Data Mining Model to Predict Academic Performance at the Universidad Nacional de Colombia

Camilo Ernesto López Guarín

Universidad Nacional de Colombia

Facultad de Ingeniería,

Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2013

Data Mining Model to Predict Academic Performance at the Universidad Nacional de Colombia

Camilo Ernesto López Guarín

Thesis work to obtain the degree of:

Master in Systems and Computer Engineering

Advisor:

Elizabeth León Guzmán, Ph.D.

Co-advisor:

Fabio Augusto González Osorio, Ph.D.

Research Line:

Intelligent Systems

Research Group:

MIDAS

Universidad Nacional de Colombia

Facultad de Ingeniería,

Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2013

Abstract

The present research proposes an approach to Educational Data Mining at the Universidad Nacional de Colombia through the definition of models that integrate clustering and classification techniques to analyze academic data, corresponding to the students who joined the University to the programs of Agricultural and Computer and Systems Engineering between 2007-03 and 2012-01. These techniques are intended to acquire a better understanding of the attrition during the first enrollments and to assess the quality of the data for the classification task, which can be understood as the prediction of the loss of academic status due to low academic performance. Different models were built to predict the loss of academic status in different scenarios such as: in the first four enrollments regardless when; at a specific academic period using only the admission process data and then, using academic records. Experimental results show that the prediction of the loss of academic status is improved when adding academic data.

Keywords: Educational Data Mining, dropout, Education

Resumen

La presente investigación propone un acercamiento a la Minería de Datos Educativa en la Universidad Nacional de Colombia mediante la definición de modelos que integran técnicas de agrupamiento y clasificación para el análisis de datos académicos reales pertenecientes a los estudiantes de Ingeniería Agrícola e Ingeniería de Sistemas que ingresaron entre 2007-03 y 2012-01. Se pretende con estas técnicas obtener un mejor entendimiento de la desvinculación por desempeño académico en los primeros semestres de la carrera y evaluar la calidad de los datos para la tarea de clasificación, que puede entenderse como la predicción de la pérdida de calidad de estudiante. Se construyeron diferentes modelos para la predicción en diferentes escenarios, como: en las primeras cuatro matrículas sin importar cuando; en un periodo académico específico usando solo los datos de admisión y después usando los registros académicos. Resultados experimentales muestran que la predicción de la pérdida de calidad de estudiante mejora al usar información académica.

Palabras clave: Minería de Datos, deserción, educación

Contents

Introduction

In recent years, three emerging fields are using data and technology approaches to improve Education and Learning. Academic Analytics, which uses a business intelligence approach to Education in order to improve decision making and organizational efficiency; Learning Analytics, which looks to empower the actors of the learning process; and Educational Data Mining, which is a branch of Data Mining specialized on Educational needs from the learner or the organization.

In educational settings, Data Mining techniques have been applied in both, Learning and Administrative/policy-oriented issues [5, 6]. In Learning, the process can be split into learner-oriented and educator-oriented. In the first one, the focus is on supporting the student to learn more effectively by suggesting new contents; in the latter, the goal is to provide the educator a tool to empower him so he can guide the learner more effectively.

Kotsiantis et al. applied in [14] different classification methods for predicting dropout from a class based on demographic and performance data from students with Naive Bayes being the best option. Superby, Vandamme and Meskens [15] studied the phenomenon of academic failure of first-year students. They present the variables that are more correlated to academic success based on the model used by Parmentier [16], which explains that the academic result of a student is influenced by three set of factors: personal history, involvement in his own studies and the student's perceptions. Also, this work includes an application of Data Mining techniques to classify the first-year students into three categories: low, medium and high-risk students. In [17], three different datasets are used to predict dropout: Pre-university information, academic performance, and a combination of both. In general, the results were better for the third dataset, followed closely by the second. The authors implemented cost sensitive learning in order to avoid False Negatives. Kotsiantis goes further in [19] by implementing a local cost-sensitive technique to manage the imbalanced datasets; the results were better than those

presented in his previous work [14]. Bayer et al. [18] used both, student and social data from a Data Warehouse in the University to predict student dropout. Data Mining models had better results with the student and social data and the lower results came from using social data only.

The research in Colombia of Data Mining techniques applied to education is limited to studies developed in the Universidad de Nariño. They applied C4.5, a classification algorithm based on decision trees, to predict both, the academic performance of a student and the possibility of a dropout. They also developed an algorithm for discovering Association rules called EquipAsso [21]. In the Universidad Sergio Arboleda there is another example with a different approach; in this case they had a focus closer to marketing rather than Computer Science [22] and were interested in identifying the profiles of the students and dropouts of the university. They used K-Means to accomplish their goal.

The Universidad Nacional de Colombia has conducted its own studies on drop out in 2007 and 2010 for the undergraduate [24] and graduate [25] programs respectively; however, these studies don't contemplate the last Academic Reform which was implemented in the year 2008 to improve the academic environment of its students. Probably the main change is the inclusion of academic credits which provided the students with more flexibility to choose their own curriculum and facilitate their mobility to other universities, national or international. Along with this, there are other changes such as the inclusion in the admission process of Math, literacy and English tests in order to level the first year students, the possibility to cancel a course at any time during the semester or the easier connection between undergraduate and graduate studies.

For the university it is of great interest to understand how these changes have affected the academic performance of its students, and because of that the offices of Academic Affairs periodically develop follow-up studies to see this impact, for instance studies of academic failure, dropouts, admitted student characterization among others. These studies allow generating a diagnosis on specific variables and how they evolve over time but neglect the possible unnoticed interaction between them, i.e., the patterns in a given time, a behavior that has emerged due to the changes that have been incorporated. Data

Mining models are a suitable tool to encompass these emerging behaviors and extend the understanding of the impact of the academic reform.

The present research aims to answer several questions. On the one hand, to find if patterns can be found in the Student data through application of descriptive and predictive models, and if it's possible to identify which factors affect in the academic success or in the student dropout event. On the other hand, the implementation of the Academic reform and its consequences may have modified our behavior as students, varying the relations among the variables and, therefore, possibly making the models to be specific for certain periods.

Objectives

The objectives of the research are listed below:

General Objective:
To design and develop a Data Mining model to predict the loss of academic status at the Universidad Nacional de Colombia.

Specific Objectives:

- To review the literature in Educational Data Mining.
- To collect, prepare, and define a proper representation of the data to apply Data Mining techniques.
- To characterize, using descriptive Data Mining techniques, a student population from the Universidad Nacional de Colombia, Bogotá Campus.
- To formulate a Data Mining model for predicting loss of academic status.
- To systematically evaluate the model.

Methodology

Data was collected from three different sources: the Academic Information System (SIA), the Direction of Admissions, and the Bogotá campus' Division of Registry. After the proper data preprocessing two data mining models were built, the first one to characterize the students based on their demographic data collected during the admission process by using descriptive Data Mining techniques. The second model made use of the

characterization mentioned above and the academic records of previous academic periods in order to design classification models to predict the loss of academic status considering different scenarios corresponding to the moment at when the loss of academic status is predicted: at any time in the first four enrollments; at a specific enrollment using only the admissions data and then adding the academic records; and finally a comparison between the models that use all the data, from the cohorts from 2007-03 to 2012-03, with those that use only the data after the academic reform, i.e. academic periods of 2009-01 and later.

Contributions

These are the main contributions of this research

- A state of the art was written regarding the prediction of academic success using data mining techniques.
- A preprocessed dataset of real data, which consists of the identification of duplicate records, attribute selection, and data integration among others.
- A student characterization model for the admitted students to the Agricultural and Computer and Systems Engineering Programs.
- A classification model for predicting the loss of academic status due to low academic performance.

