ACADEMIC PERFORMANCE PROFILES: A DESCRIPTIVE MODEL BASED ON DATA MINING

David L.la Red Martínez, Doctor
Marcelo Karanik, Doctor
Mirtha Giovannini, Specialist
Noelia Pinto, Engineer
Resistencia Regional Faculty, National Technological University, Argentine

Abstract

Academic performance is a critical factor considering that poor academic performance is often associated with a high attrition rate. This has been observed in subjects of the first level of Information Systems Engineering career (ISI) of the National Technological University, Resistencia Regional Faculty (UTN-FRRe), situated in Resistencia city, province of Chaco, Argentine. Among them is Algorithms and Data Structures, where the poor academic performance is observed at very high rates (between 60% and about 80% in recent years). In this paper, we propose the use of data mining techniques on performance information for students of the subject mentioned, in order to characterize the profiles of successful students (good academic performance) and those that are not (poor performance). In the future, the determination of these profiles would allow us to define specific actions to reverse poor academic performance, once detected the variables associated with it. This article describes the data models and data mining used and the main results are also commented.

Keywords: Academic performance profiles; data warehouses; data mining, knowledge discovery in databases

Introduction

The speed of technological developments in the new information society or cyber society involves a large number of questions about technical, economic, sociological, cultural and political order (Joyanes Aguilar, 1997). One of the most important doubts is whether education systems are able to provide the quantity and quality of professionals in order to satisfy the need of highly qualified personnel of this information and knowledge society (IKS), especially in areas related to ICT (Information and Communication Technologies).

Therefore, in the university education field, we propose the constant challenge of maintaining academic quality (even improve it). Due to this, we are constantly revised contents, strategies and teaching methods in search of to ensure appropriate quality standards which result in the formation of highly qualified professionals useful to society.

Clearly, academic performance is a critical factor take into account that, frequently, underachievement is associated with a high dropout rate.This is precisely what has been repeatedly observed in subjects of the first level of the Engineering in Information Systems career (ISI) of the National Technological University Resistencia Regional Faculty (UTN-FRRe), located in the city of Resistencia, Chaco province, Argentine,including Algorithms and Data Structures where underachievement is observed at very high rates (between 60% and 80% in recent years).

Specifically, academic performance is defined as the productivity of the individual, qualified by their activities, features and more or less correct perception of the assigned tasks (Forteza, 1975). Academic achievement is often debated, because there are many factors that can affect student performance, and determine precisely the performance is not a trivial task. Academic performance is affected by a multitude of heterogeneous factors (internal and external) that influence student performance. Therefore, evaluating the performance of all students in the same way does not provide information that can be used to detect and correct cognitive, apprehension, and discernment problems, etc. Then, a valid alternative is to try to establish whether there are common characteristics to certain groups of students. Thus, the profiling becomes a strategy of significant value when we should take action to improve the performance of students.

Profiling is a widespread activity in many areas, and it is analogous to the process of identifying and classifying patterns. Nowadays, there are many methods to determine and classify patterns used in the area of Artificial Intelligence and Machine Learning (Mitchell, 1997). These algorithms return value information when making decisions. The information is organized in large data warehouses (DW) that, after pre-cleaning, it is analyzed by algorithms that perform data mining (DM).

In this context, you can use the analysis process in two clearly defined approaches: descriptive or predictive. In the first approach, we try to characterize situations so you can understand the reason for what it is the main variable that describes the behavior of a particular situation. The second approach involves the use of the model to establish in advance a problematic situation.

For the profiling of academic performance, a descriptive model can explain the variables that affect student performance. In this way we can have multiple perspectives on the same problem and treat problems having a global vision. However, the predictive model attempts to establish future problematic situations. That is, the model tries to determine early on the performance profile of a student (according to the processed data) in order to carry out actions that help to correct any problems.

In this paper, we propose the use of DM techniques on information about student performance of the Algorithms and Data Structures professorship, in Information Systems Engineering career that is dictated in the Resistencia Regional Faculty at the National Technological University (Chaco, Argentine). This article is structured as follows: in Section 2 are detailed concepts and works related to the measurement of academic performance. The concepts related to DW and DM are presented in Section 3. In Section 4 are described the scope of the proposal and the model used. In Section 5 the results are shown. Finally, in Section 6 some conclusions are presented in relation to the work done.

Academic Performance

While academic performance characterizes the student individually, you can get some general characteristics of groups of students from it. This is where the importance lies in having reliable methods of performance evaluation. There are several ways to assess student achievement. In general, it involves determining the actual production of a student regarding formal activities. However, when we try to put into operation the performance, we choose the reductionism (González, 1998). Another way is to use indicators such as graduation rates, differentiated by types of institutions and analyzing student achievement from individual data (García & San Segundo, 2001) or through entry qualifications to university, performing the analysis of data using the statistical technique of ROC (Receiver Operating Characteristic) curve (Vivo Molina et al., 2004).

The cognitive aspects were the basis of the early research on the learning process; after researchers discovered the importance of affective components and their decisive influence on learning; finally the cognitive and affective aspects came together, giving birth to the construct called self-regulated learning (Herrera Clavero et al., 2004).

The university academic performance has also been studied using the production function approach to estimate the determinants of academic performance (Di Gresia, 2007). For example, the determinants of learning have been analyzed through a production function approach suggesting that school achievement depend on genetic and socioeconomic factors, quality of teaching, the conditions of the school and the group of students (Delfino, 1989).

