Evaluation of Algorithms to Predict Graduation Rate in Higher Education Institutions by Applying Educational Data Mining

Oswaldo Moscoso-Zea

Universidad Tecnológica Equinoccial, Faculty of Engineering Sciences, Quito, Ecuador

Pablo Saa

Universidad Tecnológica Equinoccial, Faculty of Engineering Sciences, Quito, Ecuador

Sergio Luján-Mora

University of Alicante, Department of Software and Computing Systems, Alicante, Spain

Abstract: Nowadays, researchers analyse student data to predict the graduation rate by looking at the characteristics of students enrolled and to take corrective actions at an early stage or improve the admission process. Educational data mining (EDM) is an emerging field that can support the implementation of changes in the management of higher education institutions. EDM analyses educational data using the development and the application of data mining (DM) methods and algorithms to information stored in academic data repositories. The purpose of this paper is to review which methods and algorithms of DM can be used in the analysis of educational data to improve decision making. Furthermore, it evaluates these algorithms using a dataset composed of student data in the computer science school of a private university. The core of the analysis is to discover trends and patterns of study in the graduation rate indicator. Finally, it compares these methods and algorithms and suggests which has the best precision in certain scenarios. Our analyses suggest that random trees had better precision but had limitations due to the difficulty of interpretation while the J48 algorithm had better possibilities of interpretation of results in the visualization of the classification of data and only had slightly inferior performance.

Keywords: data mining, data warehouse, educational data mining, academic development.

Introduction

In today’s information era, data are collected and stored in large repositories. The huge amounts of information that educational institutions generate every day calls for improved ways of storing and analysing data. The process of converting data into information, and information into knowledge, has to be done by following a comprehensive method to produce the expected outputs.

Higher education institutions are generating large amounts of data from their organizational systems and applications which could be more effectively used to discover trends and predict events in education. In the same manner, as in other industries, the right data management and data visualization can grant stakeholders with insights to improve organizational processes. Knowledge obtained from data analytics and data mining (DM) are enablers to ensure quality in the educational process, and therefore, it offers directors different viewpoints to improve the education generally. However, DM is not a solution itself at this point, instead, it is a tool which supports the decision-making process through the acquisition of knowledge in order to solve different problems (Buldu, 2010). The production and dissemination of organizational knowledge is a strategic objective that supports higher education institutions in the roadmap for planning, modernization, and improvement of academic and research indicators (Baepler & Murdoch, 2010) (O. Moscoso-Zea & Lujan-Mora, 2017). Thus, it is extremely necessary to establish mechanisms to store data of the highest possible quality and apply EDM methods.

Student success is an essential objective of higher education institutions, so the technological infrastructure of these institutions must include data warehouses to support sound data storage and analysis, as one of the core technologies in this field. A data warehouse is a data repository modelled with a multidimensional design and used specifically for analysis (Oswaldo Moscoso-Zea, Sampedro, & Luján- Mora, 2016). The information dispersed in different operational databases is migrated to the data warehouse using extraction, transformation, and loading (ETL) processes. An approach that guides this knowledge creation process is called knowledge discovery in databases (KDD). KDD uses DM as the core element for knowledge creation. Because of this, DM has been applied to different industries and fields of study in the last years with promising results. Some of the fields of analysis of DM are marketing, health, finances, and insurance, among others. The application of DM in educational contexts is known as educational data mining (EDM). EDM is a discipline in evolution that focuses on the design of models to improve learning experiences and organizational efficiency (Huebner, 2013). Improvements will come from the interpretation of the analysis of the variables in the dataset and instituting evidence-based improvements to teaching and learning practices. EDM is complementary to other approaches for understanding the learning process and uses software tools to discover trends and patterns in educational data to improve decision making in higher education institutions.

Currently, there are different initiatives for improving education using data analytics. EDM is one of these initiatives to improve students, lecturers, and staff performance. To mention some examples and case studies, in the following list, we present initiatives from institutions that are working in this field of study:

- Austin Peay State University takes the algorithmic approach to higher education one step further. Before students register for classes, a robot adviser assesses their profiles and nudges them to pick courses in which they are likely to succeed (Parry, 2012).
- In Arizona State University thousands of students take math courses through a system that mines performance and behavioral data, building a profile on each user and delivering recommendations about what learning activity they should do next. This addresses the continuous problem of students being unprepared for college math (Wishon & Rome, 2016).

- Purdue University has been using DM to determine that frequent evaluations in early stages can change the habits of students with low grades; by an academic alert system that tracks the performance of students (Baepler & Murdoch, 2010).
- Paul Smith's College uses learning analytics to increment the graduation rate of its students (Bichsel, 2012).
- Georgia University carries out experiments using analytic techniques to predict graduation rates in online courses (Morris, Wu, & Finnegan, 2015).

The work at the universities mentioned above exemplify the value of EDM and have encouraged us to continue with this line of research. Consequently, this paper presents the main methods and algorithms of EDM that have been developed by scientists. Furthermore, it describes the experiments carried out to analyse one key performance indicator in a private university: the students' graduation rate.