Document Structure

The rest of the document is organized as follows: Chapter 1 presents a background, with a presentation of three recent fields that study the application of data driven techniques in the Educational context, a literature review regarding the topics of predicting academic success and use of clustering methods, and finally the data mining methods used in this work; the second chapter introduces the University, its context, and the data set; then, an exploratory data analysis is presented along with the general data mining model. Chapter 3 describes the student characterization model; the classification model for predicting academic success is presented in chapter 4; finally, the conclusions and future work are presented.

1. Background

This chapter introduces the use of analytics in Education, particularly in three fields of research: Educational Data Mining, Academic Analytics, and Learning Analytics. Subsequently, the work related to prediction of academic success is presented. To finalize, the methods used in the Data Mining models are also presented.

1.1 Data-driven techniques in Education

Technology has been an enabler for education. The first thoughts of this influence might be commonly related to a way for communicating, for delivering content or interacting with students by using video and other media to support a message, or creating virtual learning environments that facilitate communication; there is also the possibility to maximize access to education with online courses. However, these are not the only possibilities, Education, as many other fields, can also be improved by the use of data and analytics to enable a better decision making.

Analytics involves the use of data and quantitative analysis in the decision making process. This is supported by the recent increase of volumes of data and computational resources, which is changing the paradigm of science, from theoretical models, to computational models and finally to a data-intensive science [1]. New tools coming from Data Mining, Machine Learning or Statistics can be applied during the process of exploratory analysis by discovering new patterns that possibly were not considered by experts, and reducing the number of traditional data collection - hypothesis testing techniques to only a few interesting patterns [2].

There is a shift in the way we are taking advantage of data, and education has not escaped from it.

1.1.1 Analytics in the educational context

The application of Data Mining and other Analytics into the educational context has increased in the last decade. Ferguson presents in [3] three drivers for this to occur: first, the volumes of data that are collected in educational institutions have greatly augmented, whether from Course or Learning Management Systems or Student Information Systems; second, the use of e-learning: although have helped collecting data it also have brought some learning issues such as possible lack of motivation and difficulties for the educators to receive direct feedback regarding the mood, level of interest, or even the understanding of the students; and finally, the political concerns: countries are getting more understanding about the importance of higher education for its development and have an interest to improve it, to offer better learning opportunities that lead to better academic results. Three communities have stood out in the application of analytics in Education: Educational Data Mining (EDM), Academic Analytics (AA), and Learning Analytics.

- Educational Data Mining

Educational Data Mining is the oldest. It started on workshops, first on the International Conference on Intelligent Tutoring in 2000 and 2004, and in the International Conference on Artificial Intelligence in Education and the National Conference on Artificial Intelligence in 2005. In 2007, four different workshops on EDM were organized and then, since 2008, the International Conference on Data Mining is held on a yearly basis, and in 2009 the first edition of the Journal of Educational Data Mining was released; these two components allows the international Educational Data Mining society to help and support the development of the field.

EDM is defined by the International Educational Data Mining Society in [4] as “an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.” Among the methods of EDM, Baker proposes in [5] a classification for EDM methods as follows: Prediction, Clustering, Relationship mining, Distillation of data for human judgment, and Discovery with models. In a closer look, these are the usual tasks of Data Mining: Classification, Clustering and Association Rules Analysis with the inclusion of exploratory tasks which precede Data Mining in the Knowledge Discovery Process, which is understandable, given that EDM is

an application of Data Mining. Romero and Ventura, on the other hand, suggest in [6] a different taxonomy based on the following educational tasks: Analysis & Visualization, Providing feedback, Recommendation, Predicting Performance, Student Modeling, Detecting Behavior, Grouping students, Social Network Analysis, developing Concept Map, Planning & Scheduling, and Constructing Courseware.

- Academic Analytics

Academic Analytics (AA) were introduced by Goldstein and Katz in [7] as an application of business intelligence practices in Academia. In their research, the authors studied how technology is used to support the decision making process, and the term emerged as a broader concept that included not only the technology, but also the application and culture around it, so the term is about "how academic enterprises use information to support decision making". Campbell and Oblinger provides a similar understanding in [8], they say that “Academic analytics marries large data sets with statistical techniques and predictive modeling to improve decision making.” In this paper, the authors also further develop the benefits of an analytical approach based on data and facts to support the decision making process in an institution of higher education, instead of a decision based purely on intuition or the accumulated experience. In particular, Data Mining is presented as an alternative to extract knowledge from the large amounts of data; it presents the potentials and concerns for the different stakeholders including: students, faculty, student affairs, Executive Officers, and IT. Among the potentials that can be found across the different stakeholders, take for instance the possibility of increasing the student success, or to support the enrollment process.

- Learning Analytics

Learning Analytics (LA,) is the most recent field. It is defined by the Society for Learning Analytics Research (SoLAR) as the “measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [9]. The focus is on the learner and the learning process, how the actors can be empowered to improve the learning outcome by using different kind of information.

One characteristic of this field is how it is presented as a common place for technical, pedagogical and social domains. This can be seen in [3], where the author presents the

research challenges for Learning Analytics. It includes tasks for improving how the information regarding the learning is handled to learners and educators such as: Visualization and Dashboards, and formative feedback, which aims to understand how people are engaged with their own learning; but also mentions technical challenges like data managing and standardization, or the use of new data sources, e.g. mobile devices, contextual data, biometric data.

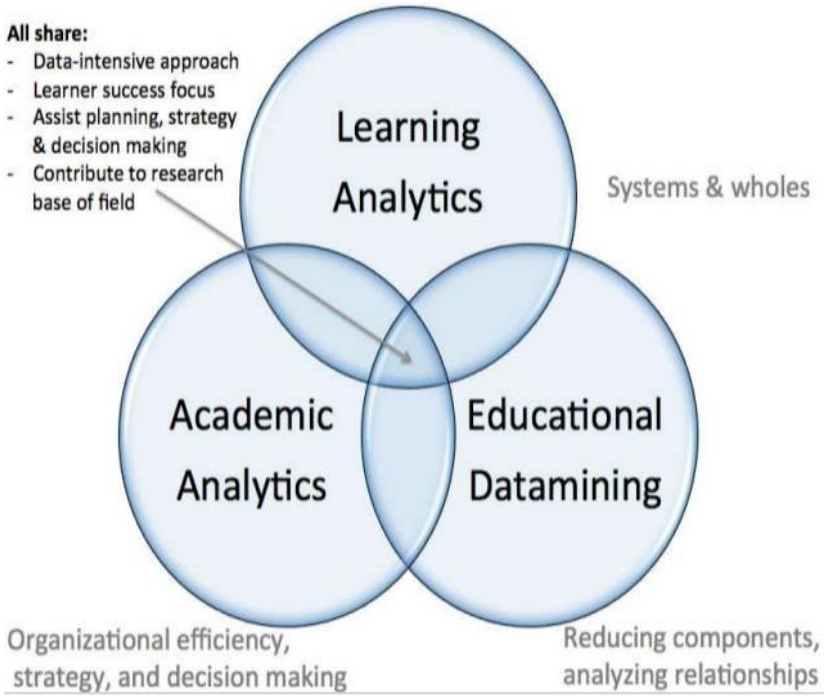
The Learning Analytics & Knowledge conference was introduced in 2011 and is held annually since then.

