Eng. Educ.* **2019**, *35*, 1316–1325.
6. Nobre, T.; Trigo, A.; Sanches, P. Business intelligence system for analysis of access to higher education: The case of the Polytechnic Institute of Coimbra. In Proceedings of the 2014 9th Iberian Conference on Information Systems and Technologies (CISTI), Barcelona, Spain, 18–21 June 2014; pp. 1–6.
7. Hsinchun, C.; Roger, H.L.C.; Veda, C.S. Business Intelligence and Analytics: From Big Data To Big Impact. *MIS Q.* **2018**, *36*, 1293–1327.
8. Márquez-Vera, C.; Romero Morales, C.; Ventura Soto, S. Predicting school failure and dropout by using data mining techniques. *Rev. Iberoam. Tecnol. del Aprendiz.* **2013**, *8*, 7–14. [CrossRef]
9. Gairín, J.; Triado, X.M.; Feixas, M.; Figuera, P.; Aparicio-Chueca, P.; Torrado, M. Student dropout rates in Catalan universities: Profile and motives for disengagement. *Quality High. Educ.* **2014**, *20*, 165–182. [CrossRef]
10. Riofrio, G.; Encalada, E.; Guaman, D.; Aguilar, J. Business intelligence applied to learning analytics in student-centered learning processes. In *Proceedings of the 2015 Latin American Computing Conference (CLEI), Arequipa, Peru, 19–23 October 2015*; IEEE: Piscataway, NJ, USA, 2015; pp. 1–10.
11. Turner, M.; Bailey, J.; Linkman, S.; Budgen, D.; Pearl Brereton, O.; Kitchenham, B. Systematic literature reviews in software engineering—A systematic literature review. *Inf. Softw. Technol.* **2008**, *51*, 7–15.
12. Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Inf. Softw. Technol.* **2015**, *64*, 1–18. [CrossRef]

13. Moscoso-Zea, O.; Paredes-Gualtor, J.; Luján-Mora, S. A Holistic View of Data Warehousing in Education. *IEEE Access* **2018**, *6*, 64659–64673. [CrossRef]
14. Scholtz, B.; Calitz, A.; Haupt, R. A business intelligence framework for sustainability information management in higher education. *Int. J. Sustain. High. Educ.* **2018**, *19*, 266–290. [CrossRef]
15. Valdiviezo-díaz, P.; Cordero, J.; Reátegui, R. A Business Intelligence Model for Online Tutoring Process. In *Proceedings of the 2015 IEEE Frontiers in Education Conference (FIE), El Paso, TX, USA, 21–24 October 2015*; IEEE: Piscataway, NJ, USA, 2015; pp. 1–9.
16. Kumaran, S.R.; Othman, M.S.; Yusuf, L.M. Applying Theory of Constraints (TOC) in business intelligence of higher education: A case study of postgraduates by research program. In Proceedings of the 2015 International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, 27–28 October 2015; pp. 147–151.
17. Maia, A.; Portela, F.; Santos, M.F. Web Intelligence in Higher Education: A Study on the usage of Business Intelligence techniques in Education. In Proceedings of the 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), Barcelona, Spain, 6–8 August 2018; pp. 176–181.
18. Jayakody, J.; Perera, I. Enhancing competencies of less-able students to achieve learning outcomes: Learner aware tool support through Business intelligence. In Proceedings of the 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Bangkok, Thailand, 7–9 December 2016; pp. 154–160.
19. Kollwitz, C.; Dinter, B.; Krawatzeck, R. Tools for Academic Business Intelligence and Analytics Teaching: Results of an Evaluation. *Anal. Data Sci.* **2017**, 227–250. [CrossRef]
20. Villegas-Ch, W.; Luján-Mora, S.; Buenaño-Fernandez, D. Towards the Integration of Business Intelligence Tools Applied to Educational Data Mining. In Proceedings of the IEEE World Engineering Education Conference (EDUNINE), Buenos Aires, Argentina, 11–14 March 2018; pp. 1–5.
21. Hamed, M.; Mahmoud, T.; Gómez, J.M.; Kfouri, G. Using Data Mining and Business Intelligence to Develop Decision Support Systems in Arabic Higher Education Institutions. *Mod. Acad. Teach. Res. Bus. Econ.* **2017**, 71–84. [CrossRef]
22. Musa, S.; Ali, N.B.M.; Miskon, S.B.; Giro, M.A. *Success Factors for Business Intelligence Systems Implementation in Higher Education Institutions—A Review*; Springer International Publishing: Johor, Malaysia, 2019; Volume 843, ISBN 978-3-319-99006-4.