Classification methods and algorithms of EDM are applied in this analysis. In addition, this study compares selected methods and algorithms and suggests which has better precision in the graduation scenario using a predefined dataset. The conclusions of this paper inform educators on how to choose the right methods and algorithms for further studies of this key performance indicator.

Thereupon, the research question of this work is:

"Which are the best methods and algorithms of EDM that can be used to analyse the graduation rate in university students?"

The results and findings from this question can be valuable for researchers in education to reduce the time of experimentation and the analysis of students' graduation rate. Thus, providing a clear vision of the methods and algorithms that had better accuracy in the analysis of a dataset for graduation.

Background

This section describes the state of the art of the EDM topic along with the existing methods and algorithms of EDM that are needed for the design of the experiments. Moreover, this section provides the reader with a summary of some key concepts of this field, and guides researchers to select the right algorithms for their experiments with EDM. Additionally, this section helps the non-expert reader to understand the methodology and the results presented in the rest of the paper, where different EDM algorithms are evaluated and compared.

Data Mining

The term DM is also known as "Data Archeology", "Data Collection", "Knowledge Extraction", or "Data Analysis". DM is an approach that uses different information technologies, systems, and tools to analyse and extract knowledge from information contained in data repositories of organizations. DM is a fundamental part of the knowledge creation process. The most commonly used framework in a DM project is the Cross Industry Standard Process for Data Mining (CRISP-DM), as is depicted in Figure 1 (Chapman et al., 2000).

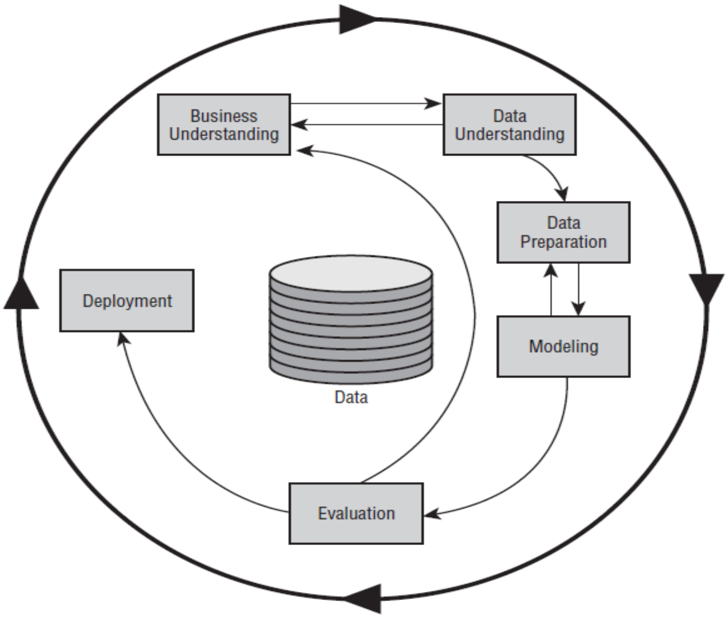

Figure 1: Phases of CRISP-DM. Source: (Chapman et al., 2000)

The implementation of DM follows the six CRISP-DM phases. These phases are:

(1) Understanding the business: Comprehension of the mission, vision, and goals of the business and how the DM project will benefit the organization. In this phase, it is important to have clearly identified the requirements of the project.

(2) Understanding data: Identifying the data tables and fields that will be subject to analysis.

(3) Preparing data: Migrating the required data to a dataset that will be used in the analysis. A data cleansing process must be performed during this transformation.

(4) Creating models: Designing and planning the model that will be used for analysis.

(5) Evaluation: Experimenting and selection of algorithms and tools. The output of this phase is the knowledge created with the existing model.

(6) Deployment: Presentation of the results. If the results are not relevant to the requirements, a new model should be proposed.

Knowledge Discovery in Databases

The KDD method is an approach to discover useful knowledge from a group of data. This process is composed of five main steps as is shown in Figure 2 (Fayyad, Piatetsky-Shapiro, & Smyth, 1996a), and described in the following numerals:

(1) Data selection: Selection of data sources from operational databases, and the migration to a target data repository, in this case, the data warehouse.

(2) Data pre-processing: Cleansing and pre-processing of data by deciding strategies to put the data in the right format, removing duplicates, and handling missing fields.

(3) Data transformation: Creating datasets with the needed variables for reducing the complexity of analysis.

(4) Data mining: Application of methods and algorithms to the dataset in order to predict trends and discover patterns in data.

(5) Interpretation and evaluation: Understanding the results and the creation of explicit knowledge by means of visualization of data in reports and dashboards.

In Figure 2, the basic steps comprising the KDD process are illustrated, but not the potential iterations and loops that can be established between any two steps.