It has been shown in several studies that the most related factor to educational quality are the studentsthemselves as co-producers, measured by household socioeconomic status where they come from (Maradona & Calderón, 2004) and it has shown that students productivity is higher for women, for younger students and those from households with more educated parents (Porto & Di Gresia, 2000), having great importance the relationship between hours worked and academic performance (Fazio, 2004).

There is a paradox regarding the availability of time to study. It has shown the contrast between those who work and study and who only studies, finding no differences in academic performance of the two sets (Reyes, 2004). All this result in the fact that, in general, empirical studies confirm the correlation between higher levels of education and positive attributes after studies (McMahon, 2002).

There are several studies using mathematical techniques for performance evaluation. In this sense, we studied the ability of linear and logistic regression in predicting the performance, and success or academic failure, based on variables such as attendanceand class participation. It was concluded that past performance is a good predictor of future performance and that attendance and specially participation are variables with significant weight in predicting performance (García Jiménez et al., 2000).

It has also been shown that variables such as study planning, intelligence, teacher support, study, time, environmental conditions of study, and involvement were part of the prediction equation of multiple regressions, which explain 25.70% of variance of academic performance in high school (Marcelo García et al., 1987).

Clearly, having adequate mathematical methods is an advantage when assessing the performance. The issue is to use them properly. The problem of finding good predictors of future performance so that, academic failure is reduced in graduate programs, has received special attention in the US (Wilson & Hardgrave, 1995), from which it follows that the classification techniques such as discriminant analysis or logistic regression are more appropriate than multiple linear regression predicting success or academic failure.

The diversity of studies on academic performance shows that there is no single way to evaluate it. The diversity of studies on academic performance shows that there is no single way to evaluate it. Moreover, problems can vary depending on the regional context and the social reality in which the student is inserted. That is, there are no tools that can be applied to all areas and the results may not be extensible to explain all possible situations. This clearly indicates the need to identify profiles in specific educational institutions by adapting the tools to each particular situation. For purpose of this article we opted for data mining techniques applied to a data warehouse loaded with socio-economic and academic information for students of the subject mentioned above.

Data Handling

The correct data organization added to a suitable model of managing them, can provide a clear view of the drawbacks in the performance of students. In this sense, there are tools in the area of Artificial Intelligence, specifically to the Business Intelligence (BI), such as Data Warehouses (DW) and Data Mining (DM), used to discover hidden knowledge in large volumes of data that can be used to determine patterns and profiles properly.

In the following subsections we make a brief review of the most remarkable aspects of these tools storage and data processing.

Data Warehouse

A DW is a collection of data-oriented topics, integrated, nonvolatile, time variant, which is used to support the process of managerial decision making (Inmon, 1992), (Inmon, 1996), (Simon, 1997). Because a DW cannot be acquired, it must be built according to certain methodology. The technique used in the creation of DW depends on to whom it focuses as main point its development, it can be towards data management, goals or users (Data-Driven, Goal-Driven and User-Driven), (Gutting, 1994).

Data-Driven: Unlike classic management systems of user requirements, this approach considers that in a DW what you handle are data, considering the needs of users in second term (Poe, 1996). The data model consists in few dimensions and groups of events. The dimension represents the basic structure of the design. The facts are based on time and have some level of granularity.

Goal-Driven: The goals and objectives in principle guiding the development process and, unlike Data-Driven model, it contains more dimensions and few facts, which are based on time and have a low level of granularity.

User-Driven: It considers that the main factors to take into account are the needs of users, because they are the ones who finally use the system. The model consists of a few facts, which have a moderate level of granularity.

Regardless of development models mentioned, the methodologies to be followed for the development of DW depend largely on the size of DW to create and the promptness with which the DW is required. Two main methodologies for the development of a DW are the Rapid Warehousing and Big Bang.

*Rapid Warehousing***,** also known as evolutionary or incremental methodology, considers that the construction and implementation of a DW is an evolutionary process, where part of a DW is created with the integration of data marts (Widom, 1995). In contrast, the *Big Bang* tries to resolve all known problems to create a large DW, before releasing it for evaluation and testing (Harinarayan et al., 1996).

The functional blocks that correspond with a comprehensive information system that uses a DW are summarized in Figure 1:

- *Operational level:*refers to the operational and transactional systems of the organization and sources that are part of the process of Data Warehousing.
- *Access level to information:* it is the user interaction layer whose purpose is to convert the data stored in easy and transparent information to the end users tools.
- *Access level to data:*it communicates the level of access to information with the universally operational level.
- *Level of directory data (metadata*): it is a repository of metadata from stored data that provide information about the origin and transformation of them in the process of Data Warehousing.
- *Level process management*: planning of tasks and processes for building and maintaining the updated Data Warehouse.
- *Level of application message:*it determines the information transport throughout the computing environment of the organization as a middleware but beyond purely network protocols.
- *Data Warehouse level (physical):*it is the highly flexible central repository of information where copies of operational data or external optimized reside for access to the query.
- *Level data organization:*it includes all the necessary processes to select, edit, summarize (usually summarize), combine and load into the Data Warehouse, and in the access Layer to information operational and external data

