DATA MINING TO FIND PROFILES OF STUDENTS

David L.la Red Martinez, Doctor
Universidad Nacional del Nordeste, Departamento de Informática
Universidad Tecnológica Nacional, Departamento de ISI
Carlos E. Podesta Gomez, Ingeniero
Instituto Superior Curuzú Cuatiá, Departamento de Informática

Abstract

Education-oriented data mining allows to predict determined type of factor or characteristic of a case, phenomenon or situation. In this article the mining models used are described and the main results are discussed. Mining models of clustering, classification and association are considered especially. In all cases seeks to determine patterns of academic success and failure for students, thus predicting the likelihood of dropping them or having poor academic performance, with the advantage of being able to do it early, allowing addressing action to reverse this situation. This work was done in 2013 with information on the years 2009 to 2013, students of the subject Operating Systems tertiary career Superior Technical Analyst (TSAP) Higher Institute of Curuzú Cuatiá (ISCC), Corrientes, Argentina,

Keywords: Academic performance, data mining, profiling students

Introduction

The career of Technician Programmer Analyst (TSAP) Higher Institute of Curuzú Cuatiá (ISCC) has consistently been the first in number of students, considering the totality of the ISCC: 36.71%, and produces more graduates: 51.57 % of that institution; the respective percentages for reports for the years: 2006, 2007, 2008, 2009 and 2010 by the Department of students of ISCC.

A more detailed analysis, we found relatively low graduation rates in respect of new enrollees in TSAP; these percentages vary if we consider only the terminal title (Senior Technical Analyst) or if one also considers the intermediate title (Computer Systems Operator). Regardless of the intermediate title data are: 2006: 10.25%, 2007: 11.55%, 2008: 10.75%, 2009: 11.45%, and 2010: 10.45%. Considering the intermediate title: 2006: 21.81%, 2007: 23.22%, 2008: 21%, 2009: 23%, 2010: 22%.

The relatively low rates of new graduates enrolled respect mentioned in the previous section, we might consider the "overall academic performance" of a race, are also observed in many subjects of the TSAP, considering "special achievement" or simply "academic performance" the results of student assessments completed during a course, and the final condition reached by them in the framework of Resolution N° 1551-1501 Organic Framework Regulation (RAM) for Higher Institutes (system of evaluation and promotion: Article 85 and 86): promoted, regular or free.

Operating Systems for the subject securities in recent years are as follows: Students promoted and regularized for which surrendered some part test: 2006: 16.25% 2007: 27.45% 2008: 30.55%, 2009: 28.50% 2010: 30.39%. It has also been observed that a considerable proportion of students enroll to study the subject, but then do not complete the course (55.39% in 2010).

Given the above situation is considered of great importance to conduct an investigation that would determine the variables that affect the relatively low academic performance of students in Operating Systems TSAP ISCC, belonging to the Directorate General of Higher Education (DGES), identify profiles of successful students (those who promote or regularize the subject), as well as profiles of students who do not succeed (the remaining free status). Having determined the profiles of students with poor academic performance, could face action to avoid potential academic failure. To determine the profiles of students was considered appropriate techniques of Data Warehouses (Data Warehouse: DW) and Data Mining (Data Mining: DM).

The contribution is important for the institution in which the research has been done and other similar, because investigations of this type never have been made in the area.

In this context, characterized by massive, lack of resources in the right proportions, poor academic performance, the application of ICTs would be an important complement to traditional teaching - learning, constituting a priori an effective tool to try to resolve the situation above [1], [2], [3] and [4].The work was based on the following hypothesis: the use of teaching-learning tools based on the New Technologies of Information and Communications (NICT), impacts affects the academic performance of students of Operating Systems Superior Technical Analyst ISCC, but this use of ICTs is influenced by various socio-economic and attitudinal variables.

The overall objective was: meet the variables that affect the academic performance of students in Operating Systems on the use of ICTs in the Superior Technical Analyst Higher Institute of Curuzú Cuatiá. The specific objectives were to determine how these variables affect the academic use of ICTs and academic performance of students: a) the educational level of the

parents, b) the socio-economic level, c) the possession of a PC, d) the area in which student's access to ICTs, e) the general attitude towards the study.

Antecedents and Theoretical Framework

We have included many references to work around the world. In the paper works not only related to data mining have been included as background and theoretical framework in education but also works related with the description of the problem, with the methodological aspects, with the selection of variables to consider, with classical statistical methods and the methods of data mining, etc.

Higher education institutions are beginning to use analytics for improving the services they provide and for increasing student grades and retention. The U.S. Department of Education's National Education Technology Plan, as one part of its model for 21st-century learning powered by technology, envisions ways of using data from online learning systems to improve instruction. Educational data mining and learning analytics are used to research and build models in several areas that can influence online learning systems [5].

Reference [6] shows that Florida's Education Data Warehouse (Florida, USA), plays an important role in the state by enabling policymakers, educators, and researchers to track student progress from prekindergarten through graduate school. The data warehouse has improved Florida's ability to track and assess student progress and performance over time and has significantly improved the availability of information for tracking students' academic progress. The warehouse provides a more efficient and consistent process for compiling longitudinal student data.