The three fields have similarities and some particularities, George Siemens, presented in [10] a characterization of the three fields, with all sharing: a data-intensive approach, a focus on the learner success, and an objective to support or assist planning, strategy and decision making. On the differences, Siemens described different focus for each one. AA focused on Organizational efficiency, LA in systems and wholes and EDM in reducing components and analyzing relationships. His proposal is shown in Figure 1-1. Ferguson, on the other hand presents in [3] the three factors that have driven the development of the application of analytics in education, as mentioned above, and each one of these drivers have a corresponding challenge, i.e. Big Data and a technical challenge, online learning and the opportunities to optimize it, and Political concerns and the interest in improvement. Through this evolution, Learning Analytics, Educational Data Mining and Academic Analytics have shared a lot; however, a difference can be presented as how the challenges before mentioned are taken. EDM is solving the first challenge while LA and AA are solving the second and third respectively.

As it can be seen, there are several similarities in the work of these communities with some overlaps in their research fields; this has motivated AA and LA to encourage a joint work. In 2012, during the second International Conference on Learning Analytics & Knowledge a plenary panel was held titled: “Educational Data Mining meets Learning Analytics” [11] in which representatives from both communities presented their thoughts about their discipline and the relation between the two, understanding the differences and how both can complement each other’s work. In the same conference, Siemens and Baker, the current representatives of the LAK and EDM communities respectively, presented in [12] how the two communities have evolved, along with the similarities and

distinctions between them. They both have an interest in improving education through a data-intensive approach by improving the quality of the analyses of large-educational data. The differences come from the focus of both communities which tend to differ, EDM has a greater focus on automated discovery while LAK focus on leveraging human judgment, however, several EDM and LAK research areas often overlap and researchers conduct research that could be placed on the other community's side. Based on this, the authors made a call to both communities to communicate and collaborate in order to continue growing together.

Figure 1-1: Differences and similarities, Siemens [10].



1.1.2 Predicting Academic success

Dropout prediction and the analysis of its influencing factors is a well-studied subject since the late 1960s and early 1970s [13]. Most of the works cite two researches from 1975, those from Astin and Tinto. The former presents characteristics that increase the chances of completing the studies; these are individual student's characteristics at the time when he enters college and during the course, as well as institutional characteristics. The latter introduced a model of student retention at universities in which the event of

dropping out is explained by the level of integration, both, social and academic, of an individual with the institution.

Another way to study academic success is to study the academic performance in a given course; it uses similar approaches for a different outcome, instead of studying the failure at completing the course is the study of failure at passing the course. Both of these use information about the student's past and present to predict his academic success in a class, a year, or a full program of studies.

The use of data mining techniques, on the other hand is more recent, back to 2003 when Kotsiantis et al. [14] applied different Machine Learning Techniques (C4.5, backpropagation neural network, Naive Bayes, 3NN, Maximum Likelihood Estimation and Support Vector Machines) to predict dropout in data from students in the course of Introduction to Informatics in a Distance Learning Institution. They used curriculum-based data, i.e. sex, age, marital status, occupation, computer literacy, and association between computer use and current job; and student performance data, this was represented by the activities where the student participated, namely: attendance to the first two out of three optional face-to-face meetings with a tutor and the results of the first two out of four written assignments, but only three of them were mandatory.

The algorithms were trained using older data and tested in five different training sets. The first one used only the curriculum-based data; the rest added incrementally the four features of the student performance data. The accuracy results were improved when new information was added, i.e. the academic performance data. C4.5 and Naive Bayes performed better when only the demographic attributes were used (63%). Naive Bayes and Artificial Neural Networks had the best accuracy results when the full data was included (83%).

This was a pattern for early approaches, where a lot of comparisons were made among different Data Mining techniques. Superby et al. studied in [15] the phenomenon of Academic failure of first-year students from three Belgian universities. The data comes from surveys filled out by 533 students in November 2003, at the beginning of the academic year. They present the variables that are more correlated to academic success based on the model used by Parmentier [16], which explains that the academic result of a

student is influenced by three set of factors: personal history, involvement in his own studies and the student's perceptions. The variables of Personal history had the highest correlation coefficients, followed by those regarding the involvement in the studies.

In a second part, there is an application of Data Mining techniques to classify the first-year students into three categories: low, medium and high-risk students. The results, according to the researchers were not remarkable, varying between 51%-57% of accuracy. The algorithms used were: decision trees, random forests, neural networks, and linear discriminant analysis.

Dekker et al. compares in [17] Decision trees, a Bayesian classifier, a logistic model, a rule-based learner, and the Random Forest. They analyzed three different datasets are used to predict dropout in first-year Electrical Engineering students: Pre-university information, which is mainly the previous academic performance; the academic performance, i.e. the number of attempts of every course and the higher grade; and a combination of both. The results were very similar for those datasets including the grades data, which implies that the pre-university data does not add much independent information. Decision trees provide with good results between 75 and 90% of accuracy. It was necessary to implement Cost-sensitive learning in order to avoid False Negatives.

Recent researches take into consideration new sources of data, going beyond the surveys, and national test scores, and start trying with other data.

In addition to student data (e.g. year of birth, admission year, capacity-to-study test), and semester related data (e.g. courses, credit management and grades); Bayer, in [18] created new attributes by using social network analysis, the social data is represented by a sociogram which shows the engagement in the school community with the ties being direct relations like friendship, email conversation, publication co-authoring; or indirect ones like marking a post in the forum as favorite or uploading a file into someone's repository. The features in the Social Network analysis are related to the network structure, i.e. degree, total, in and out, and to the direct neighbors' data (GPA, credit management).

Different types of machine learning algorithms were chosen. They employed a decision tree learner, a lazy learner, a rule learner, a support vector machine, and a Naive Bayes classifier. Data Mining models had better results when the student and social data was used, and the lower results came from using social data only. On the other hand, different approaches were also applied: feature selection increased the accuracy in the different techniques, except for Naive Bayes; cost sensitive methods, on the other hand lowered the results; and finally, only historic data was considered, that is, a model was learned by using exclusively the prior data, e.g. only the n first semesters. In the latter, the results of accuracy improved when more data, i.e. more semesters, were considered, however, the True Positive rate fluctuated in time, having the highest values in data from the second semester.

The more relevant features according to the paper are: the relation between gained credits and credits to gain, the GPA and weighted GPA, capacity-to-study (Learning potential.) There are two things to analyze here: first, that none of these are from the social features, and second that, depending on University regulation, these are probably the reason for a student to be dismissed.

Another new approach was the consideration of different data management techniques to overcome the special characteristics of the data, e.g. handling imbalanced data sets. Kotsiantis [19] revisited the study of [14] but handling the problem of an imbalanced dataset by implementing a local cost-sensitive technique; six different algorithms for managing imbalanced datasets were applied to the Naive Bayes model, given the results from his previous work, where it had the best results. The results were better than those presented previously.

Middle school students' data is analyzed by applying Data Mining techniques to predict school failure in [20]. There are three main sources of information: A survey conducted to the students to gather personal and Family information; a survey from CENEVAL (National Center for Evaluation), which provided socioeconomic data; and the scores for the course in several subjects. Datasets are integrated into one dataset comprised of 670 records (610: PASS, 60: Fail) and 77 attributes, which was analyzed through five rule-based learning and five decision tree algorithms, and used 10-fold cross validation to

evaluate performance, measuring: Accuracy, True Positive (TP) and True Negative (TN) rates and the Geometric Mean (GM), which is specially used in imbalanced datasets.

An initial mining was performed, which led to high accuracy results (between 93.1 and 97.6%); however it is important to remind the imbalance in the data classes. The TN results vary between 25 and 78.3% and the GM between 49.9 and 87.5%. Considering the fact that not all the attributes were used, the authors applied ten different Feature Selection algorithms, ranked the most popular and used those fifteen in a new experimentation. The results were improved in measures such as TN (41.7 - 81.7%) and GM (64.2 - 89%) but not so much in Accuracy (93.1 - 97.3%) and TP. To deal with the imbalanced dataset issue, two approaches were considered: a supervised data filter that adds more synthetic records of the minority class (SMOTE -Synthetic Minority Over-sampling Technique), and a cost-sensitive function. The second approach proved to be more effective reaching GM values between 74 and 94.6% much better than those from the balanced data approach (59 - 92.1%).