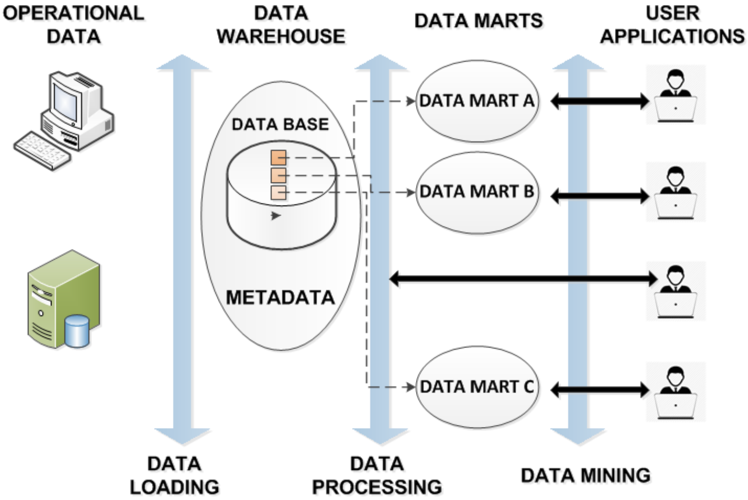

Figure 1. Basic architecture of a DW.

Clearly, the model and the methodology to be used depend on the scope of application, and they are very important decisions when working with these technologies. This type of data warehouse provides adequate support for the implementation of techniques pursuit of knowledge and it is suitable for determining student profiles, standing out for this purpose using data mining techniques.

Data Mining

DM is the stage of knowledge discovery in databases (KDD). It is the consistent use of specific algorithms that generates a list of patterns from pre-processed data (Fayyad et al., 2001), (Hand et al., 2000), (Frawley et al., 1992). DM is closely linked to the DW since they provide historical information with which mining algorithms obtain the information needed for decision-making (IBM Software Group, 2003). The set of data analysis techniques that allow us to extract trends, patterns and regularities to describe and understand in a better way the data are part of DM. It also allows extracting patterns and trends to predict future behavior (Simon, 1997), (Berson & Smith, 1997), (White, 2001).

As it was mentioned above, using DM, descriptive and predictive models can be generated (Agrawal & Shafer, 1996). Even though the techniques used are several, one of the most prominent is the clustering (or data pooling) (Grabmeier & Rudolph, 1998), (Ballard et al., 2007).

Currently there are several DM methodologies; the most widespread are the CRISP-DM and SEMMA. The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology is organized into six stages. At the same time, each stage is divided into various tasks (Chapman et al., 1999). The six stages are:

- *Understanding of business:*this includes understanding the project objectives and demands from the business perspective.
- *Understanding of data:*it involves collecting initial data and continues with activities that let us discover relevant subsets to form hypotheses related to information hiding.
- *Data preparation:*this covers all activities necessary to build the final dataset from initial raw data.
- *Modeling:*it involves selecting several modeling techniques and parameters calibrated to optimal values.
- *Evaluation*: it is to thoroughly assess the model and revision the steps executed to create and compare the model properly with established objectives.
- *Deployment:*this goes from the generation of a report to implementing a repeatable data mining process through the organization.

This model is an idealized sequence of events. In practice, many of the tasks can be performed in a different order and even require going back to some of them. The SEMMA methodology is used to discover unknown patterns business. The name refers to the five basic stages (Sample, Explore, Modify, Model and Assess) (Matignon, 2009), (SAS Institute, 2009). The stages consist of the following:

- *Sample:*Get a subset of the data sufficiently representative to obtain useful information from them and process them easily.
- *Explore:*Look for patterns in the data.
- *Modify:*Create and transform variables or even eliminate those that are unnecessary.
- *Model:* Select and apply the model that best fits the data.
- *Assess:*Determine if the results are useful and reliable. To test the results against known data.

Both methods provide an adequate framework to address the profiling of students. With these methodologies we try to explain the behavior of certain variables and to identify relevant issues within the academic performance, for which a model was developed we describe below.

Used Model

In this section, the scope will be described, and it is also briefly the data model project profiling of students according to their academic performance; a more detailed description can be seen in (La Red Martínez et al., 2014a, 2014b).

Scope of Application

The scope of application of the draft determining profile of students was the Resistencia Regional Faculty of the National Technological University (UTN-FRRe), where the Engineering in Information Systems is dictated. This career has several subjects in the first year that are specific to the profession of engineer systems and their contents are not addressed during high school. This causes that the student finds a certain amount of new issues, ignoring how to address them. These subjects usually are those that generate the greatest attrition in the early year's career, since its approval is compulsory to pursue other subsequent years. One of these subjects is Algorithms and Data Structures, whose study involves a high content of logical procedures and handling of structures for different types of variables.

In this sense, we sought to determine using techniques of Data Warehouse and Data Mining, the variables that explain the unequal academic performance by students of Algorithms and Data Structures of Engineering in Information Systems career of the UTN-FRRe, with the purpose of setting actions to improve such academic performance of students.

In this context, the results achieved in the assessments made during the course completed in 2013 are considered as academic performance (loading, filtering and information processing was performed in 2014). We sought to determine to what extent unequal academic performance of students in Algorithms and Data Structures is influenced by the following variables: middle school of origin, educational level of parents, socio-economic status, age, gender, general attitude towards study, existence of course for entry, study regimen (annually - quarterly), use of supporting tools (virtual campus).

Low, medium, and high student profiles of achievement were searched using data mining on a data warehouse.