In [7], the idea proposed is to perform an analysis considering number of parameters for the derivation of performance prediction indicators needed for faculty performance assessment, monitoring and evaluation. The data mining methodology used for extracting useful patterns from the institutional database is able to extract certain unidentified trends in faculty performance when assessed across several parameters. The authors consider that the traditional approach uses cumulative values of all parameters taken into consideration. This necessitates using data mining concepts for the performance evaluation so that hidden trends and patterns in faculty performance can be unearthed and can be a benefactor for the management in restoring potential faculties, encouraging faculty growth, honoring and awarding faculties.

Reference [8] shows studies conducted to identify the possible parameters that contributed to the successfulness of student grade in academic especially in computer science course.In [9] six parameters were selected for the Students' Academic Performances (SAP) which include:

Grade Point Average (GPA), race, gender, hometown, family income and university entry mode.Reference [10] have a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it has a specific requirements not presented in other domains; they described the cycle of applying data mining in educational systems.

In [11], different methods and techniques of data mining were compared during the prediction of students' success, applying the data collected from the surveys conducted at the University of Tuzla, the Faculty of Economics, among first year students and the data taken during the enrollment. The success was evaluated with the passing grade at the exam. The impact of students' socio-demographic variables, achieved results from high school and from the entrance exam, and attitudes towards studying which can have an effect on success, were all investigated.In [12], they predicted a student's academic success (classified into low, medium, and high risk classes) using different data mining methods (decision trees and neural network).

According to [13], the new information society or cyber society raises a number of questions of technical, economic, sociological, cultural and political order.One question is whether education systems are able to produce the quantity and quality of graduates needed to withstand the demands of highly trained staff of this Information and Knowledge Society (IKS) in different areas, especially those related to ICTs.It is here where it appears the problem of performance or academic performance.

Reference [14] examined to which extent different motivational concepts contribute to the prediction of school achievement among adolescent students independently from intelligence.In [15], standard t-test and ANOVA were applied to investigate the effect of different factors on students' achievement.In general, empirical studies confirm the correlation between higher levels of education and positive attributes after studies [16].

The problem of finding good predictors of future performance so that academic failure is reduced in graduate programs has received special attention in the U.S. [17], having found that the classification techniques such as discriminant analysis or logistic regression are more appropriate than the multiple linear regression predicting academic success / failure.

In addition to traditional tools before you point used for the study of academic performance, there are other from the Business Intelligence (BI), such as Data Warehouses (DW) and Data Mining (DM), used for discovering hidden knowledge in large volumes of data.A DW is a collection of data-oriented issues, integrated, nonvolatile, time variant, which is used to support the process of managerial decision making [18], [19].

The process of the formation of significant models and assessment within Knowledge Discovery in Databases (KDD) is referred to as DM [20].KDD is an interdisciplinary area focusing on methodologies for extracting useful knowledge from data. Extracting knowledge from data draws on research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing to deliver advanced business intelligence and Web discovery solutions [5].

Data mining is the field of discovering of implicit and interesting patterns for large data collections [21].The DM is the discovery stage in the process of KDD, is the step consisting in the use of specific algorithms that generate a list of patterns from the pre-processed data [22].The DM is closely linked to the DW because they provide historical information with which mining algorithms obtain the information needed for decision-making [23].The DM is a set of data analysis techniques that allow to extract patterns, trends and regularities to describe and understand the data and extract patterns and trends to predict future behavior [19], [24].The DM generated models can be descriptive or predictive [25]; its techniques are different, one of the most used is the clustering (or grouping of data) [26]. The demographic cluster is an algorithm developed by IBM that automatically solves the problems of defining distance / similarity metrics, providing criteria for defining an optimal segmentation.

Educational Data Mining (EDM) develops methods and applies techniques from statistics, machine learning, and data mining to analyze data collected during teaching and learning. EDM tests learning theories and informs educational practice. Learning analytics applies techniques from information science, sociology, psychology, statistics, machine learning, and data mining to analyze data collected during education administration and services, teaching, and learning. Learning analytics creates applications that directly influence educational practice [5].

In [27], using the decision tree predicted the result of the final exam to help professors identify students who needed help, in order to improve their performance and pass the exam.In [20], the relationship between student's university entrance examination results and their success was studied using cluster analysis and k-means algorithm techniques.Reference [28] shows that Educational Data Mining (EDM) is concerned with developing methods and analyzing educational content to enable better understanding of students' performance.

Reference [8] shows studies conducted to identify the possible parameters that contributed to the successfulness of student grade in academic especially in computer science course.In [9] six parameters were selected for the Students' Academic Performances (SAP) which include: Grade Point Average (GPA), race, gender, hometown, family income and

university entry mode.Reference [29] applies the kernel method as data mining techniques to analyze the relationships between students's behavioral and their success and to develop the model of student performance predictors.Reference [30] shows a case study that used data mining to identify behavior of failing students to warn students at risk before final exam.