Regarding a Data Mining approach, there are only a few of examples in Colombia; at the Universidad de Nariño and Universidad Sergio Arboleda, Bogotá Campus. In the first one, Timarán [21] applied C4.5, a classification algorithm based on decision trees to predict the academic performance of a student and the possibility of a dropout, and association rules discovery. Pinzón [22], on the other hand, has a different interest, with a focus closer to marketing rather than Computer Science. They were interested in identify the profiles of the students and dropouts of the university. To do this, they applied K-Means to both sets separately. However, Dropout in Higher Education has been largely studied by the Ministry of Education and the Universidad de Los Andes [23]. Also, the Academic Vice-Presidency and the National Welfare Direction conducted studies in 2007 and 2010 for the undergraduate [24] and graduate [25] programs respectively.

In the study of 2007 regarding the undergraduate programs, the authors defined four profiles based on the academic and economic vulnerability. For the academic vulnerability, those students with an admission score below the median compared to the admitted to a particular program are classified as high risk. In the case of economic vulnerability, the Basic tuition score was used to separate the two risk categories.

1.1.3 Student Profiling

Clustering techniques have also been used to create a descriptive profile of the students. In particular, regarding to the classification objective, clustering have been used to reduce the complexity of the data. For instance Tsai et al. [26] evaluated the results of computer proficiency tests of undergraduate students in a National University in Hong Kong. They used three clustering methods, i.e. K-Means, SOM, and BIRCH, to identify groups; the best clustering model was selected and used as input for a predictive model. BIRCH with k=5 provided the best results, although these were considered based on the similarity of the distribution of the data between training and test results.

After the selection of the clustering model, the decision tree algorithm C5.0 was then applied to each cluster for extracting rules regarding the performance of the students in the two components of the test, skills and discipline based. The performance of the algorithm was evaluated separately on each cluster by the accuracy, with results varying between 78.6 and 82.9% for the discipline based test and 79.9 and 86.9% for the skills based test.

A different approach by Bresfelean et al [27]. was used to identify a student profile for exam success or failure, as part of their work for the Institution managers, in order to offer a better knowledge of the students' situation. The data was collected from online and written surveys, as well as university databases; among the attributes included are: General student data, scholastic situation, scholarship information, interruption of studies, tuition, and opinions. They applied Farthest-first, a variation of K-Means, and C4.5, a classification algorithm. For the clustering method, K was set to two clusters, corresponding to the categories for failing and passing students.

In [28], a prediction of final marks was performed by applying several classification methods directly to the Learning Management System data through a module developed by the authors. Data regarding user activity (posts sent and read, quizzes taken and their results, number of assignments, and time dedicated to assignments or quizzes) was used to predict the final mark. There were three configurations of the experiments using: numerical data, and categorical data, i.e. Fail, pass, good, excellent; with and without resampling the data set in order to deal with an imbalanced data set. Statistical classifiers, decision trees, rule induction algorithms, fuzzy rule learners and neural networks, a total

of 25 classifiers were compared on the global percentage of correctly classified and geometric mean. On average, the results on accuracy were similar for imbalanced data sets, close to 61% but varying between 50 and 67% for the numerical data, and 53 and 66% for the categorical data. Most of the algorithms had worst results when using a balanced data set: however, it is important to notice that imbalanced data sets can produce classifiers with acceptable accuracy results just by classifying everything as the most common class.

Talavera and Gaudioso [29] used data from a LMS, particularly the interactions of the students in an unstructured environment, together with survey information regarding students' background and interests in order to characterize similar behaviors in these collaboration spaces. The EM algorithm was used to create the clusters, which were evaluated by using an external feature, similar to the profitability in marketing studies of customer segmentation; in this case, the external feature is student performance, in terms of the final grade in a specific course. This feature was also used to determine the number of clusters.

Clustering has been used to predict the students' final mark, i.e. Pass or Fail, as done by López et al. [30] used clustering algorithms on forum activity data. They collected the data through a module for Moodle regarding activity, e.g. messages, threads, replies, words: some other attributes were created such as an evaluation of the content, a measure done by the course teacher to the content created by the student, and network measures like centrality and prestige, to perform Social Network Analysis. The forum data, from 114 first year students, was analyzed with clustering algorithms with 2 groups and then were evaluated regarding the classification performance; the experiment was repeated by using Feature Selection. The EM algorithm obtained results similar to those from classification methods like Naive Bayes, Decision Trees, Logistic regression and Neural Networks.

1.2 Data Mining methods

Data Mining is presented in [31] as the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data," and is often characterized by the following common tasks: descriptive techniques, that aim to identify the structure of the data, and predictive techniques, where the goal is to identify the value

of a given variable. Clustering and Association Rules analysis are examples of the first group, whilst decision trees are an example of the latter. These techniques, and the algorithms used in this thesis will be presented below.

1.2.1 Clustering

Clustering is the process of forming groups in such a way that objects from a group are more similar between them than to objects from different groups [32].

- K-Means - K-Medoid

K-Means is one of the most widely used partitional clustering algorithms with more than 50 years of use [33]. It partitions the dataset into K different groups, being K a parameter previously defined by the user. K-Means is an iterative algorithm, it starts with the definition of the initial K centroids which can be understood as the prototypes of the different groups, it is then followed by two steps that are repeated until convergence. In the first step, all data points are assigned to a cluster for which the similarity to the centroid is minimized. In the second step, the centroids of the new formed clusters are recomputed. The final result is a product of several iterations of these two steps, until the centroids don't change anymore.

The quality of the clustering depends highly on the initial configuration; different initial conditions may produce different results, even finishing in a suboptimal clustering. The choice of K, the number of groups is another important issue that requires special consideration.

K-Medoid has a slight difference to K-Means. Instead of considering the centroid, i.e. the mean of a set of points, as cluster prototype, it considers the medoid, i.e. most representative object among the group.
Association Rules

1.2.2 Association Rules Analysis

The process of discovery of association rules is a very well-known problem in the data mining community because of its capabilities in the exploratory analysis. It allows the analysts to find hidden relations in the data. Originally this methodology was thought for

market basket transactions, in which a transaction contains a set of items which were purchased at the same time; however, its use is much broader and it has been applied to other application domains such as web mining, medical diagnosis, bioinformatics, and scientific data [32]. The discovery of association rules consists in finding patterns between disjoint itemsets within a set of records. The problem was introduced by Agrawal, Imielinski and Swami in [34] in the context of market basket analysis motivated by the increasing possibility of storing information of the items purchased in a per-transaction basis. The objective was to be able to find rules like "A customer who buys products X and Y will also buy products Z and W with a probability of c%".

The problem statement [35], [36] considers a set of binary attributes, called items $(I = i_1, i_2, \ldots, i_m)$, a database of transactions (D) where each transaction has an identifier TID and is a set of items called an itemset. A k-itemset is k-length itemset. The support of an itemset is given by the count of transactions where the item can be found. A frequent, or large, itemset is one that is in more transactions than a minimum support (*minsup*) value specified by the user. An association rule is an implication of the form $X \Rightarrow Y$, where X and Y are frequent itemsets and $X \cap Y = \emptyset$ The rule $X \Rightarrow Y$ holds in the transaction set (D) with confidence c, if c% of transactions in (D) that contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the transaction set (D) if s% of transactions in (D) contain $X \cup Y$. A rule is considered frequent if its support is greater than the minsup given by the user. If the rule has a confidence greater than minconf, also provided by the user then the rule is considered strong.