In similar work (La Red Martínez et al., 2010, 2012), a model of data analysis was proposed to integrate academic and contextual information. In the next section it is described the analysis model of data used in this work.

Data Model

As it was mentioned in the previous section, the overall objective was to determine, using techniques DW and DM, the variables that explain the unequal academic performance by students of Algorithms and Data Structures in Information Systems Engineering career from UTN-FRRe. To achieve this, the following activities are performed:

- Gather information on the current situation regarding the academic performance of students.
- Filter and debug information in the current databases.

- Establish the relevant variables to describe the situation under study.
- Determine how each of the variables that were set to assess the situation of the student.
- Determine how each of the variables that were set to evaluate the academic context affects.
- Establish actions aimed at improving the academic performance indices of students.

Using the User-Driven technique we have pursued to determine performance profiles (low, medium and high) based on the results obtained by students in assessments, and then, relationships and correlations were look for among the variables mentioned in the previous section.

In the first instance, the information for the students of the Algorithms and Data Structures course was taken from the database of the academic system, from which the specific data of students and their grades were extracted, those that were considered as indicators of academic performance. Data under the socio-economic situation of the student and his family, as well as attitudinal aspects regarding the study and ICT were collected through a survey using a system of forms in an application online. This information was preprocessed, making a cleanup of inconsistent and missing data. The universe was made up of students able to study the subject in 2013 (we are working on the reporting burden of the course of 2014 and previous years, about 300 students per year) and the unit of analysis was each of those students. The cases selected were students able to study the subject and who regularly attend classes (we have not considered those students who enroll to study but then they do not study).

DW structure, as it is shown in Figure 2,consists of a fact table and several dimension tables.We can distinguish two types of columns in a fact table: fact columns and columns keys. The fact columns store the issue measuresto be controlled and the columns keys are part of the key table.

A table of dimensions or dimensional entity is a table or entity that stores details about facts. It includes descriptive information about the numerical values of a fact table. Furthermore, the dimension tables describe the various aspects of a subject under study. Each table of dimensions contains a single key and a set of attributes that describe the dimension. The columns in a dimension table are used to create reports or to display query results. For example, the textual descriptions of a report are created from the labels of the columns in a dimension table.

The model used in this work consists in the fact table called Students and more dimension tables associated with it, which include features that you want to study. In Figure 2, this structure is graphically represented.

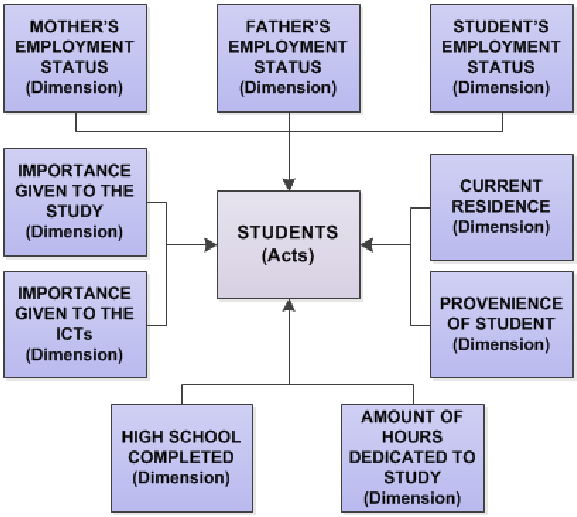
MOTHER'S EMPLOYMENT STATUS (Dimension)
FATHER'S EMPLOYMENT STATUS (Dimension)
STUDENT'S EMPLOYMENT STATUS (Dimension)
IMPORTANCE GIVEN TO THE STUDY (Dimension)
STUDENTS (Acts)
CURRENT RESIDENCE (Dimension)
IMPORTANCE GIVEN TO THE ICTs (Dimension)
PROVENIENCE OF STUDENT (Dimension)
HIGH SCHOOL COMPLETED (Dimension)
AMOUNT OF HOURS DEDICATED TO STUDY (Dimension)

Figure 2. DW model used.

The fact table includes specific student information and academic performance, while the dimension tables contain information that makes the description of socio-economic background of the student and family, their academic background in high school, and their attitude towards study and towards ICT.

To carry out this work we used the suite of tools included in the IBM Data Warehouse Edition (DWE) V.9.5, with robust business intelligence infrastructure from IBM. It consists in several products but for the purposes of this paper we have used the following components:

- DB2 Database Management System Enterprise Server Edition (DB2), which allows multi-user connections and presents high scalability.
- Design Studio (DS), a platform as a tool used by business analysts and managers data warehouses to design rules of workloads, processing data flows and analytical flows for data mining and text analysis.
- Intelligent Miner (IM), a set of functionalities that allow performing an analysis of information according to concepts of Business Intelligence (BI).

It was considered interesting to study data with data mining techniques, both descriptive and predictive, for what it was used the CRISP-DM methodology, referred in section 3.2; the scheme of work was similar to described in (La Red Martínez et al., 2014c, 2014d, 2014e), adapted to the particularities of the UTN - FRRe. The developed actions are indicated in the following section, where DM processes performed are commented, and also the main results obtained.

Detection of Academic Performance Profiles

The work done was divided into several stages such as selection, purification and data preparation, data mining, description of the results, which are explained below.