Reference [31] shows how using data mining techniques can help discovering pedagogically relevant knowledge contained in databases obtained from Web-based educational systems or Online Learning Systems.Reference [32] describe different types of data mining techniques, both classical and emergent, used for educational tasks by different stakeholders.Reference [33] provides a technical overview of the current state of knowledge in educational data mining. It helps education experts understand what types of questions data mining can address and helps data miners understand what types of questions are important in education decision making.In [34], they define learning analytics, how it has been used in educational institutions, what learning analytics tools are available, and how faculty can make use of data in their courses to monitor and predict student performance.Reference [35] presents the capabilities of data mining in the context of higher educational system by i) proposing an analytical guideline for higher education institutions to enhance their current decision processes, and ii) applying data mining techniques to discover new explicit knowledge which could be useful for the decision making processes.

Reference [36] presents a data-based user modeling framework that uses both unsupervised and supervised classification to build student models for exploratory learning environments.Reference [37] proposes a methodology based on data mining and self-evaluation in order to predict whether an instructor will or will not accept the students' proposed marks in a course.Reference [38] presents an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system.In [39], the variables considered are: status of the student, educational level of parents, secondary education, socio-economic level, and others. DW and DM techniques were used to search profiles of students and determine success or failure academic potential situations. Classifications through techniques of clustering according to different criteria have become. Some criteria were the following: mining of classification according to academic program, according to final status of the student, according to importance given to the study, mining of demographic clustering and Kohonen clustering according to final status of the student. The experience was developed at the Northeastern National University, Argentine.

Methodology Used

We used a quantitative logic approach, working with measurement of variables, hypothesis production and use of intelligent data mining, for the purpose of extracting hidden knowledge in the data.In the research carried out there have been several hypotheses, which were then verified with data mining methods; also used methods of data mining that do not requires prior hypothesis, except for the decision of which variables to include in each mining process (e.g.: supervised classification and unsupervised classification).

We sought to fulfill the above objectives previously working with the hypothesis already mentioned in the Introduction.The universe consisted of the students able to study the subject Operating Systems TSAP ISCC career.The unit of analysis consisted of each student in a position to take the subject Operating Systems. The selected cases were students able to attend this course (about 200 students).

The data generated by this research will be added to data of other similar research conducted at the Northeast National University (Argentine), at the National Technological University (Argentine), the Catholic University of Santiago del Estero (Argentine) and the National University of the East (Paraguay); the above investigations will continue to incorporating data for several years more. It is considered that this justifies using a structure of DW, which will continue to grow in the future and will enable more research work.

Quantitative data obtained (integrated into a DW) were analyzed with DM tools, in order to investigate relationships between variables with non-traditional methods.Has been used the IBM Data Warehouse Edition (DWE) V.9.5, including the Intelligent Miner (IM).

Methodology of Definition of Used DW

It is important to remember that a DW cannot be acquired, must be built following certain methodology.The technique used in the creation of the DW depends on to whom main point focuses its development, can be to the management of data, goals or users [40]. The proposed models are: “Data-Driven”, “Goal-Driven” and “User-Driven”. The following describes in general terms what constitutes each.

Data-Driven: This model considers that in a DW handled data, in contrast to the classical systems, that are managed requirements, which are the last aspect to be considered in the decision-making process, considering the needs of users in second term. The data model consists of few dimensions and groups of facts. The dimension represents the basic structure of the design. The facts are based on time and have low level of granularity. Goal Driven: This model considers that the development process revolves

around the objectives and targets set out in principle. Unlike the previous model, it contains more dimensions and few facts, which are based on time and have a low level of granularity.User Driven: It is considered that the main factor to take into account is the needs of users, as are those who ultimately use the system. The model consists of a few facts, which have a moderate level of granularity.

The two main methodologies for the development of a DW are “Big Bang” and “Rapid Warehousing”.

Big Bang: This methodology tries to solve all known problems creating large DW, before releasing for evaluation and testing [41].Rapid Warehousing: This is also known as evolutionary or incremental methodology and considers the construction and implementation of a DW is an evolutionary process, which is to quickly create a portion of a DW with the integration of data marts [42].

In this work we have followed “User Driven” model and “Big Bang” methodology.

Structure Description of Used DW

Variables in the fact table: University ID card, National identity document, Career, Sex (gender), Age, Marital status, Date of birth, Country, Province, City, Date of the survey, Blood group, First partial, First recuperatorio, Second partial, Second recuperatorio, Extraordinary, Final situation of the student after completed, School year.

Variables that make up the dimension importance awarded to the study: University ID card, National identity document, Importance awarded to the study.Variables that constitute the dimension of student hometown: University ID card, National identity document,Province of residence, City of residence.Variables that constitute the dimension use of ICTs in consideration of the student: University ID card, National identity document,Use of ICTs in consideration of the student.