23. Othman, M.S.; Kumaran, S.R. *An Implementation of Metaheuristic Algorithms in Business Intelligence Focusing on Higher Education Case Study*; Springer: Cham, Switzerland, 2018.
24. Rodzi, N.A.H.M.; Othman, M.S.; Yusuf, L.M. Significance of data integration and ETL in business intelligence framework for higher education. In Proceedings of the Proceedings International Conference on Science in Information Technology: Big Data Spectrum for Future Information Economy, Yogyakarta, Indonesia, 27–28 October 2015; pp. 181–186.
25. Christozov, D. Business analytics as a tool to transforming information into an Informing System: The case of the on-line course registration system. *Inf. Sci.* **2017**, *20*, 167–178.
26. Peng, M.Y.-P.; Tuan, S.-H.; Liu, F.-C. Establishment of Business Intelligence and Big Data Analysis for Higher Education. In *Proceedings of the International Conference on Business and Information Management, Beijing, China, 23–25 July 2017*; ACM: New York, NY, USA, 2017; pp. 121–125.
27. Valdez, A.; Cortes, G.; Castaneda, S.; Medina, J.; Haces, G. Development and Implementation of the Balanced Scorecard for a Higher Educational Institution using Business Intelligence Tools. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 164–170. [CrossRef]
28. Elhissi, Y.; Haqiq, A. Information system at the Moroccan University: A business intelligence tool for management and communication of scientific research. In Proceedings of the 2016 International Conference on Information Technology for Organizations Development (IT4OD), Fez, Morocco, 30 March–1 April 2016.
29. Zheeng, G.; Zhang, C.; Li, L. Bringing Business Intelligence to Healthcare Informatics Curriculum: A Preliminary Investigation. In Proceedings of the 45th ACM Technical Symposium on Computer Science Education, Atlanta, GA, USA, 5–8 March 2014; pp. 205–210.
30. Wixom, B.; Ariyachandra, T.; Goul, M.; Gray, P.; Kulkarni, U.; Phillips-Wren, G. The current state of Business Intelligence in academia: The Arrival of Big Data. *Commun. Assoc. Inf. Syst.* **2011**, *29*, 299–312. [CrossRef]
31. Hevner, A.; Chatterjee, S.; Series, I.; Antonelli, P.; Mathew, R. Design Science Research in Information Systems. *Des. Res. Inf. Syst. Integr. Ser. Inf. Syst.* **2010**, *22*, 9–22.

32. Arnott, D. Cognitive biases and decision support systems development a design science approach. *Inf. Syst. J.* **2006**, *16*, 55–78. [CrossRef]
33. Rodriguez-Ascaso, A.; Boticario, J.G.; Finat, C.; Petrie, H. Setting accessibility preferences about learning objects within adaptive elearning systems: User experience and organizational aspects. *Expert Syst.* **2017**, *34*, 1–12. [CrossRef]
34. Villegas-Ch, W.; Luján-Mora, S. Analysis of data mining techniques applied to LMS for personalized education. In Proceedings of the IEEE World Engineering Education Conference: Engineering Education-Balancing Generalist and Specialist Formation in Technological Carriers: A Current Challenge, Santos, Brazil, 19–22 March 2017; pp. 85–89.
35. Wang, Y.J. Multi Dimension Knowledge Mining in Heterogeneous Data Resources. *Adv. Mater. Res.* **2012**, *433–440*, 5256–5262. [CrossRef]
36. Sen, A.; Ramamurthy, K.; Sinha, A.P. A Model of Data Warehousing Process Maturity. *IEEE Trans. Softw. Eng.* **2012**, *38*, 336–353. [CrossRef]
37. Zhang, Z.; Wang, X.; Qi, G.; Yao, W. Study on the method of building data persistence layer based on the data dictionary. *Int. Conf. Comput. Commun. Technol. Agric. Eng.* **2010**, *1*, 320–323.
38. Rahmadi Wijaya, B.P. *An Overview and Implementation of Process in Data Warehouse*; IEEE: Piscataway, NJ, USA, 2015; pp. 70–74.