Selection Stage

At this stage, different sources of information (internal and external) were selected which served as the basis for data mining stage. As a source of internal information, it was used the information from the corporate and professorship database, where the qualifications of the partial exams of students and their condition at the end of the completed study (Free, Regular, Promoted) are stored.

As it was raised at previous sections, for the study of the academic performance of students the results of the instances of planned evaluations per subject should not only be considered, but also other cultural, social and economic factors that affect student performance.

Therefore, to obtain external information was decisive direct participation of the student, because it was necessary to know information about personal issues that could not be achieved otherwise. To this end, a web application was used that allowed to have an On-Line Survey consists of questions related to family status, history of secondary education, among other issues, including student attitudinal aspects regarding the study and ICT.

Academic, socioeconomic and attitudinal data obtained in the above manner were used for the construction of DW which is then used for data mining processes.

Stage of Purification and Data Preparation

The quality of the patterns obtained with data mining is directly proportional to the quality of the data used (Spositto et al., 2010). Based on this, the objective of this stage was the detection, correction, and removal of anomalous data. This exploratory tasks was performed by analyzing each of the records stored in the Survey table (intermediate table from which, after this stage, the socioeconomic and attitudinal data were loaded into DW), achieving the elimination of inconsistencies through manual corrections, generated by the improper use of on-line system by students.

Once refined information obtained by each student, we proceeded to the manual loading of qualifications corresponding to the three midterms, examination recovery and the final condition of the student at the end of the course. It is necessary to clarify that if a student has not given any examination, the field was left blank.

As a final activity at this stage, and with full information, we proceeded to load DW through data streams from the Survey table. At the end of the process and before starting the next phase, 242 records were available.

Stage of Data Mining

At this stage, DM techniques were selected to use, creating corresponding miningflows, in which, the respective algorithms are parameterized.

Firstly, we have started with the supervised classification technique with decision trees, whose goal is to make classifications on known data, models which can then be used to predict or classify new or unknown values.

A Tree Classification algorithm is used to compute a decision tree. Decision trees are easy to understand and modify, and the model developed can be expressed as a set of decision rules. This algorithm scales well, even where there are varying numbers of training examples and considerable numbers of attributes in large databases. Decision Tree Classification generates the output as a binary tree-like structure (IBM, 2013).

A Decision Tree model contains rules to predict the target variable. The Tree Classification algorithm provides an easy-to-understand description of the underlying distribution of the data. The core algorithm for building decision trees called ID3 (Quinlan, 1986) which employs a top-down, greedy search through the space of possible branches with no backtracking.

Description of Results Achieved

The analysis of the results was based on consideration of the variable related to the final situation of the student as mining parameter, which reflects their status in the matter at the close of school term. It was considered as *Free* students, those that not approved neither midterms nortests to recover; *Regular*, who managed to approve 3 exams (by retrieving them or not) with greater than or equal to 60% grade, but did not reach at least 75% in all cases. Finally, the students in the *Promoted* state are those who approved all partial greater than or equal to 75% grade.

Taking into accountthe above, we have obtained the following results: 81.42% of students in Free condition, 10.62% as Regular student, and only 7.96% as Promoted student. Thus, and always by basing the analysis according to Status parameter, it was considered different criteria for grouping data for the description of classes:

- Dependence of secondary school.
- Number of hours dedicated to the study.
- Importance given to study.
- Academic level of their mother.

- Academic level of their father.
- Use of ICT.

Then, for better readability, in Table 1 we describe how all classes are characterized (Regular, Free and Promoted) taking into account the criteria of Secondary School Dependency (Private Faith, Provincial and Municipal, Private Particular, among others).

Class	Attribute	Percentage
Regular 10.62%	PrivateReligious	8%
	Provincial and Municipal	67%
	Private Particular	25%
	Others	0%
Free 81.42%	PrivateReligious	22%
	Provincial and Municipal	61%
	Private Particular	11%
	National	4%
	Others	2%
Promoted 7.96%	PrivateReligious	22%
	Provincial and Municipal	78%

Table 1. Characterization of Classes considering Dependency of High School.

Table 2 shows the characterization of all kinds of students (Regular, Free and Promoted) considering the criterion of Number of Hours Dedicated to Study for them.

Class	Attribute	Percentage
Regular 10.62%	Up to 10 inclusive	17%
	Over 10 Up to 20	33%
	Over 20	50%
Free 81.42%	Up to 10 inclusive	27%
	Over 10 Up to 20	27%
	Over 20	46%
Promoted 7.96%	Up to 10 inclusive	22%
	Over 10 Up to 20	22%
	Over 20	56%

Table 2. Characterization of Classes taking into account the Number of Hours Dedicated to Study.

Table 3 shows the characterization of all kinds of students (Regular, Free and Promoted) considering the criterion of Importance Given to the Study for them.

Class	Attribute	Percentage
Regular 10.62%	More than Work	33%
	More than Fun	50%
	More than Family	17%
Free 81.42%	More than Work	18%
	More than Fun	64%
	More than Family	17%
Promoted 7.96%	More than Work	11%
	More than Fun	89%

Table 3. Characterization of Classes taking into account the Importance Given to the Study.

Table 4 shows the characterization of all kinds of students (Regular, Free and Promoted) considering the criterion of Latter Studies of Mother.

Given the same previous criterion but analyzing the Father situation, the results are detailed in Table 5.