Variables that constitute the dimension student's secondary studies: University ID card, National identity document, Name of the College, Dependency of the College, Province to which the school belongs, City College to which it belongs, Title awarded by the College, Graduation date of the student.Variables that make up the dimension student's current residence: University ID card, National identity document, Type of student residence, Address of current residence, Province of residence, City of residence.

Variables that make up the dimension hours dedicated to the study on the assessment of the student: University ID card, National identity document, Hours dedicated to the study on the assessment of the student.Variables that make up the dimension of employment situation of the

mother of student: University ID card, National identity document, Educational level of the mother, Labor situation of the mother, Weekly hours worked, Work activity branch of the mother, Occupational category of the mother.

Variables that make up the student employment status dimension: University ID card, National identity document, Occupational category, Category of economic activity, Weekly hours worked, Working relationship with your chosen career, Social security system, Occupational category, Employment situation.Variables that make up the dimension of employment situation of the parent of student: University ID card, National identity document, Educational level of the parent, Labor situation of the parent, Weekly hours worked, Work activity branch of the parent, Occupational category of the parent.

The study was carried out on data obtained through surveys of students, considering also the results of the different instances of evaluation envisaged during the course of Operating System.

Used DM Methodology

Currently, there are several DM methodologies; the most used are the SEMMA and the CRISP-DM.

SEMMA methodology was developed by SAS Institute to discover unknown business patterns. The name refers to the five basic stages of the process [43].The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology is organized in six stages, each of which in turn is divided into several tasks [44]. The main steps are including: domain understanding, data understanding, data preparation, modeling, evaluation and deployment.

In this work the CRISP-DM methodology was used.

Data mining functions and algorithms

The IBM DWE used provides mining functions to solve various problems [45]:

Associations: The Associations mining function finds items in your data that frequently occur together in the same transactions.Classification: With the Classification algorithms, you can create, validate, or test classification models. For example, you can analyze why a certain classification was made, or you can predict a classification for new data.Clustering: The Clustering mining function searches the input data for characteristics that frequently occur in common. It groups the input data into clusters. The members of each cluster have similar properties.Regression: Regression is similar to classification except for the type of the predicted value. Classification predicts a class label, regression predicts a numeric

value. Regression also can determine the input fields that are most relevant to predict the target field values.Sequence Rules: The Sequence Rules mining function finds typical sequences of events in your data.Time Series: The Time Series mining function enables forecasting of time series values.

Main Results

They will then show the main results (only the main results) obtained with different used DM techniques: clustering (segmentation), association generators (association rules) and decision trees (classification prediction).

Results with Clustering

The Clustering mining function searches the input data for characteristics that frequently occur in common. It groups the input data into clusters. The members of each cluster have similar properties [45].IM (in DWE) provides the Clustering mining function. The Clustering mining function includes the following algorithms: a) Demographic clustering (distribution-based); b) Kohonen feature maps (center-based); c) Enhanced BIRCH (distribution-based).

These Clustering algorithms group data records on the basis of how similar the data records are.A data record might, for example, consist of information about a customer. The Clustering algorithm groups similar customers together. At the same time it maximizes the differences between the different customer groups that are formed in this way.

The algorithms of the Clustering mining function provide common parameters and algorithm-specific parameters:

Generic clustering: The Clustering mining function searches the input data for characteristics that frequently occur in common. It groups the input data into clusters. The members of each cluster have similar properties. There are no preconceived notions of what patterns exist within the data. Clustering is a discovery process.

Demographic clustering: Demographic clustering is distribution-based. It provides fast and natural clustering of very large databases. Clusters are characterized by the value distributions of their members. It automatically determines the number of clusters to be generated. Typically, demographic data contains many categorical variables. The mining function works well with data sets that consist of this type of variables. You can also use numerical variables. The Demographic Clustering algorithm treats numerical variables by assigning similarities according to the numeric difference of the values.

Clustering with Kohonen Feature Maps: The Kohonen Feature Map algorithm is center-based. It normalizes input variables to the value range [0;1]. Categorical input variables are encoded by using nominal

encoding. Therefore, categorical input variables with lots of different values can slow down the mining run considerably. The Kohonen Feature Map tries to put the cluster centers in places that minimize the overall distance between records and their cluster centers. Euclidean distance is used to determine the distance between a record and the centers.

Enhanced BIRCH Clustering: The enhanced BIRCH algorithm is distribution-based. BIRCH means balanced iterative reducing and clustering using hierarchies. It minimizes the overall distance between records and their clusters. To determine the distance between a record and a cluster, the log-likelihood distance is used by default. If all active fields are numeric, you can select Euclidean distance.

The mining parameter setup for generating clusters is the following: maximum number of clusters: 20, algorithm: demographic, similarity threshold: 0.6.

Influence of sex (gender) in the use of ICTs by students and their academic performance

The largest group contains 16% of the total population. The smallest group contains the 4.17% of the total population. The overall quality of the model, measure of homogeneity of the clusters is 0.749, which indicates that, on average, tuples in a same cluster have 74.9% of the same value in the active attributes.