39. Moscoso-Zea, O.; Saa, P.; Luján-Mora, S. Evaluation of Algorithms to Predict Graduation Rate in Higher Education Institutions by Applying Educational Data Mining. *Australas. J. Eng. Educ.* **2019**, *24*, 4–13. [CrossRef]
40. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. *KDD* **1996**, *96*, 82–88.
41. Lujan-Mora, S.; Palomar, M. Reducing inconsistency in integrating data from different sources. In Proceedings of the 2001 International Database Engineering and Applications Symposium, Xi'an, China, 9–11 July 2001; pp. 209–218.
42. Xi, R.; Lin, N.; Chen, Y. Compression and Aggregation for Logistic Regression Analysis in Data Cubes. *IEEE Trans. Knowl. Data Eng.* **2008**, *21*, 479–492.
43. Lai, C.-T.; Chang, H.-H.; Lee, K.-T.; Huang, J.-Y.; Lee, W.-P. Predicting Drug Side Effects Using Data Analytics and the Integration of Multiple Data Sources. *IEEE Access* **2017**, *5*, 20449–20462.
44. Boumiza, S.; Souilem, D.; Bekiarski, A. Workflow approach to design automatic tutor in e-learning environment. In Proceedings of the 2016 International Conference on Control, Decision and Information Technologies (CoDIT), St. Julian's, Malta, 6–8 April 2016; pp. 263–268. [CrossRef]
45. Lujan-Mora, S.; Trujillo, J. Applying the UML and the Unified Process to the Design of Data Warehouses. *J. Comput. Inf. Syst.* **2006**, *46*, 30–58. [CrossRef]
46. Winter, R.; Winter, R.; Strauch, B. A Method for Demand-driven Information Requirements Analysis in Data Warehousing Projects. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6–9 January 2003.
47. Karr, A.F.; Sanil, A.P.; Banks, D.L. Data quality: A statistical perspective. *Stat. Methodol.* **2006**, *3*, 137–173. [CrossRef]
48. Buenaño-Fernandez, D.; Villegas-CH, W.; Luján-Mora, S. The use of tools of data mining to decision making in engineering education—A systematic mapping study. *Comput. Appl. Eng. Educ.* **2019**, 1–15. [CrossRef]
49. Sarawagi, S.; Agrawal, R.; Megiddo, N. Discovery-driven exploration of OLAP data cubes. In *International Conference on Extending Database Technology*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 168–182.
50. Yan, Y.; Huang, C.; Wang, Q.; Hu, B. Data mining of customer choice behavior in internet of things within relationship network. *Int. J. Inf. Manag.* **2018**, *50*, 566–574. [CrossRef]
51. Huang, T.C.K.; Wu, I.L.; Chou, C.C. Investigating use continuance of data mining tools. *Int. J. Inf. Manag.* **2013**, *33*, 791–801. [CrossRef]
52. Pandey, A.K.; Rajpoot, D.S. A comparative study of classification techniques by utilizing WEKA. *Int. Conf. Signal. Process. Commun.* **2016**, 219–224.
53. Timarán Pereira, R. Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos. In Proceedings of the In Memorias de la 8a Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI, Nariño, Colombia, 14 August–22 October 2009; pp. 1–5.

54. Palacios-Pacheco, X.; Villegas-Ch, W.; Luján-Mora, S. Application of Data Mining for the Detection of Variables that Cause University Desertion. In Proceedings of the Communications in Computer and Information Science; Springer: Cham, Switzerland, 2019; Volume 895, pp. 510–520.
55. Larson, D.; Chang, V. A review and future direction of agile, business intelligence, analytics and data science. *Int. J. Inf. Manag.* **2016**, *36*, 700–710. [CrossRef]
56. Combita Niño, H.A.; Cómbita Niño, J.P.; Morales Ortega, R. Business intelligence governance framework in a university: Universidad de la costa case study. *Int. J. Inf. Manag.* **2018**, *50*, 405–412. [CrossRef]
57. Hirji, K.K. Discovering Data Mining: From Concept to Implementation. *ACM SIGKDD Explor. Newsl.* **1999**, *1*, 44–45. [CrossRef]
58. Syeda Farha Shazmeen, S.F.S. Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis. *J. Comput. Eng.* **2013**, *10*, 1–6. [CrossRef]
59. Lopez Guarin, C.E.; Guzman, E.L.; Gonzalez, F.A. A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *Rev. Iberoam. Tecnol. del Aprendiz.* **2015**, *10*, 119–125. [CrossRef]