DM and KDD are closely related, both in terms of methodology and terminology. However, DM is the analysis step of the KDD process. More specifically, DM "is the application of specific algorithms for extracting patterns from data" and KDD "refers to the overall process of discovering useful knowledge from data" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b). The additional steps in the KDD process are critical because they guarantee that useful knowledge is derived from the data.

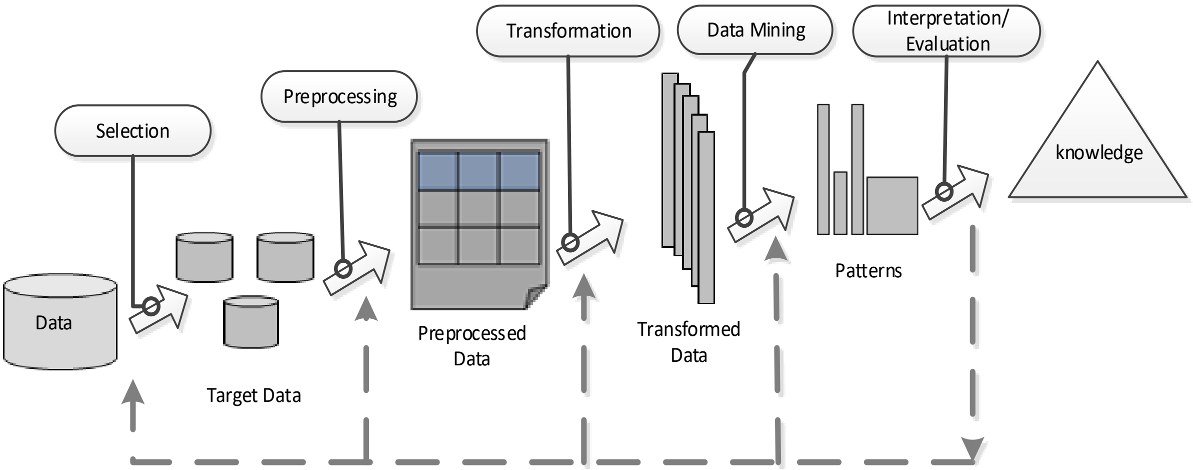

Figure 2: An overview of the steps of a Knowledge Discovery in Databases process. Adapted from: (Fayyad et al., 1996a)

Educational Data Mining

EDM is the name of the data mining process applied to education. It uses methods, technological tools, and algorithms to investigate data from educational databases. These academic repositories process data about students, academic and administrative staff, and academic processes as admission, registration, student welfare, research, among others. A popular definition for the EDM field is proposed by the International Society of Educational Data Mining: "*EDM is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings which they learn in*" (International Educational Data Mining Society, 2018). The main idea of EDM is to create knowledge from the data gathered from students and educators to improve the educational processes.

The knowledge discovery process that we suggest to follow is shown in Figure 3. This figure depicts in the first step the pre-processing of raw data obtained from an educational environment.

These raw data are then modified (a new dataset is created) and used with EDM methods or algorithms. In the next step, the model is defined, and the experiments are carried out. The results of experimentation allow the evaluation and refinement of the process with the results of the analysis (Romero & Ventura, 2013).

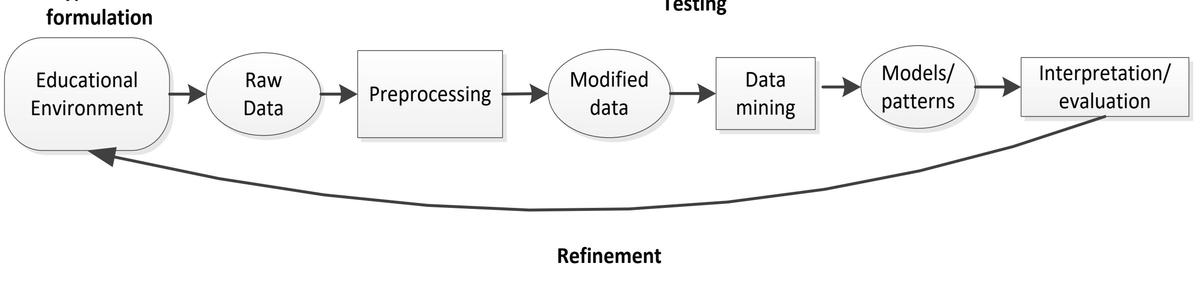

Figure 3: Knowledge Discovery with EDM. Source: (Ventura & Romero, 2013)

EDM explores the data in the organizational context of a higher education institution to transform these data into useful knowledge. Consequently, it is key in the study and improvement of academic indicators as dropout rate, graduation rate, restructuration processes and organizational management (Bienkowski, Feng, & Means, 2012).

Methods and Algorithms of EDM

The field of EDM comprises methods, algorithms, and techniques to perform different experiments and to design models. The output of the model implementation allows researchers to predict or obtain patterns to improve performance indicators.