One of the clusters, which represents 16% of the total population, has predominantly students with the following characteristics: male, final status of 6, note which approved the course, single marital status, their hometown is Curuzú Cuatiá (84%), its origin is Corrientes province (96%) for the male population (predominant in that cluster) ICTs facilitate the teaching of the subject by 58%, while 27% displayed the importance of same as applied to the professional field.

Another cluster, with 11.46% of the population is completely female students, 21% have achieved a final position of 7, 8 and 9; this group can be seen that although no regularity of 6 common note in the male population, women have obtained higher grades, marital status is single in all cases, the city of origin is Curuzú Cuatiá for 86 %, the home province of Corrientes is 100%, 27% believe that ICTs are a reality, while 64% think that the importance of these lies in its application to the professional field.

Influence of educational level of parents in the use of ICTs by students

The largest group contains 31% of the total population. The smallest group containing 3.84% of the total population. A cluster corresponding to 31% of the total population, indicates that 23% of parents of students have completed elementary school, while 14% have completed secondary school;

regarding the degree of use of ICTs by students, 56% define the use thereof as facilitators of the learning process, 28% consider that they will be essential in professional practice, which allows to assert a priori a high degree of acceptance in relation to the use of these technologies (84%).

Another cluster, corresponding to 13% of the total population, shows that the level of education of the parents is 100% complete secondary school; regarding the degree of use of ICTs by students, can be seen a strong response in relation to the importance that the student assigned to the use of these tools (98%), linking them fundamentally to its academic formation process.

Another group, corresponding to 11.39% of the total population, shows that 95% of parents of students have completed elementary school, while 3% had completed university studies and 2% non-university higher education complete; 59% of students believe that ICTs facilitate the learning process, while 26% say it will be essential for professional practice.

Whereas previously indicated, can be extracted as a comment that as improves the level of education of parents, this undoubtedly influences the opinion that the student has regarding the use of ICTs.

Influence of type of training received in high (secondary) school in the use of ICTs by students

The largest group contains 38% of the total population. The smallest group containing 3.36% of the total population.In the cluster corresponding to 38% of the total population, it appears that the predominant qualification profile is related to the administrative management of organizations (35%); respect of the opinion that the student has in relation to the use of ICTs, we can see that 100% define these tools as facilitators of the teaching; a priori we can say that it does influence the type of degree obtained by the student at the end of high school, as the student whose degree profile is oriented to the administrative management of companies, has a better opinion regarding the use of these technologies.

Influence of the fact that the students work in addition to studying, in the use of the ICTs

The largest group contains 18.61% of the total population. The smallest group containing 5.12% of the total population.The cluster corresponding to 18.61% of the total population, with regard to the employment situation of the student, shows that 100% of this population does not work; with respect to the use of the ICTs, it shows that 100% of the population agree that facilitate the teaching process.

In another cluster, corresponding to 8.54% of the total population, compared to the number of hours worked by the student in the week, it can

be seen that 100% of the population works in tasks that consuming an average of more than to 5 clock hours per day; referred to the situation of the use of ICTs by students, it can be said that although the importance attached to the use of these tools in terms of their use does not clearly indicate that there is an influence as students working and which does not, however it can be stated that there is a more concrete opinion on the student who works and studies, based on the fact that students who work and study also expresses interest in using these tools in the professional field.

Influence of the general attitude toward study in the use of ICTs by students

The largest group contains 19.72% of the total population. The lower group contains 5.45% of the total population.The cluster corresponding to 19.72% of the total population, has predominantly students who spend more than 10 and up to 20 hours even to the study, are, also with regard to the use of the ICTs, saying which facilitate the process of teaching and learning and the importance assigned to the study is more than fun; with respect to the number of hours devoted to the study by the student, it can be seen that 100% of the population expressed a commitment between 10 and 20 hours; with respect to the importance that the student assigned to study, it can be seen that 100% of this population appears to give one importance greater than the fun; with respect to the use of the ICTs from the learner, it can be observed that 100% of the population reported that they facilitate the learning process.

Another cluster, corresponding to 10.14% of the total population, with respect to the amount of hours a week dedicated to the study by the student, it can be seen that 100% of the population expressed a commitment between 10 and 20 hours; with respect to the importance that the student assigned to study, it can be seen 98% of this population appears to give one importance greater than the fun, while 1% more than the family; with respect to the use of the ICTs from the learner, it can be seen that 100% of the population reported that they will be essential to the professional practice.

Another grouping, corresponding to 5.45% of the total population, with respect to the amount of hours a week dedicated to the study by the student, it can be seen that 88% of the population expresses more than 20 hours, while a 2% up to 10 hours inclusive; with respect to the importance that the student assigned to the study, 77% of the population believes that it is more important than the fun, on the other hand 1% more than the family and 22% more than work; with respect to the use of the ICTs from the learner, it is observed that 70% of the population believes that they facilitate the teaching process, moreover 15% believes that they will be essential for the professional practice and 11% believes that they are a reality today.

It can be seen that the degree of commitment and importance attached by students to their studies has a direct relationship with the same attitude about the use of ICTs.