A common strategy for finding all the frequent and strong rules is to decompose the problem into two well-defined problems:

1. Find the frequent itemsets considering syntactic (Which items are of interest?) and support (Significant participation in the database) constraints.
2. Based on the frequent itemsets: Generate the rules $X \Rightarrow Y$ with a confidence above a threshold.

- Frequent itemsets

One of the most famous algorithms for the first subproblem is Apriori [AS94]. It starts by identifying the frequent 1-itemsets in the first pass, F1. The k^{th} pass takes the frequent (k-1)-itemsets previously identified and generates, by a self-join, the candidate itemsets C_k. These and all of their supersets are pruned when any of its subsets is not frequent. The frequent itemsets are then stored in the leaves of a hash tree; internal nodes contains hash tables. This structure is used in the rule discovery stage. The pruning stage is very important because the possible number of combinations grows exponentially with the size of I. This stage is based on the fact that every subset of a frequent itemset is also frequent. If there is an itemset which is infrequent, then all of its supersets are infrequent.

The Apriori approach is effective in finding the frequent itemsets; however, there might be too many of them to handle them properly.

- Rule generation

In general terms, the rule generation is a straightforward process and a huge amount of rules can be generated easily by combining the items of a frequent itemset. A common approach for rule generation partionates frequent itemsets Y into two nonempty subsets, X and $Y - X$ such that $X \Rightarrow Y - X$ satisfies the confidence threshold. The support threshold is accomplished for sure because every subset of a frequent itemset is also frequent. A frequent k-itemset can produce up to $2^k - 2$ association rules.

So, this leads us to a data mining problem of second order, as can be seen in [37], in which the rule discovery process can identify a large amount of rules. Statistical measures such as support and confidence can prune the rules. An exhaustive study of 21 objective measures can be found in [38], where Tan et al. proposed some properties for such measures and concluded that there is no one measure better than others in all application fields. Instead, the right measure should be selected based on the properties required in the specific application domain. Although there is not a best measure they identified two situations where all measures have a similar behavior and become highly correlated with each other. (a) When the support constraint is low, and (b) when the contingency tables are standardized.

1.2.3 Classification

In a classification task, the groups are already known, so the objective is to assign a record to a predefined label or class. It can be seen as: Given a set of known attributes, estimate an unknown value; when this value is categorical, it is known as classification, when is numerical it is known as regression.

An important feature of a classification model is that it is built using part of the data, the training set, which is used to learn the model. In this subset all the attributes are known, including the class. After the model is built, it is used to assign a label to new records where the class attribute is unknown.

- Decision Trees

A decision tree is a representation made out of nodes and arcs where an internal node presents a decision based on attribute values, and the arcs represent the choice made in the node. It ends on a leaf node, which represents the label or the class to be assigned. To classify a record with a decision tree, it starts by the root node and goes down one level at a time depending on the results of the conditions tested on every node; when it ends on a leaf node, the record is classified according to the label of that leaf node.

In this work, the C4.5 algorithm is used which is based on Hunt's algorithm [39]. It has an important feature which is its ability to manage both, discrete and continuous attributes.

The idea behind the building of a decision tree is the following:

1. If the stop criterion is satisfied all the records in the set are from the same class Y, then the node is a leaf node and is labeled Y.
2. If the records are from different classes, the algorithm selects the attribute that can split the records into smaller and purer subsets. The default splitting criterion is the gain ratio, but there are other measures that are used for selecting the best split, for instance: Entropy, Gini and Classification error.

This procedure is performed recursively and it is done until all the nodes are from the same class or have the exact same attributes; however, this can lead to a 100% pure configuration in which case it is most likely to have overfitted the model. In such cases,

the model has a very low number of classification errors on the training set but it is quite large when applying the model to a previously unseen set, or a test set. To overcome this situation, it is common to have an earlier stopping condition, i.e. prepruning, for instance, when there are a minimum number of records in the splitted subset. Another approach is to prune after the tree is fully grown, i.e. post-pruning, for instance, by replacing a subtree with its most used branch, or with a leaf node with the label defined by the majority class.

- Bayesian Classifier

A Bayesian Classifier [32] considers a probabilistic relationship between the class and the attributes, instead of a deterministic relationship where a given set of attributes not always have an identical label outcome. The classification task, to classify a record depending on its attributes values, can be expressed as the probability of a record of being from the class Y, given that the record has a set of attributes X. That is $P(Y = y|X = x)$.

This can be calculated by using the Bayes Theorem

$$P(Y = y|X = x) = \frac{P(X=x|Y=y) * P(Y=y)}{P(X=x)}$$

where

$P(Y = y|X = x)$ is the posterior probability of Y,

$P(X = x|Y = y)$ is the class-conditional probability,

$P(Y = y)$ is the prior probability, or the probability that the class is labeled as y,

$P(X = x)$ is the probability of the set of attributes, or the evidence.

In order to classify an instance, the classifier looks to maximize the posterior probability. Naive Bayes makes a strong assumption, that the attributes are conditionally independent given the class. With that in mind, the class-conditional probability can be expressed as:

$$P(X_i|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

And the posterior probability as:

$$P(Y|X) = \frac{\prod_{i=1}^{d} P(Y) * P(X_i|Y = y)}{P(X)}$$

The selected class is the one that maximizes the numerator, because the denominator is the same for all classes.

2. Problem and data understanding

In this chapter, the Universidad Nacional de Colombia and the data sets are presented, along with a description of the data sources and an exploratory data analysis of the population chosen for this study. Finally, the general Data Mining model is presented which will be further developed in chapters 3 and 4.

2.1 Universidad Nacional de Colombia

The Universidad Nacional de Colombia, Bogotá campus, is the largest University in the country with 49 undergraduate programs. The admission process is held every semester with more than 50000 applicants to the first academic period of the year and more than 30000 to the second; however, only close to 3200 are admitted. These students are selected based exclusively on one criterion, their performance in the admission test, an academic exam that evaluates five components: Mathematics, sciences, social sciences, text analysis and image analysis. Every academic program has a previously defined number of places that are occupied by the students with higher marks, ensuring a high academic quality of the students.

Once the students are part of the university, they can lose their student status by academic or non-academic reasons. Among the first type are the losses due to a low academic performance, take for instance failing in more than 2/3 of the subjects in one academic period, failing a subject three times, or failing it a second time with an insufficient GPA; these were reviewed after the Academic reform of 2008, and are not considered anymore; instead, based on the credits of a particular subject, there is a Weighted GPA requisite, it cannot be lower than 3.0. In addition, a quota of academic credits was included. The non-academic reasons are mostly voluntary retirements, students who don't enroll in the academic period. Transfer between programs or campuses are considered here.

To develop and test the model, a sample of the population was used, corresponding to students admitted and enrolled to two engineering programs, Computer and Agricultural Engineering. The former was chosen based on the previous experience on the program and an understanding of its dynamics; however, after an early Exploratory Data Analysis it was clear that the program didn't show so much of the variety: the student population gender (less than 10% are women), or the option in which the student chose the program, this attribute can be seen as a proxy for the student's motivation to join the program (more than 95% of the students chose computer and systems engineering as the first option). Based on this, the agricultural engineering program was chosen to complement the model.

2.2 Data sets

Data was collected from two sources, the Integral System of the National Direction of Admissions (DNA) and the Academic Information System (SIA). DNA collects the information from the biannual admission process and includes the admission test scores results, the options for enrollment and some socio-demographic attributes. On the other hand, SIA includes data of the academic life of the student; three datasets were used regarding grades and credits, loss of academic status, and student enrollment records per academic period.