This section presents the algorithms and classifies the methods used by researchers in this scientific field. For example, the following list describes a very popular classification presented by (Baker & Siemens, 2014):

- Prediction: Is used in the design of a model that allows inferring in some aspects of the data. It is based on the combinations of other features of the data. For example, information on student dropouts can be analysed to take corrective and preventive actions targeted to probable dropout candidates. There are three types of methods in this group: classification, regression, and latent knowledge estimation.
    - Classification: is a supervised learning approach in which the system learns from the information loaded and then uses this knowledge to infer and classify new observations.
    - Regression: is a statistical measure used to determine the relationship between one dependent variable and a series of other mutable variables (independent variables).
    - Latent Knowledge Estimation: is the estimation of the building blocks of existing knowledge that has not been harnessed to produce new knowledge

- Structure Discovery: Is a question of finding a structure for the data without a prior idea of the solution. The researcher task is to identify the natural structure of the data. Within this classification, we can mention clustering, factor analysis, social network analysis and discovery of domain structures.
- Relationship Mining: Is useful to discover relationships between certain variables in a dataset. Within this category, we can mention association, correlation, sequential pattern mining and causal mining of data.
- Model Discovery: The results of the mining analysis are used as inputs for further analysis. Normally a model is obtained through prediction methods.

Moreover, different EDM methods are presented by Kumar and Mitra (Kumar & Vijayalakshmi, 2011; Mitra, Pal, & Mitra, 2002). Besides, the key areas of application in education, classified by the EDM method, are shown in Table 1.

Table 1: Educational Data Mining Algorithms and Applications
(Kumar & Vijayalakshmi, 2011, p. 154)

Method	**Applications in Education**
Prediction	• Detecting student behaviors • Developing domain models • Predicting and understanding student educational outcomes
Clustering	• Discovery of new student behavior patterns • Investigating similarities and differences between schools
Relationship Mining	• Discovery of curricular associations in courses and sequences of courses
Discovery with models	• Discovery of relationships between student behaviors, and student characteristics or contextual variables • Analysis of research question across a wide variety of contexts
Distillation of Data for Human Judgment	• Human identification of patterns in student learning, behavior, or collaboration • Labelling data for use in later development of prediction models

The paper “Top 10 algorithms in data mining” presented a ranking of the ten best and most influential DM algorithms (Wu et al., 2008). Therefore, the first decision we made was to use this proposal to initiate our research. From the proposed methods, we excluded CART and C4.5, two popular classification algorithms based on induction trees, because they are specific of the decision trees algorithms which were also proposed in this paper. The eight remaining algorithms are:

- Decision Trees: Organize data forming branches of influences for decision making. The tree trunk represents the initial decision. This decision starts with a yes and no question; for example, if the student will graduate or not. The next divergent branches are graduation and no graduation, and each further election should have their own divergent branches that conduct to an endpoint. One widespread decision tree is J48 which is an open source implementation of Ross Quinlan's C4.5 decision-tree classification algorithm (A. Kumar & Sahni, 2011).
- K-Means Algorithm: Divides data collected in clusters separated by common characteristics based on group analysis.
- Apriori Algorithm: Controls transactional data. For example, the algorithm could predict which courses might be taken together by students in a semester.
- Expectation–maximization (EM) Algorithm: Defines parameters by analysing data and predicts the possibility of a future output or a random event within the parameters of data. For example, the EM algorithm could intent to predict the graduation rate of a cohort of students based on the analysis of data of previous cohorts.
- Page Rank Algorithm: Ranks and estimates the relevance of a piece of data within a big set of data. An example is the ranking of a website within a big set of all the websites on the internet. This algorithm was originally created for search engines.

- Adaboost Algorithm: Anticipates the behavior using observed data in order to be sensitive to statistical extremes.
- Nearest Neighbour Algorithm: Recognizes patterns in the location of data and associates the data with a bigger identifier.
- Naive Bayes Algorithm: Predicts the output of an identity-based on the data of known observations. For example, if the height of a person is 1.50 meters and the size of the shoes is 7, the algorithm could predict with a determined probability that a person is a woman.

Method

Setting

The case study presented in this work is based on a dataset obtained from the data warehouse of a higher education institution. The definition of the dataset was created after a requirement analysis with the stakeholders. The primary analysis is realized to discover and predict trends in the graduation rate performance indicator; the goal is to detect on time students that have a risk to dropout and therefore, take corrective actions so they can graduate. The data analysed was limited to records of computer science engineering students. The period of analysis was from 2002 to 2015.

Data Source

The data source contains personal and educational information collected at the beginning of the studies from the academic system of the university, and data that was collected during the course of the studies. Data selection and pre-processing are performed using dissimilar criteria for the representation and application of classification algorithms such as decision trees, Bayesian networks, and decision rules. The influential class defined for the analysis is “Graduation”. This indicator is evaluated according to data of enrolment: Students who started their studies and finished within the regular period of studies are those who contribute to the graduation rate. For example, if there were 20 students who started the studies in a school and only 15 students finished in the regular period then the graduation rate is 75%. This definition is used by the boards of academic evaluation, accreditation, and quality assurance of higher education in the country of the university (Ceaaces, 2015). This evaluation board is a public entity that carries out continuous evaluation and accreditation processes for higher education institutions. The detailed process of this research is shown in the following KDD steps.