Results with Association Generators

Mining association aims to find the elements that are consistently associated with others in a meaningful way. Discovered relationships are expressed as association rules. The role of association mining and associations is also assigns probabilities. The first part of an association rule is called the body of the rule and the second part the head of the rule.

Association rules have the following attributes: a) confidence: confidence value represents the validity of the rule (a rule has 70% confidence if at 70% of the cases in which the body of the rule is also present in a group, the head of the rule is present in the group); b) support: the support value is expressed as a percentage of the total number of records or transactions; c) elevation: elevation value indicates to what extent the value of confidence is higher than expected; It is defined as the ratio of the value of confidence and the value of support from the head of the rule; the value of the rule head support can be considered as the value expected for confidence and indicates the relative frequency of the head of the rule in the whole transactions.

You can make the associations or the sequences that are found among items more meaningful if you group the items in categories. You can group these categories again into subcategories. The result is a hierarchy of categories with the items on the lowest level. This is called a taxonomy [45].

112 rules are obtained, some of which are listed below; the mining parameter setup for generating associations is the following: maximum rule length: 2, maximum number of rules: 10, minimum confidence: 25%, minimum support: 2%, number of bins: 5.

If the student is male gender, single marital status involves 91% of cases.If the student is female gender, single marital status involves 85% of cases. If the final status of the student is 6, which occurs in 31%, implies a single marital status in 86% of cases.If the student is female gender implies that opine that ICTs facilitate the teaching process in 56% of cases.If the student believes that the use of ICTs is essential for professional practice, which occurs in 25%, means that your marital status is single in 88% of cases.If sex (gender) of student is male, implies that its final status will be 6 in 37.5% of cases.If sex (gender) of student is female, implies that its final status will be 6 in 35.44% of cases.

If the student believes that the use of ICTs is essential for professional practice, which occurs in 14%, implies that the student is female gender in 49% of cases.If they student opinion is that the use of ICTs

facilitates the teaching and the hours are devoted to the study up to 10 hours inclusive, what happens in 12.54%, implies that gender student is male in 50.31% of cases.If your marital status is single and the student believes that the use of ICTs is essential for professional performance, which occurs in 13%, implies that the gender of the student will be male in 52% of cases.If they student opinion is that the use of ICTs facilitates the teaching and the hours devoted to the study are over 10 and up to 20 inclusive, which occurs in a 13.43%, implies that the student is female gender at 49.68% of the cases.

If they student opinion is that the use of ICTs facilitates the teaching and the hours devoted to the study are over 10 and up to 20 inclusive, which occurs in a 13.60%, implies that the student is male gender at 50.31% of the cases.If the student is female gender and the final status is 6, what happens in a 14.46%, implies that the marital status of the student will be single in 82% of cases.If the final status of the student is 6 and the hours are devoted to the study and 10 to 20 inclusive, which occurs in 15%, implies that the marital status of the student will be single in 86% of cases.If the final state is 6 and is male, which occurs in 17%, implies that the marital status of the student will be single in 90% of cases.If female and spends up to 10 hours to study inclusive, which occurs in 19%, implies that the marital status of the student will be single in 85% of cases.If it is single and hours devoted to the study are to 10 inclusive, which occurs in 22% of cases, implies that the opinion on the use of ICTs is to facilitate the learning process in 56% of cases.

If students have the opinion that the use of ICTs facilitates the teaching and the hours devoted to the study are more than 10 to 20 inclusive, which occurs in 24%, implies that the marital status of the student will be single at 88 % of cases.If the use of ICTs facilitates the process of teaching and student gender is male, which occurs in a 25.63%, implies that the marital status of the student will be singles in 91.25% of cases.

Results with Decision Trees

IM supports a decision tree implementation of classification [45]. 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.

Decision Tree Classification generates the output as a binary tree-like structure, which gives fairly easy interpretation to the marketing people and easy identification of significant variables for the churn management.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 results have been summarized and grouped according to final rating (class); has been considered high-performance academic to the final ratings between 7 and 10, academic performance medium to the final score of 6 and low academic performance to the final score from 0 to 5.

The results summarized from the profile of the students considered high academic performance, corresponding to 25.78% of the population, are the following: a) most lives with the family group, b) generally do not work, c) a minority group works up to 20 hours a week, d) in the majority of cases the work relationship with the chosen career is partial, e) the degree of primary and secondary schooling of parents is relatively low, registering cases of tertiary or university education, f) mostly the parents occupancy rate is relatively high, g) in most cases the goal of students is to study to learn to learn or to fully learn the subject, h) the majority considers the use of the ICTs associated with the process of teaching and learning and as essential for professional practice, i) the majority are single, registering a good percentage of married, j) the majority correspond to the male gender, k) a minority group gives the study more priority than job.

The results summarized from the profile of the students considered average academic performance, corresponding to 36.44% of the population, are the following: a) the majority living with the family group, b) usually do not work, c) a minority group works up to 20 hours a week, d) in most cases the relationship between work and career choice is partial, e) the degree of primary and secondary schooling of parents is relatively low, not registering cases of tertiary or university schooling, f) mostly the parents occupancy rate is relatively low, g) in most cases the goal of students is to study to pass the subject, h) the majority considers the use of the ICTs associated with the teaching-learning process, i) most are single, recorded a good percentage of married, j) the majority correspond to the male gender.