The four datasets used in this study are commonly requested reports generated by the SIA, the Direction of Admissions, and the Bogotá campus' Division of Registry. The reports, originally in Excel format were imported to a MySQL database created for this research.

The admissions data set fields considered in this study can be grouped in three categories that will be described briefly:

> **Academic potential:** Admission test score in five modules (i.e. Sciences, Math, Image, Text and Social studies) and classification levels for Basic Math and Literacy.

Demographic and socio-economic: Age at Admission, Gender, city of origin, '*estrato*' (i.e. socio-economic classification), ethnicity.

Previous academic information: high school type (e.g. public, private), type of access (e.g. regular, special admission program), option in which the student chose the program (from 1, first option, to 3), and the previous program, if exists.

All SIA datasets includes fields to identify the academic period, the student and the program in which is enrolled. Other fields of these datasets will be presented below.

The enrollment report is composed of records of the students enrolled at a given academic period. It includes different fields regarding student data, number of enrollments and general academic performance, such as GPA, weighted GPA, and total approved credits (however these values are only from the last academic period available before the query was performed, for that reason it was not considered in this research).

The grades report has data of the courses taken in each academic period by a student and their final results. Some of the fields regarding the courses are: course ID, course name, course section, number of credits, numeric grade (0 to 5), qualitative grade (approved and not approved) and the typology of the course, i.e. professional, foundation, optional electives, and leveling courses.

The loss of academic status report registers when a student's academic history is blocked. Some of its fields are the code of the blocking, the description, date and academic period; active, if the blocking is still active or not; and, in some cases, the information of the unlocking of the academic history: code, description, date and academic period. The codes for blocking a student academic history were classified into academic, non-academic, and others in a process together with representatives of the National Direction of Undergraduate Programs, the National Direction of Graduate Programs, the National Planning Office, and the Bogotá campus' Office of Academic Affairs.

The loss of academic status is considered academic when is related to a low academic performance, non-academic if the academic performance requirements were still fulfilled

but the student didn't enroll in that academic period. The 'others' category is used for administrative reasons. The academic category is the only one considered in this research. A summary of these categories for undergraduate students is presented below in Table 2-1.

Table 2-1: Categories of loss of academic status for undergraduate students.

Academic	Non-academic	Others
▪ To fail in more than 2/3 of the subjects in one academic period ▪ To fail a subject three times ▪ To fail a subject two times and have an insufficient GPA ▪ To have a Weighted GPA lower than 3.0 ▪ To have an insufficient quota of academic credits	▪ Transfer Program or campus. ▪ Withdrawal for not renewing enrollment in the limits set by the University. ▪ Cancellation of registration due to suspension. ▪ Expulsion from the University. ▪ Illness substantiated.	▪ Double degree ▪ Campus of National presence ▪ 033 agreement transfers

The specific preprocessing for the student characterization and the classification models will be described in chapters 3 and 4 respectively.

2.3 Exploratory Data Analysis: the student population

As it was mentioned above, the present research studies the Agricultural and Computer and Systems Engineering programs, specifically to the cohorts from 2007-03 to 2012-02. Both admit around 100 new students on each academic period. The population of admitted students is described below.

In terms of academic potential, which is measured by the results in the admission test, Computer and systems engineering admitted students show a good performance, especially in Math and Image analysis components where, in average, they were always above the Campus average.

Table 2-2 presents the average of the admission test results for the different components. The bold font, in color, is used when the average of the program is greater than the average of the Faculty and the colored cells represent an average above the Campus average.

Table 2-2: Computer and Systems Engineering. Admission test results by component.

Bold font represents an average greater than Faculty's average and colored cell greater than Campus' average

Period	Image	Sciences	Math	Text	Social	Total
2007-03	**25.637**	**36.660**	**42.733**	**25.325**	**44.741**	**175.097**
2008-01	**26.770**	38.270	46.527	**20.907**	**47.885**	180.359
2008-03	**11.448**	10.971	**12.063**	**11.131**	**11.334**	**711.271**
2009-01	**11.730**	11.334	11.674	11.305	**11.354**	718.518
2009-03	**11.503**	11.323	11.625	**11.186**	**11.269**	**697.219**
2010-01	11.315	**11.794**	11.951	**11.343**	**11.643**	**726.776**
2010-03	**11.214**	11.214	11.634	11.067	11.152	679.479
2011-01	11.775	11.447	11.793	**11.241**	11.382	722.788
2011-03	**11.317**	11.260	11.747	11.114	11.034	690.405
2012-01	**11.691**	11.464	11.962	**11.253**	11.155	724.901
2012-03	11.311	11.350	11.351	10.988	11.061	677.941

Agricultural engineering admitted students have lower results compared to both, the Faculty and the campus, but their results are still good among all the applicants, being close to one standard deviation in all components. The tests, from 2008-03 have a mean of 10 and a standard deviation of 1. Table 2-3 presents the results for Agricultural engineering admitted students.

Table 2-3: Agricultural Engineering. Admission test results by component.

Bold font represents an average greater than Faculty's average and colored cell greater than Campus' average

Period	Image	Sciences	Math	Text	Social	Total
2007-03	22.947	32.188	35.875	23.507	42.055	156.570
2008-01	24.912	35.181	41.664	19.286	46.335	167.377
2008-03	10.986	10.812	11.261	10.902	11.123	654.168
2009-01	11.238	10.873	11.201	11.057	11.157	663.601
2009-03	11.003	11.119	11.215	11.017	10.906	648.776
2010-01	11.245	11.023	11.434	10.907	11.057	664.552
2010-03	10.935	10.991	11.063	10.875	11.026	639.368
2011-01	11.629	10.895	11.178	10.774	11.005	669.416
2011-03	10.937	11.271	11.226	11.154	11.017	665.730
2012-01	11.120	11.220	11.419	10.918	10.947	669.454
2012-03	11.038	11.046	10.974	10.641	10.884	636.460

There are similarities between the admitted students of both programs in several variables such as estrato where almost 80% of the population belongs to estratos 2 and 3 and less than 5% are from the higher estratos, i.e. 5 and 6. The region of origin is also very similar; around 70% of the students are from Bogotá. The type of school is almost equally divided into public and private schools, with a percentage of 46%. The age distribution is similar for both programs. Overall, around 55% of the admitted students are 17 or younger, this fraction increases in the processes for the first academic period of the year, e.g. 2007-01, 2010-01.

Besides the test results, the differences are present in gender distribution and the option in which the applicants selected the program. Regarding gender, there is a high imbalance, a common situation in the Faculty of Engineering. In Agricultural Engineering, 26% of the admitted students are women and in Computer and Systems Engineering only 7% are. On the other hand, the option of enrollment presents a large difference. 95% of computer and systems engineering admitted students chose it as their first option, but only 22% of the agricultural engineering admitted students did that. Half of the students joined the program as their second choice.

This can be reflected in the enrollments in first semester. Table 2-4 presents the number of enrollments in first semester

Table 2-4: Number of first semester enrollments. Not considering transfers

Period	AE	CE
2007-03	55	77
2008-01	66	65
2008-03	38	72
2009-01	56	75
2009-03	43	71
2010-01	70	82
2010-03	70	80
2011-01	77	86
2011-03	46	84
2012-01	67	83
2012-03	74	95

AE: Agricultural Eng. CE: Computer and systems engineering

Table 2-5 shows the number of enrollments and the number of loss of academic status due to low academic performance. It is important to notice that the number of enrollments shown below only considers the enrolled students with a corresponding record in the admissions dataset who have a grade report in the given academic period.