Class	Attribute	Porcentaje
Regular 10.62%	Complete High School	17%
	Complete Non-University Higher Study	17%
	Incomplete University Study	17%
	Complete UniversityStudy	25%
	No Answer	25%
Free 81.42%	No Studies	1%
	IncompleteElementarySchool	2%
	Complete ElementarySchool	5%
	Incomplete High School	8%
	Complete High School	23%
	Incomplete Non-University Higher Study	3%
	Complete Non-University Higher Study	15%
	Incomplete University Higher Study	15%
	Complete University Higher Study	17%
	PostgraduateStudies	7%
	No Answer	3%
Promoted 7.96%	Incomplete High School	22%
	Complete High School	11%
	Incomplete University Higher Study	11%
	Complete University Higher Study	33%
	PostgraduateStudies	22%

Table 4. Characterization of Classes considering Latter Studies of Mother.

Class	Attribute	Porcentaje
Regular 10.62%	IncompleteElementarySchool	8%
	Complete High School	17%
	Complete Non-University Higher Study	8%
	IncompleteUniversityStudy	25%
	Complete UniversityStudy	25%
	No Answer	17%
Free 81.42%	No Studies	2%
	IncompleteElementarySchool	3%
	Complete ElementarySchool	2%
	Incomplete High School	11%
	Complete High School	22%
	Incomplete Non-University Higher Study	2%
	Complete Non-University Higher Study	4%
	Incomplete University Higher Study	21%
	Complete University Higher Study	21%
	PosgraduateStudies	1%
	No Answer	11%
Promoted 7.96%	Complete ElementarySchool	11%
	Complete High School	22%
	Complete UniversityStudy	44%
	PosgraduateStudies	11%
	No Answer	11%

Table 5. Characterization of Classes considering Latter Studies of Father.

In the case of the use of ICT, the results are expressed in Table 6.

Class	Attribute	Percentage
Regular 10.62%	They are a reality today	8%
	They facilitate the process of teaching	50%
	It will be essential your domainfor professional practice	42%
Free 81.42%	They are a reality today	4%
	They facilitate the process of teaching	53%
	It will be essential your domainfor professional practice	42%
Promoted 7.96%	They are a reality today	11%
	They facilitate the process of teaching	56%
	It will be essential your domainfor professional practice	33%

Table 6. Characterization of Classes considering the use of ICT.

Finally, it is important to refer to the overall quality of the model used to classify the Final Status Student, which turned out to be 0.944, meaning that when estimating the situation based on the variables considered in the model, the estimate is correct in 94.4% of cases; global quality value of a model close to 0 indicates a very poor with respect to the model predictive accuracy, reliability and predictive, reliability and possibility of classifying data. A value of overall quality of a model close to 1 indicates an excellent model that correctly classifies data and records and it is the most reliable.

Conclusion

It will start with discussions and comments regarding the main results obtained so far, conclusions are then indicate themselves and it will end with the main lines of future work.

Comments and Discussions

The processes of educational data mining made have produced a considerable volume of information, whose detailed study will consume a considerable amount of time, not only to the members of the research project but other areas since, as it is supposed, academic performance is influenced by socioeconomic and cultural background of the students and attitudinal aspects of them regarding the study and use of ICT.

In the following comments are considered *high* academic performance to that achieved by students with final status of *promoted*, *medium* performance to students with situation of *regular*, and performance *low*, the situation of students with *free*; at the same time *academic success* to *high* and *medium* performance and *academic failure* to *low* performance are considered.

Next, some of the aspects considered appropriate to emphasize will be discussed.

Considering the type of secondary school from which students come (Table 1), it was observed that for all categories of academic achievement most students come from School of Provincial and Municipal level, but with significant differences in the percentages, high academic performance is considered as 78%, middle 67%, and low 61%. This indicates the type of secondary school the student attended is related to the academic performance achieved by it, observed that the highest percentage of participation of schools Provincial and Municipal level (State) falls under the category of higher academic performance.

In the face of the amount of hours per week that students devoted to the study (Table 2) was observed that 56% of those who have high academic performance have spent more than 20 hours per week to study, this percentage drops to 50% for the medium academic performance and 46% for poor academic performance. In addition, 22% of those who had a high academic performance have spent up to 10 hours per week to study, this percentage increases to 27% for poor academic performance. This indicates a direct relationship between the dedication to study and academic success.

Considering the importance that students give the study (Table 3), it was observed that 89% of those who have high academic achievement have given more importance to study than fun. This percentage drops to 50% for the average academic performance and 64% for poor academic performance, being 64.6% for the total population. In addition, 11% of those who have high academic achievement have given more importance to study the work, this percentage increase to 33% for the medium academic performance and 18% for poor academic performance. This indicates a relationship between academic success and the importance given to the study before the fun and work.

With respect to recent studies of the mother (the highest level) (Table 4), it was observed that 22% of those who have high academic performance have mothers with postgraduate studies, this percentage is reduced to 7% for poor academic performance, being 7.08% for the total population. In addition, 33% of those who had a high academic performance are children of mothers with completed university studies, this percentage decreases to 25% for the medium academic performance and 17% for poor academic performance. This indicates a relationship between academic success and the level of education achieved by the mother.