Data Analysis

The KDD method was used in this work for the knowledge discovery process and the CRISP-DM method for EDM experimentation. This section shows the development of the experiments using the KDD process; therefore, the following steps are described: Data Selection, Data Pre-processing, Data Transformation, Data Mining, and finally Interpretation and Evaluation.

Data Selection

The data used for this analysis was previously collected in a data warehouse of the higher education institution. The next step was to filter the relevant information obtained from a requirement analysis with the stakeholders and create an external table to the data warehouse with the necessary fields for the analysis of graduation rate. In this step, the data from the data warehouse from the year 2002 to the year 2015 was analysed. The focus group for the analysis was the students of the computer science engineering faculty. The total number of students in this group was 441 students. From these students, 330 entered the university from the first semester, and the remaining 111 students came from other universities by means of validating courses and subjects.

There were two phases of analysis. The first phase was done including dropout fields and the results were presented in the Research in Engineering Education Symposium 2017 held in Bogota Colombia (O. Moscoso-Zea, Vizcaino, & Luján-Mora, 2017). In the experiments carried out in the latter paper, J48 was the algorithm that had better results for student dropout. In the case of the graduation rate, the best algorithm was random tree. In all these cases, the comparison was made with the percent of correct and incorrect classification. We concluded in the symposium paper that the J48 was the algorithm that obtained the best results in the experiments after the visualization of the trees.

The second phase of analysis which is presented in this paper was performed updating the original dataset. In the new dataset, we remove the dropout field to improve precision. We remove this attribute because it does not contribute to the graduation analysis. After that, the tests were performed using both the dataset and seven attributes to improve the accuracy of the results.

The comparison of both analyses presents the following results: the graduation rate showed that random trees perform better with the given dataset. However, since it was very difficult to observe facts from visualization of random trees, we once again recommended the J48 algorithm.

Data Pre-processing

After having the information of the focus group of analysis, the data was cleaned, verifying formats and checking that information was correct. This process was performed using Microsoft SQL Server Integration Services and Microsoft Analysis Services.

Data Transformation

In order to create the dataset, the ETL process was performed. The ETL process included an extraction process in which data were extracted from the data warehouse. Different dimensions and fact tables were the sources of these data. The ETL process included as well a transformation process that was a less complicated activity due to the fact that a transformation process was previously performed for the creation of the data warehouse. However, different SQL operations (aggregation and normalization) were executed in the information in order to structure the dataset fields with yes, no or 1, 0 in order to facilitate further analysis. Once the data were transformed, in the final step of the ETL process data were loaded into the newly created table (dataset). The dataset was the main input of DM in the different analysis tools used in this investigation. This dataset contained information of 18 attributes of students as shown in Table 2. This study was applied to use classifiers in the graduation class.

Table 2: Initial dataset for analysis of student attributes and variables

Attribute	**Datatype**	**Description**
STUDENTID	Varchar	Student identifier
STUDENT_NAME	Varchar	Student First Name
STUDENT_LASTNAME	Varchar	Student Las Name
CAMPUSID	Number	Campus Id
HIGHSCHOOL_ID	Number	High school Id from student
MARITAL_STATUSID	Number	Id for Marital Status
MARITAL_STATUS	Varchar	Marital Status
SCHOOL_ID	Number	School or Career of Student
DISABILITY_ID	Number	Id that shows if the student has a disability
FIRST_LEVEL	Varchar	Shows if a student started university from the first level or came from another university
FINISH_STUDIES	Varchar	Shows if a student finished its studies but did not finish the thesis
GRADUATE	Varchar	Shows if the student finished the studies and the thesis
ENROLLMENT_STATUS	Varchar	Shows the enrolment status of the student
HIGHSCHOOL_TYPE	Varchar	Shows if the high school is public or private
SEX	Varchar	Sex of the student
PROVINCE_ORIGEN	Varchar	Province where the student was born
ENROLLMENTYEAR	Number	Year of enrolment
NUMBER OF PERIODS	Number	Number of semesters the student have enrolled

As previously commented, the objective behind the analysis was to detect on time students that have a risk to dropout and therefore, take corrective actions so they can graduate. Moreover, with these experiments, we will have a better roadmap in the future to choose which algorithm to use in the experiments that predict educational events and to improve decision making. There are other variables that will be incorporated in the future to understand problems with the teaching or learning process.

Data Mining

The DM process starts with the creation of a comma-separated file from the resulting dataset. This file is the input for the three tools used in our experiments: WEKA, Orange 3 and Rapid Miner. These are three of the most used free software tools for general DM that are available today and they offer most of the desired features for a fully-functional DM platform (Jović, Brkić, & Bogunović, 2014). After performing a feasibility analysis realized to these tools, WEKA was selected as the tool considered for this study.