The results summarized from the profile of the students considered low academic performance, corresponding to 37.73% of the population, are the following: a) the most lives with the family group, registering a significant minority who lives independently, focusing especially on the class corresponding to the rating of 2, b) generally do not work, but a significant group does, in this category is the largest number of students who work, c) a minority group works up to 20 hours a week and another smaller group more than 36 hours per week, d) in the majority of cases the work relationship with the chosen career is partial or non-existent relationship, e) the degree of primary and secondary schooling of parents is relatively low, registering cases of tertiary or university education, f) mostly the parents occupancy rate is relatively high, registering an important minority group with a low occupancy rate, g) in most cases the goal of students is to study to pass the subject and a minority group makes learning to learn or learn

integrally the subject, h) most consider the use of ICTs associated with the teaching-learning process and a minority group as essential for professional practice, i) most are single, j) the majority correspond to the female gender.

There are interesting correlations, for example, showing the incidence of the first quarter note in the final situation of the student, so the impact of the type of residence regarding the final status of the student, the education level of parents in relation to the hours devoted to the study and final status of the student, the incidence of the use of ICTs in relation to the final situation of the student.

Conclusion

It is critical to know from the beginning of the academic activities which students are candidates to poor academic performance and what factors influence it, to address early action to reverse this situation. In this investigation we have only covered a few methods of extracting knowledge through DM. However, there are many more possibilities offered by this and other tools.

With three selected DM techniques have been obtained very good results, fulfilling the objectives and verifying the working hypothesis. Have you been evidenced characteristics of representative profiles of students with low, medium and high academic performance.The model sorting through decision tree outperformed as the patterns obtained with the method of generating clusters.

DM techniques have enabled to build predictive models, of association, of segmentation, based on historical data stored in different sources; the quality of the obtained models considered adequate. It has been possible to determine the success and failure academic profiles of Operating Systems students of TSAPS of ISCC, which has allowed to define lines of action aimed to give greater support to the students detected with risk of academic failure.

Future Lines of Research

Throughout the development of this work have appeared several lines to be addressed in future research.Some of them include the following: a) integrate the different mining flows in control flows that allow to automate the processes described in this paper; b) design hyper cubes of data incorporating new socio-economic variables; c) implement academic monitoring mechanisms of the actions that are carried out on the basis of the information provided by the mining process, for the purpose of making adjustments that are considered relevant for the implementation of actions referred to above; d) applying the model developed in this work to other subjects in the degree of ISCC TSAP especially the first year in which the

largest percentage of academic failure are recorded; e) compare this model based on data mining with the one proposed in [46], based on genetic algorithms.

References:

Jézégou, A. Presence in E-learning: Theoretical Model and Perspectives for Research. *International Journal of E-Learning & Distance Education,* Vol. 26, No 2. Canada. 2012.

Stoerger, S. Creating a Virtual World Mindset: A Guide for First Time Second Life Teachers. *International Journal of E-Learning & Distance Education,* Vol. 24, No 3. Canada. 2010.

Wallace, L. & Young, J. Implementing Blended Learning: Policy Implications for Universities, *Online Journal of Distance Learning Administration*, Volume XIII, Number IV, winter 2010 University of west Georgia, Distance Education Center. 2010.

IEEE. (2012). Learning Technology *Standards Committe. Retrived* Jan 6, 2012, from http://www.ieeeltsc.org:8080/Plone. 2012.

Duncan, A.; Cator, K. *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics*. U.S. Department of Education. USA. 2012.

OPPAGA: Office of Program Policy Analysis & Government Accountability. *Education Data Warehouse Serves Important Function; Project Planning and Management Need Strengthening,*Report No. 09-31, Florida Legislature. USA. 2009.

Singh, C.; Gopal, A. Performance Analysis of Faculty using Data Mining Techniques. *International Journal of Computer Science and Application*. USA. 2010.

Mohsin, M. F. M.; Norwawi, N. M.; Hibadullah, C. F.; Wahab, M. H. A. *Mining the student programming performance using rough set*. International Conference on Intelligent Systems and Knowledge Engineering (ISKE). China. 2010.

Aziz, A. A.; Ismail, N. H.; Ahmad, F. Mining Students' Academic Performance. *Journal of Theoretical and Applied Information Technology.* Vol. 53 No. 3. Pakistan. 2013.

Romero, C.; Ventura, S. *Educational Data Mining: A Survey from 1995 to 2005*, Expert Systems with Applications (33), pp. 135-146. Elsevier. 2007.

Osmanbegović, E.; Suljić, M. Data Mining Approach for Predicting Student Performance. *Journal of Economics and Business*, Vol. X, Issue 1, Elsevier. 2012.