Considering recent studies of the father (the highest level) (Table 5), it was observed that 11% of those who have high academic performance have parents with graduate studies, this percentage is reduced to 1% for poor academic performance, being 1.77% for the total population. In addition, 44% of those who had a high academic performance are children of parents with completed university studies, this percentage decreases to 25% for the medium academic performance and 21% for poor academic performance. This indicates a relationship between academic achievement and educational level achieved by the father.

Taking into account the views of students on the use of ICT (Table 6) it was observed that 56% of those who had a high academic achievement felt that it facilitates the learning process, this percentage is reduced to 50% for the medium academic performance, being 53% for poor. In addition, 33% of those who have high academic performance considered that the domain of ICT for professional practice will be essential; this percentage rises to 42% for medium and low academic performance. This would indicate that most students with higher academic performance would be concentrated more on the teaching – learning than in the possible future exercise of the profession.

As described above it can be said that the success and academic failure are related to the type of secondary school that the student attended, the student's dedication in weekly hours of study, the importance given to study versus fun and work, educational level of parents and students perception of ICT.

Conclusion

This paper presents an efficient model for determining student profiles as academic achievement using data warehouses and data mining techniques. This will take actions that tend to reduce academic failure, early acting with a special accompaniment of students whose profile indicates high probability of academic failure. Clearly, the model presented in this paper is suitable for the determination of profiles and constitutes a valid tool for academic management. The proposed model can be implemented in various institutions.

Future Lines of Work

We are currently working on the evaluation of prior information completed in 2013, and corresponding to the course of 2014, for subsequent incorporation into DW and analysis with data mining models of classification with decision trees and clustering; it also begins the meeting of the information for the completed 2015. It will also continue studying the influence of other variables on academic performance, such as employment status of parents and the students themselves, the student health coverage, the employment relationship with the issue of career, etc.

Acknowledgements

This work was performed within the framework of the accredited research project: “Determination of profiles of students and academic performance by using data mining”, code 25 / L059 - UTI 1719.

References:

Agrawal, R.; Shafer, J. C. Parallel Mining of Association Rules. *IEEE Transactions on Knowledge and Data Engineering*. December. USA. 1996.

Ballard, Ch.; Rollins, J.; Ramos, J.; Perkins, A.; Hale, R.; Dorneich, A.; Cas Milner, E. &Chodagam, J. *Dynamic Warehousing: Data Mining Made Easy*. IBM International Technical Support Organization. IBM Press. USA. 2007.

Berson, A. & Smith, S. J. *Data Warehouse, Data Mining & OLAP*.McGraw Hill. USA. 1997.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Renartz, T., Shearer, C., Wirth, R. *CRISP-DM 1.0.Step-by-step data mining guide*. 1999.

Delfino, J. A. Los determinantes del aprendizaje. In Petrei, A. H., editor, Ensayos en economía de la educación. *Educational Evaluation and Policy Analisys*. 1989.

Di Gresia, L. *Rendimiento Académico Universitario*. Tesis Doctoral. Universidad Nacional de La Plata. Argentina. 2007.

Fayyad, U.M.; Grinstein, G. & Wierse, A. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann. Harcourt Intl. 2001.

Fazio, M. V. *Incidencia de las horas trabajadas en el rendimiento académico de estudiantes universitarios argentinos*. Documentos de Trabajo UNLP, 52. Argentina. 2004.

Forteza, J. Modelo instrumental de las relaciones entre variables motivacionales y rendimiento. *Revista de Psicología General y Aplicada*, 132, 75-91. España. 1975.

Frawley, W., Piatetsky-Shapiro, G. & Matheus, C. Knowledge Discovery in Databases: An Overview. *AI magazine*, 13(3), 57. 1992.

García, M. M.; San Segundo, M. J. *El Rendimiento Académico en el Primer Curso Universitario*. X Jornadas de la Asociación de Economía de la Educación. Libro de Actas, págs. 435-445. España. 2001.

García Jiménez, M. V.; Alvarado Izquierdo, J. M.; Jiménez Blanco, A. La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema* Vol. 12, Supl. nº 2, pp. 248-252. España. 2000.

González, A. J. Indicadores del rendimiento escolar: relación entre pruebas objetivas y calificaciones. *Revista de Educación*, 287, 31-54. España. 1998.

Grabmeier, J. & Rudolph, A. *Techniques of Cluster Algorithms in Data Mining version 2.0.* IBM Deutschland Informations system eGmbH. GBIS (Global Business Intelligence Solutions). Germany. 1998.

Gutting, R. An Introduction to spatial database systems. *VLDB Journal*, 3, 357- 399. 1994.

Hand, D.J.; Mannila, H. & Smyth, P. *Principles of Data Mining*. The MIT Press. USA. 2000.

Harinarayan, V., Rajaraman, A., Ullman, J. Implementation data cubes efficiently. *ACM SIGMOD Record*, 25 (2), 205 - 216. 1996.

Herrera Clavero, F. et al. ¿Cómo Interactúan el Autoconcepto y el Rendimiento Académico en un Contexto Educativo Pluricultural? *Revista Iberoamericana de Educación*. España. 2004.

IBM. *IBM Knowledge Center*. Retrieved Jan 6, 2013, from http://www-01.ibm.com/support/knowledgecenter/SSEPGG_9.7.0/com.ibm.im.model.doc/c_dataminingoverview.html?lang=en. 2013.

IBM Software Group. *Enterprise Data Warehousing whit DB2: The 10 Terabyte TPC-H Benchmark*. IBM Press. USA. 2003.