Interpretation and Evaluation

As explained previously, in the IEEE conference of 2006 (Wu et al., 2008) it was stated that one of the best methods for EDM is classification (decision trees, Naïve Bayes, meta-classifiers), therefore, this method was implemented in this analysis. Many algorithms were tested using supervised and non-supervised filters applied to the dataset. As was mentioned previously WEKA was the technological tool used for analysis. WEKA works with different classifiers that can be chosen in the tool.

In this work, we have used four classifiers and five algorithms for the experiments (see Table 3).

- Decision Tress are graphical structures in which each internal node represents a condition on an attribute and each branch represents the result of the condition, in this work we have used random trees and the J48 algorithm.

  - Random Tree Algorithm is a supervised classifier that chooses randomly the attributes at each node of analysis and allows class probabilities. A random tree presents uniform trees drawn "at random" which means that each tree has an equal chance of being tested with arbitrary permutations.

  - J48 algorithm generates rules for the prediction of the target variable using decision trees.

- Rule classifiers as OneR are one of the simplest and fastest classifiers, although sometimes its results are good compared to much more complex algorithms. OneR generates a rule for each attribute and chooses the one with the minimum error.

- Bayesians classifiers as Naïve Bayes which starts with the hypothesis that all the attributes are independent of each other. Bayesian classifiers maximize the probability that a new instance of the dataset is correctly classified by presenting a probabilistic measure in the classification results.

- Metaclassifiers, generally metaclassifiers are considered complex classifiers composed of simple classifiers that include some pre-processing of the data. Stacking is based on the combination of models to build a set of different learning algorithms with different learning sets.

We have performed four tests with the five algorithms:

(1) With all the attributes of the dataset (Cross Validation),

(2) With seven attributes (Cross Validation),

(3) With all the attributes of the dataset (Percentage Division), and

(4) With seven attributes (Percentage Division).

The seven attributes chosen on test 2 and 4 to improve precision were: marital_status, first_level, finish_studies, graduate, sex, enrollmentyear, and number_of_periods.

Results

After applying the four tests in the five algorithms we selected the one that performed best in the experiments. The results for the final analysis are consolidated in Table 3. We do not include the results of all the tests and all the algorithms in this work because the preliminary tests were not conclusive; however, they helped us to fine-tune the process until satisfactory results were obtained.

Table 3: Comparison of different algorithms within graduation rate dataset, ordered by “Well ranked instances (%)”

Algorithm	**Classifier**	**Test Mode**	**Well ranked instances (%)**	**Badly-ranked instances**	**Kappa Index**	**Absolute Error (%)**
Random Tree	With 7 dataset attributes	Cross-validation	96.00%	4.00%	0.67	0.05%
J48	With 7 dataset attributes	Percentage Division	95.69%	4.30%	0.70	0.05%
Naïve Bayes	With all dataset attributes	Percentage Division	95.46%	4.53%	0.69	0.06%
One R	With 7 dataset attributes	Cross-validation	94.78%	5.21%	0.68	0.05%
Stacking	With all dataset attributes	Percentage Division	93.33%	6.66%	0.00	0.13%

The sixth column in Table 3 describes the kappa index. The kappa index considers the observed agreement with respect to a baseline agreement, in this case, is a measure between the categories predicted by the classifier and the true categories observed. This index takes into account the possible concordances due to chance. If the value is 1 then the list is classified in a complete agreement. If the value is greater than 0 then the list is classified with a degree of agreement better than chance. If the value is 0 then the list is classified randomly. The absolute error indicates the percentage of error that may exist in the classification predicted by the classifiers.

In these experiments, the random tree algorithm (96% of precision) and the J48 (95.69% of precision) were the algorithms that better performed for graduation rate. After the visualization analysis of both algorithms, J48 is suggested to be used in future experiments because results are easier to understand and read for researchers. One partial view of this analysis is shown in Figure 4. Figure 4 has in the root the base class and then starts classifying instances according to the values of the dataset variables. The view of the decision tree could help analysts to infer knowledge from the different branches observed. The random tree figure is unreadable as shown in Figure 5 and therefore, is not suggested for future experiments with this kind of datasets. The root on Figure 4 shows whether the student has graduated or not from the computer science school. The next branch shows if the student dropped out from any course in the previous levels of studies and therefore, is eliminated when obtaining the graduation rate. Some of the discoveries of knowledge from the experiments with EDM are:

(1) The graduation rate is higher for students born and living in the same city, and for students who came from another higher education institution and that have validated some subjects from the course plan.

(2) Students that attend a high school in public institutions, married students, and those who lose a scholarship have a higher risk of not graduating.