Table 2-5: Number of enrollments and academic blockings per academic period and program

Period	AE		CE	
	Enrollment	BLQ	enrollment	BLQ
2007-03	55	4	77	0
2008-01	110	9	128	7
2008-03	134	25	188	21
2009-01	152	17	236	30
2009-03	170	23	272	20
2010-01	212	36	340	28
2010-03	231	30	392	35
2011-01	276	34	443	25
2011-03	274	23	489	18
2012-01	305	39	553	38
2012-03	334	40	598	30

The number of loss of academic status due to low academic performance greatly increased after the academic reform, an implementation that started in the academic period of 2009-01. Tables 2-6 and 2-7 show the number of blockings according to the period and period of first enrollment.

Table 2-6: Academic blockings per academic period - Agricultural Engineering.

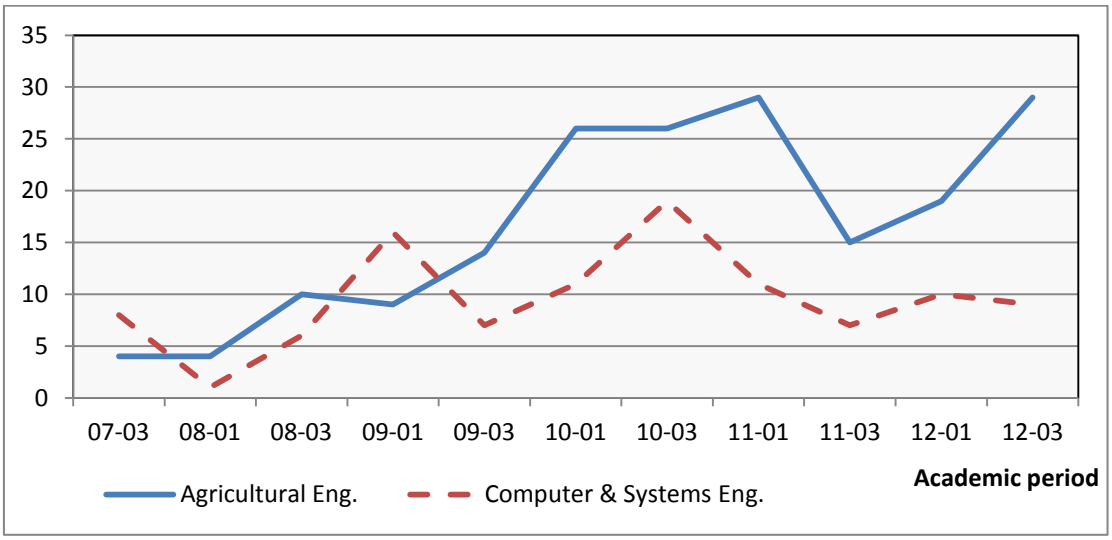
		BLQ_Acad										
		07-03	08-01	08-03	09-01	09-03	10-01	10-03	11-01	11-03	12-01	12-03
Enrollment	07-03	4	5	8	0	3	0	0	1	1	1	0
	08-01		4	7	5	3	0	1	1	1	2	0
	08-03			10	3	2	1	0	0	0	0	0
	09-01				9	1	1	0	0	0	1	1
	09-03					14	8	0	0	0	1	0
	10-01						26	3	1	0	0	2
	10-03							26	2	0	2	0
	11-01								29	6	8	1
	11-03									15	6	1
	12-01										19	6
	12-03											29

Table 2-7: Academic blockings per academic period - Computer and Systems Engineering.

		BLQ_Acad										
		07-03	08-01	08-03	09-01	09-03	10-01	10-03	11-01	11-03	12-01	12-03
first enrollment	07-03	8	6	9	3	5	1	0	0	1	2	0
	08-01		1	6	4	1	3	1	1	0	0	0
	08-03			6	8	3	3	2	2	0	1	1
	09-01				16	4	6	5	2	1	0	1
	09-03					7	4	1	2	1	3	1
	10-01						11	7	2	1	2	1
	10-03							19	5	4	3	4
	11-01								11	3	6	5
	11-03									7	11	4
	12-01										10	4
	12-03											9

As can be seen in the tables, the most critical is first semester, especially after the implementation of the academic reform. This is presented is Figure 2-1.

Figure 2-1: Academic blocking in the first enrollment.



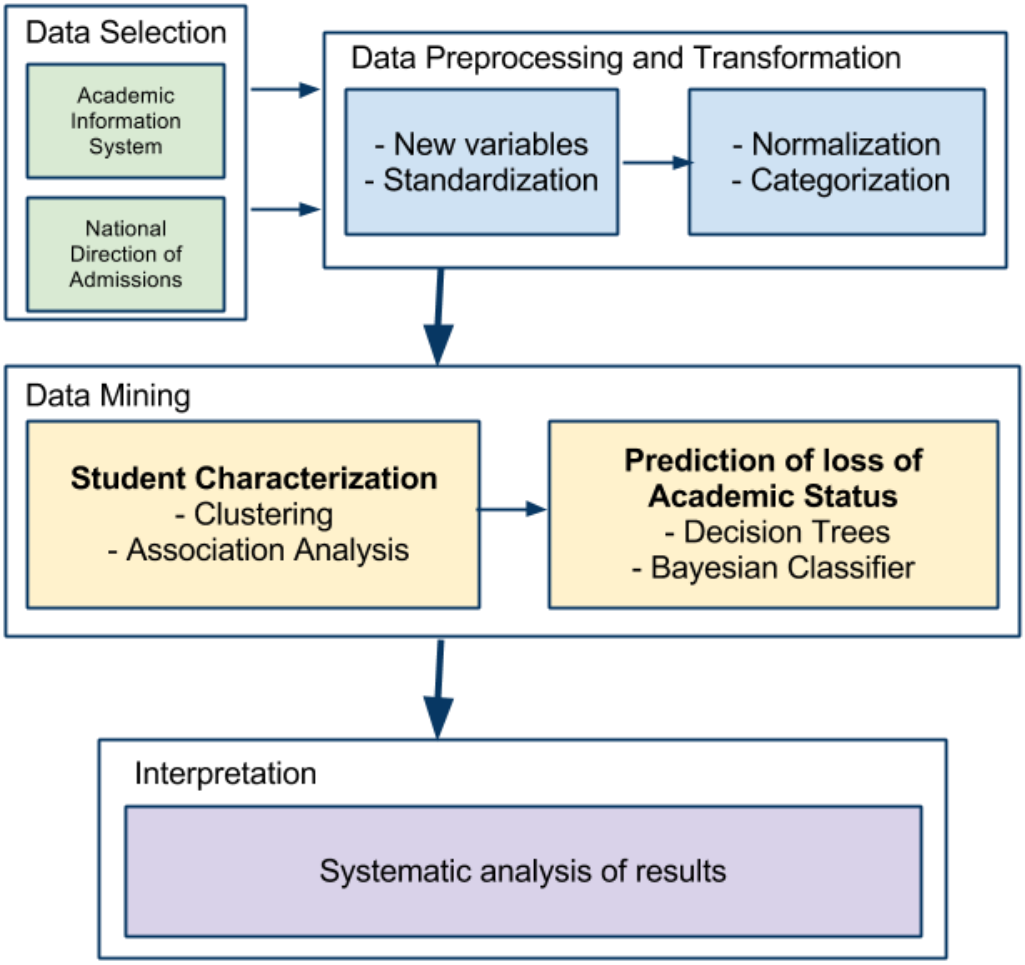
2.4 Data Mining Model

Given the interest to detect those students that are at risk of losing their academic status early in their academic life, a Data Mining model is proposed. The entire process follows the KDD process presented in [31] by Fayyad et al., which can be seen in Figure 2-2. It

starts with a data selection in which the mining view is built from data sets from the Academic Information System and the Admission process. It is followed by the Data preprocessing and transformation where the data is prepared for the application of Data Mining algorithms. These were partially described in this chapter, and will be complemented in chapters 3 and 4.