Inmon, W. H. *Data Warehouse Performance*. John Wiley & Sons. USA. 1992.

Inmon, W. H. *Building the Data Warehouse*. John Wiley & Sons. USA. 1996.

Joyanes Aguilar, L. *Cibersociedad*. McGraw Hill. España. 1997.

La Red Martínez, D. L.; Acosta, J. C.; Uribe, V. E.; Rambo, A. R. Academic Performance: An Approach From Data Mining. *Journal of Systemics, Cybernetics and Informatics*; V. 10 N° 1 2012, págs.66-72; USA. 2012.

La Red Martínez, D. L.; Acosta, J. C.; Uribe, V. E.; Rambo, A. R.; Cutro, A. L. *Data Warehouse y Data Mining Aplicados al Estudio del Rendimiento Académico*. CISCI (9na. Conferencia Iberoamericana en Sistemas, Cibernética e Informática); Memorias, Volumen I, págs. 289-294; ISBN N° 978-1-934272-94-7; Orlando, Florida, USA. 2010.

La Red Martínez, D. L.; Karanik, M.; Giovannini. *Determinación de Perfiles de Estudiantes y de Rendimiento Académico Mediante la Utilización de Minería de Datos en la UTN – FRRe*. XVI Workshop de Investigadores en Ciencias de la Computación - WICC; Universidad Nacional de Tierra del Fuego, Antártida e Islas del Atlántico Sur, Ushuaia, Tierra del Fuego, Argentina. 2014a.

La Red Martínez, D. L.; Karanik, M.; Giovannini, M.; Pinto, N. *Estudio del perfil de rendimiento académico: un abordaje desde Data Warehouring*. 2° Congreso Nacional de Ingeniería Informática / Sistemas de Información - CoNaIISI - 2014; ISSN N° 2346-9927; pág. 604-612; Universidad Nacional de San Luis, San Luis, Argentina. 2014b.

La Red Martínez, D. L.; Podestá, C. E. *Data Mining to Find Profiles of Students*; Volume 10 – N° 30; European Scientific Journal (ESJ); pp. 23-43; ISSN N° 1857-7881; University of the Azores, Portugal. 2014c.

La Red Martínez, D. L.; Podestá, C. E. Metodología de Estudio del Rendimiento Académico Mediante la Minería de Datos; Volume III – N° 01; *Revista Científica Iberoamericana de Tecnología Educativa – Scientific Journal of EducationalTechnology*; pp. 56-73; ISSN N° 2255-1514; España. 2014d.

La Red Martínez, D. L.; Podestá, C. E. Contributions from Data Mining to Study Academic Performance of Students of a Tertiary Institute; Volume 02

– N° 9; *American Journal of Educational Research*; pp. 713-726; ISSN N° 2327-6126; USA. 2014e.

Maradona, G. & Calderón, M. I. Una aplicación del enfoque de la función de producción en educación. *Revista de Economía y Estadística*, Universidad Nacional de Córdoba, XLII. Argentina. 2004.

Marcelo García, C.; Villarín Martínez, M.; Bermejo Campos, B. Contextualización del rendimiento en bachillerato. *Revista de Educación*, 282, 267-283. España. 1987.

Matignon, R. *Data Mining Using SAS Enterprise Miner*. U.S.A.: Wiley. 2009.

McMahon, W. W. *Education and Development*. Oxford University Press. 2002.

Mitchell, T. *Machine Learning*. McGraw Hill. 1997.

Poe, V. *Building a Data Warehouse for Decision Support*. New Jersey: Prentice Hall. 1996.

Porto, A. & Di Gresia, L. *Características y rendimiento de estudiantes universitarios*. El caso de la Facultad de Ciencias Económicas de la Universidad Nacional de La Plata. Documentos de Trabajo UNLP, 24. 2000.

Quinlan, J. R. Induction of decision trees. *Machine Learning*; Vol. 1(1); pp. 81-106; Kluwer Academic Publishers, Boston, USA. 1986.

Reyes, S. L. El Bajo Rendimiento Académico de los Estudiantes Universitarios. Una Aproximación a sus Causas. *Revista Theorethikos*. Año VI, N° 18, Enero-Junio. El Salvador. 2004.

SAS Institute. Disponible en: http:// www. sas.com/technologies/analytics/datamining/ miner/semma.html: Fecha de Consulta: 20/06/2009.

Simon, A. *Data Warehouse, Data Mining and OLAP*. John Wiley & Sons. USA. 1997.

Spositto, O., Etcheverry, M., Ryckeboer, H., &Bossero, J. *Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil*. En Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISCI (Vol. 29, pp. 06-2). 2010.

Vivo Molina, J. M.; Franco Nicolás, M.; Sánchez de la Vega, M. del M. Estudio del rendimiento académico universitario basado en curvas ROC. *Revista de InvestigaciónEducativa*, RIE, Vol. 22, Nº 2, págs. 327-340. España. 2004.

White, C. J. *IBM Enterprise Analytics for the Intelligent e-Business*. IBM Press. USA. 2001.

Widom, J. *Research Problems in data warehousing*. Conf. Information and Knowledge Management, Baltimore. USA. 1995.

Wilson, R. L.; Hardgrave, B. C. Predicting graduate student success in an MBA program: Regression versus classification. *Educational and Psychological Measurement*, 55, 186-195. USA. 1995.