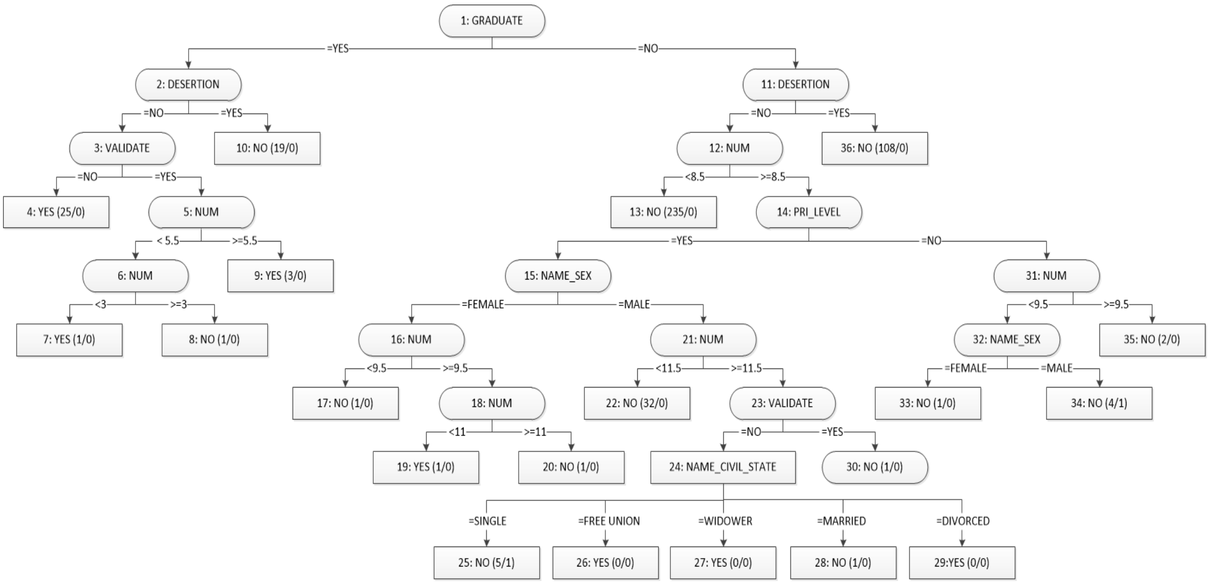

Figure 4: Partial tree view of the J48 analysis for graduation rate

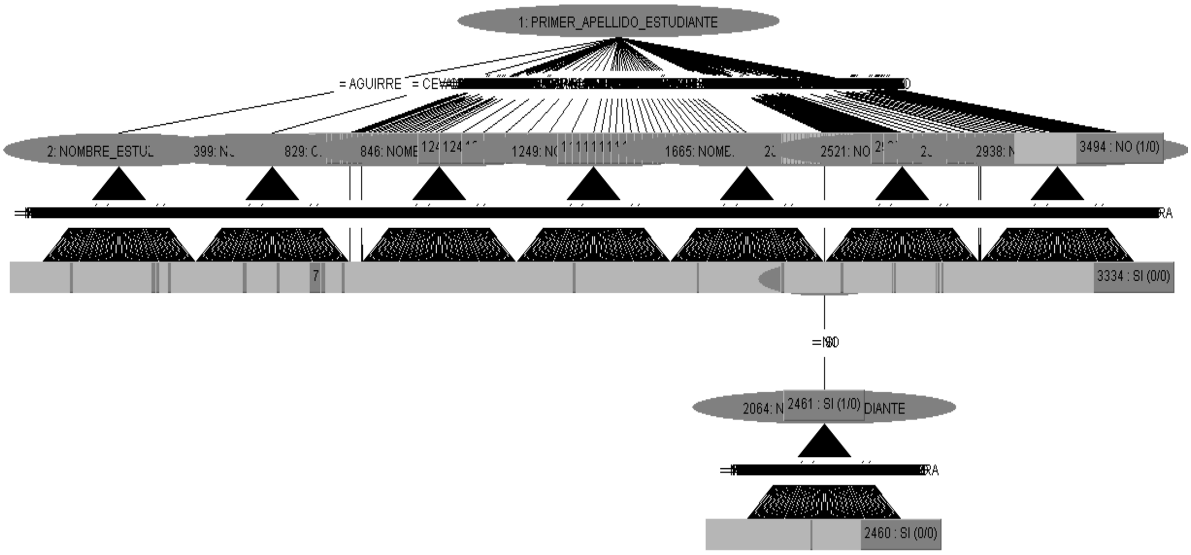

Figure 5: Tree view of the random tree analysis

Conclusions and Future Work

Predicting students' graduation rate to improve performance could be determined for most higher education institutions to help educators and learners to improve their learning and teaching process. This paper presents a wider view of EDM aiming to be a good source of information for researchers wishing to experiment with EDM in areas such as evaluation, enrolment, planning, student welfare, marketing, etc. EDM is projected as an essential discipline for university management that gives visibility to managers to improve decision making. By using DM tools such as WEKA, an evaluation of EDM methods and algorithms were performed.