Superby, J.F.; Vandamme, J.P.; Meskens, N. *Determination of Factors Influencing the Achievement of the Firstyear University Students using Data Mining Methods*. Proceedings of the 8th International Conference on

Intelligent Tutoring Systems, Educational Data Mining Workshop, (ITS`06), pp. 37-44. Taiwan. 2006.
Negroponte, N. *Being digital*. 1st ed. Knopf. USA. 1995.
Steinmayr, R.; Spinath, B. The importance of motivation as a predictor of school achievement. *Learning and Individual Differences, Journal of Psychology and Education,* 19, 80–90, Elsevier. USA. 2009.
Farooq, M.S.; Chaudhry, A.H.; Shafiq, M.; Berhanu, G. Factors Affecting Students' Quality Of Academic Performance: A Case Of Secondary School Level. *Journal of Quality and Technology Management.* Volume VII, Issue II, Page 01 - 14. 2011.
McMahon, W. W. *Education and Development*. Oxford University Press. 2002.
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.
Inmon, W. H. *Data Warehouse Performance*. John Wiley & Sons. USA. 1992.
Simon, A. *Data Warehouse, Data Mining and OLAP*. John Wiley & Sons. USA. 1997.
Erdogan, S.Z.; Timor, M. A Data Mining Application in a Student Database. *Journal of Aeronautics and Space Technologies (AST),* Vol.2, no.2, pp.53-57. Springer. 2005.
Kloesgen, W. *Handbook of Knowledge Discovery and Data Mining*. Oxford University Press, Oxford, UK. 2002.
Fayyad, U.M.; Grinstein, G. & Wierse, A. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann. Harcourt Intl. 2001.
IBM Software Group. *Enterprise Data Warehousing whit DB2: The 10 Terabyte TPC-H Benchmark*. IBM Press. USA. 2003.
Berson, A. & Smith, S. J. *Data Warehouse, Data Mining & OLAP*. Mc Graw Hill. USA. 1997.
Agrawal, R.; Shafer, J. C. Parallel Mining of Association Rules. *IEEE Transactions on Knowledge and Data Engineering*. December 1996. 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.
Kumar, S. A.; Vijayalakshmi, M. N. *Efficiency of Decision Trees in Predicting Student's Academic Performance*, First International Conference on Computer Science, Engineering and Applications, CS and IT 02, pp. 335-343. Dubai. 2011.

Baker, R. Data Mining for Education, in *International Encyclopedia of Education*, McGaw, B.; Peterson, P.; Baker, E. Eds., 3rd ed. Oxford, Elsevier. U.K. 2010.

Sembiring, S.; Zarlis, M.; Hartama, D.; Ramliana, S.; Wani, E. *Prediction of Student Academic Performance by an Application of Data Mining Techniques.* International Conference on Management and Artificial Intelligence. IPEDR vol.6. IACSIT Press, Bali, Indonesia. 2011.

Merceron, A; Ycef, K. *Educational Data mining: A Case Study*. In proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005, IOS Press. Amsterdam, The Netherlands. 2005.

Jailia, M.; Tyagi, A. Data Mining: A Prediction for Performance Improvement in Online Learning Systems. *International Journal of Advanced Research in Computer Science and Software Engineering*. Volume 3, Issue 7. India. 2013.

Romero, C. R.; Ventura, S. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man and Cybernetics*, Part C: Applications and Reviews 40 (6): 601–618. USA. 2010.

Romero, C.; Ventura, S.; Pechenizkiy, M.; Baker, R. S. J. (eds.). *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press. USA. 2010.

Dietz-Uhler, B.; Hurn, J. E. Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective. *Journal of Interactive Online Learning*, Volume 12, Number 1. USA. 2013.

Delavari, N.; Phon-Amnuaisuk, S.; Beikzadeh, M. R. Data Mining Application in Higher Learning Institutions. *Informatics in Education*, Vol. 7, No. 1, 31–54. Institute of Mathematics and Informatics, Vilnius, Lithuania. 2008.

Amershi, S.; Conati, C. Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining*, Article 2, Vol 1, No 1. Canada. 2009.

Fuentes, J.; Romero, C.; García-Martínez, C.; Ventura, S. *Accepting or Rejecting Students' Self-grading in their Final Marks by using Data Mining*. Proceedings of the 7th International Conference on Educational Data Mining, EDM 2014. London, UK. 2014.

Minaei-Bidgoli, B.; Kashy, D. A.; Kortemeyer, G.; Punch, W. F. *Predicting Student Performance: An Application Of Data Mining Methods With The Educational Web-Based System* LON-CAPA. 33rd ASEE/IEEE Frontiers in Education Conference. Boulder, USA. 2003.

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, pp. 66-72. USA. 2012.

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

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

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

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

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.

IBM. (2013). *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_datamininoverview.html?lang=en. 2013.

Miranda Lakshmi, T.; Martin, A.; PrasannaVenkatesan; V. An Analysis of Students Performance Using Genetic Algorithm. *Journal of Computer Sciences and Applications*, Vol. 1, No. 4, 75-79. USA. 2013.