The analysis performed in both phases of experimentation shows that random trees had better precision in the experiments. However, once the results were visualized it was very complex to analyse random trees (see Figure 5) due to a large number of edges in the picture. Therefore, we recommend the use of the J48 algorithm (see Figure 4) in future experiments with similar datasets since it presented a better and more comprehensive visualization and almost similar performance than random trees. Although it is difficult to understand Figure 4 as it is shown on this work, the original image was produced by software that allows zoom in and zoom out.

Some potential limitations of the work are that the analysis is done only to classification algorithms. Nevertheless, researchers recommend the use of these methods due to the fact that they perform better in these scenarios. It can be said, that generally, predicting students’ success can help educators and learners to improve their learning and teaching processes and identify students at risk.

The knowledge created with these experiments allows the institution to identify students’ at risk in early stages to take better decisions and corrective actions for educational improvement and management procedures. Furthermore, it precipitates actions and enables instructional choices for both, the student and the faculty. It also gives information on the groups that are more likely to graduate and the groups that are not. With the output of the experiments carried out, new strategies can be implemented to improve academic indicators.

This research does not include any statement of warning regarding some critical issues associated with data such as the valid use of the data with a focus on its interpretations and decisions made from such data, including any unintended misuse. Also, the ethical use of such data, because the interpretations can potentially lead to the creation or reinforcement of stereotypical views and potentially discrimination. And finally, the safe use of such data, which potentially includes sensitive information about the students.

In conclusion, future work arises for an extended research by including some critical issues, such as ethical and safe use of the data. The meta-analysis on predicting students' performance motivate us to carry out further research that can be applied in the faculty, and furthermore, in the entire university. This future work could become a guideline for higher education institutions to help them monitor students' performance in a systematic way.

Acknowledgments

This paper was based on a preliminary version published as "Evaluation of Methods and Algorithms of Educational Data Mining" (Moscoso-Zea, Vizcaino, & Luján- Mora, 2017), in the 2017 Research in Engineering Education Symposium (REES2017). In the REES version, two key performance indicators were considered: students' dropout and graduation rate.

References

Baepler, P., & Murdoch, C. J. (2010). Academic Analytics and Data Mining in Higher Education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2), 1–9.

Baker, R., & Siemens, G. (2014). Educational Data Mining and Learning Analytics. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 253–272). Cambridge University Press. http://doi.org/10.1017/CBO9781139519526.016

Bichsel, J. (2012). Analytics in Higher Education Benefits, Barriers, Progress and

Recommendations. Retrieved January 5, 2019, from https://library.educause.edu/~/media/files/library/2012/6/ers1207.pdf?la=en

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Retrieved January 30, 2019, from https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Ceaaces. (2015). *Modelo genérico de evaluación del entorno de aprendizaje de carreras presenciales y semipresenciales de las universidades y escuelas politécnicas del ecuador*. Quito.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rudiger, W. (2000). CRISP-DM 1.0. Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–37. http://doi.org/10.1609/AIMAG.V17I3.1230

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Second International Conference on Knowledge Discovery and Data Mining*. Retrieved from www.aaai.org

Huebner, R. A. (2013). A Survey of Educational Data-Mining Research. *Research in Higher Education Journal*, *19*, 1–13.

Jović, A., Brkić, K., & Bogunović, N. (2014). An overview of free software tools for general data mining. *37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO*, (March), 1112–1117. http://doi.org/10.1109/MIPRO.2014.6859735

Kumar, A., & Sahni, S. (2011). A Comparative Study of Classification Algorithms for Spam Email Data Analysis. *International Journal on Computer Science and Engineering (IJCSE)*, *3*(5), 1890–1895. Retrieved from http://www1.ics.uci.edu/~mlearn/MLRepository.html.

Kumar, S. A., & Vijayalakshmi, M. N. (2011). A Novel Approach in Data Mining Techniques for Educational Data. In *3rd International Conference on Machine Learning and Computing (ICMLC 2011)* (pp. 152–154).

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, *13*(1), 3–14.

Morris, L., Wu, S.-S., & Finnegan, C. (2015). Predicting retention in online general education courses. *American Journal of Distance Education*, *19*(1), 23–36.

Moscoso-Zea, O., & Lujan-Mora, S. (2017). Knowledge management in higher education institutions for the generation of organizational knowledge. In *Iberian Conference on Information Systems and Technologies, CISTI*. http://doi.org/10.23919/CISTI.2017.7975823

Moscoso-Zea, O., Sampedro, A., & Luján- Mora, S. (2016). Datawarehouse design for Educational Datamining. In *15th Information Technology Based Higher Education and Training (ITHET)* (pp. 1–6). Istanbul - Turkey.

Moscoso-Zea, O., Vizcaino, M., & Luján-Mora, S. (2017). Evaluation of methods and algorithms of educational data mining. In *2017 Research in Engineering Education Symposium, REES 2017* (pp. 972–980). Bogota.

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining Knowledge Discovery*, *3*(1), 12–27.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. http://doi.org/10.1007/s10115-007-0114-2