The Data Mining model is divided into two phases. In the first one, the aim is to characterize a student population by using descriptive Data Mining techniques in order to get a better sense of the population. In the second one, classification methods are used to predict the loss of academic status based on the characterization mentioned above, and the academic records registered by the students in each academic period.

Figure 2-2: General KDD Process.



3. Student characterization

This chapter describes the first phase of the Data Mining model, the student characterization. In this phase, a clustering model was built using the K-Means algorithm. The process of data preparation and experimental design are presented.

3.1 Data preparation

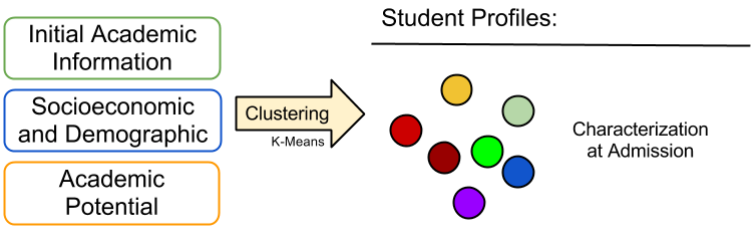
To characterize the students, only the data collected during admission process was used. The preprocessing phase includes the creation of secondary variables such as: Age at admission, from the date of Birth and the period of admission, calculating the age of the student on the first day of the second month of the semester, i.e. February or August; residency, from the city of origin (i.e. Bogotá or out of Bogotá); and finally, Boolean variables to represent the presence or absence of disabilities or belonging to ethnic groups.

The admission scores have a mean of 10 and a standard deviation of 1, except the academic periods of 2007-03 and 2008-01. These were standardized to meet these characteristics. The type of school attribute was modified to include information of the equivalency diploma (‘*validación de estudios*’), foreign schools and students who don’t report an institution.

3.2 Student characterization model

To characterize the admitted students, two techniques were used. A clustering technique, namely K-Means, is used to create the different student profiles, then, to complement the visual exploration of the clustering results, association rules analysis is used to detect rules within the clusters. The model is presented in Figure 3-1.

Figure 3-1: Student characterization model.

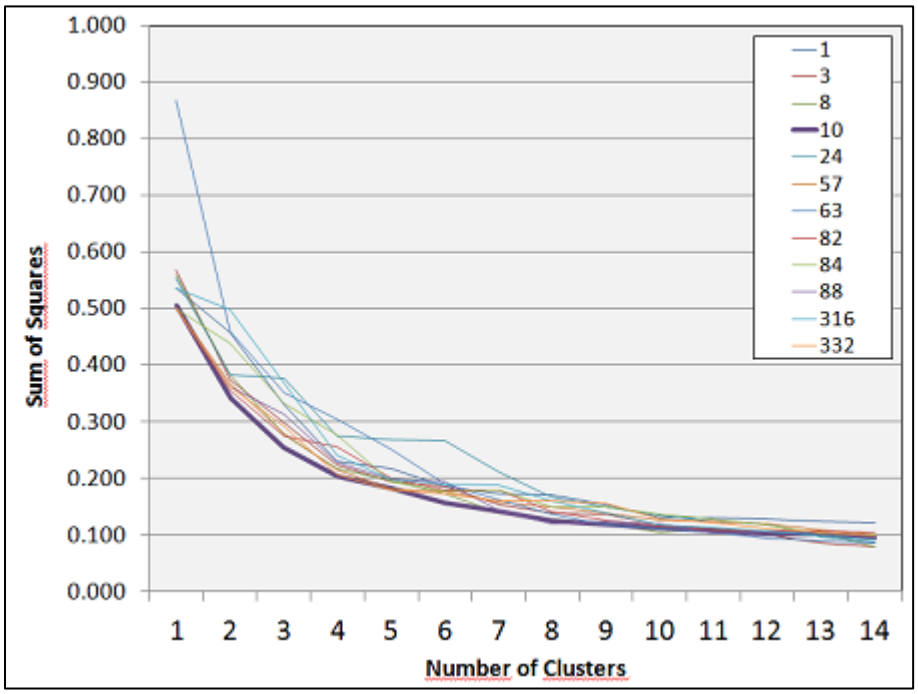


3.3 Experimental design and validation

The K-Means algorithm was used to create the clusters. The tool used to perform the experiments was RapidMiner, an open source toolbox for Machine Learning [40], the implementation of the algorithm is the operator W-SimpleKMeans, the operator is part of Weka's open source library [41]. W-SimpleKMeans uses the Euclidian distance and the sum of squares to evaluate the quality of the clustering. This implementation of the algorithm handles both, numerical and categorical values; however, it was necessary to perform an additional preprocessing to normalize all numerical attributes between 0 and 1 so, no bias were included because of the magnitude of the values. To overcome the sensitivity to the starting configuration, the model was trained using different sets of initial points, which are defined by different seeds.

To choose the model, the algorithm was run several times, varying the number of clusters, starting from 2 to 14; and the set of initial points, 10 different seeds. On each run, the sum of squares was measured to evaluate the quality of the clustering. The number of clusters is selected considering the number K, after which there is no considerable change in the sum of squares value. The results were plotted, in the X-axis are the number of clusters and in the Y-axis the corresponding sum of squares value; the seeds, or initial configurations are represented by the lines. K was chosen based on a visual inspection as can be seen in Fig. 3-2. The number of clusters was set to 8 for both programs. This inspection also allowed the selection of the seed; in the case of the figure, the seed 10 was chosen, which means that the initial configuration of centroids that minimizes the clustering was defined by that seed.

Figure 3-2: K-Means – Selection of the number of clusters.



3.4 Analysis of results

The clusters of the admitted students are presented here. First, the algorithm only uses one program, Agricultural and Computer and Systems separately. Then, the two programs are used at the same time; finally, only the students from both programs who had a fourth enrollment are considered.

3.4.1 Agricultural Engineering Clustering

In the case of Agricultural Engineering the clusters differentiated each other by these key attributes: gender, residency, type of high school formed. In terms of the academics, the clusters present a similar performance with small differences among the different components. Characteristics for each cluster are presented below. A graphic display of some of the characteristics of each cluster implementation can be found in Figures 3-3 to 3-10.

Cluster 0: Mostly women from private schools in Bogotá and a medium-high ‘estrato’. They are enrolled as second and third option. Figure 3-3.

Cluster 1: Similar to cluster 0 but with a majority of men. This is the cluster with the highest number of students who chose the program as third-option (72%). Figure 3-4.

Cluster 2: The main characteristic of this cluster is the region of origin, since all of the students come from out of Bogotá, they are also from public schools and lower 'estrato'. Figure 3-5.

Cluster 3: Mostly men from Bogotá, medium 'estrato' and older students (around 45% are older than 18). Figure 3-6.

Cluster 4: Similar to cluster 3, with a bigger presence of lower 'estrato' students and first-option students. Figure 3-7.

Cluster 5: Mostly women from public schools in Bogotá and a lower-medium 'estrato'. This cluster also has a majority of first-option students. Figure 3-8.

Cluster 6: Similar to cluster 5 but with a majority of men. There are no students enrolled as first-option. Figure 3-9.

Cluster 7: Similar to cluster 3 but with a youngest population. Academically, this cluster didn't perform well in the text analysis component, with an average lower than the mean. Figure 3-10.

Figure 3-3: Cluster 0 – description of variables – Agricultural Engineering.

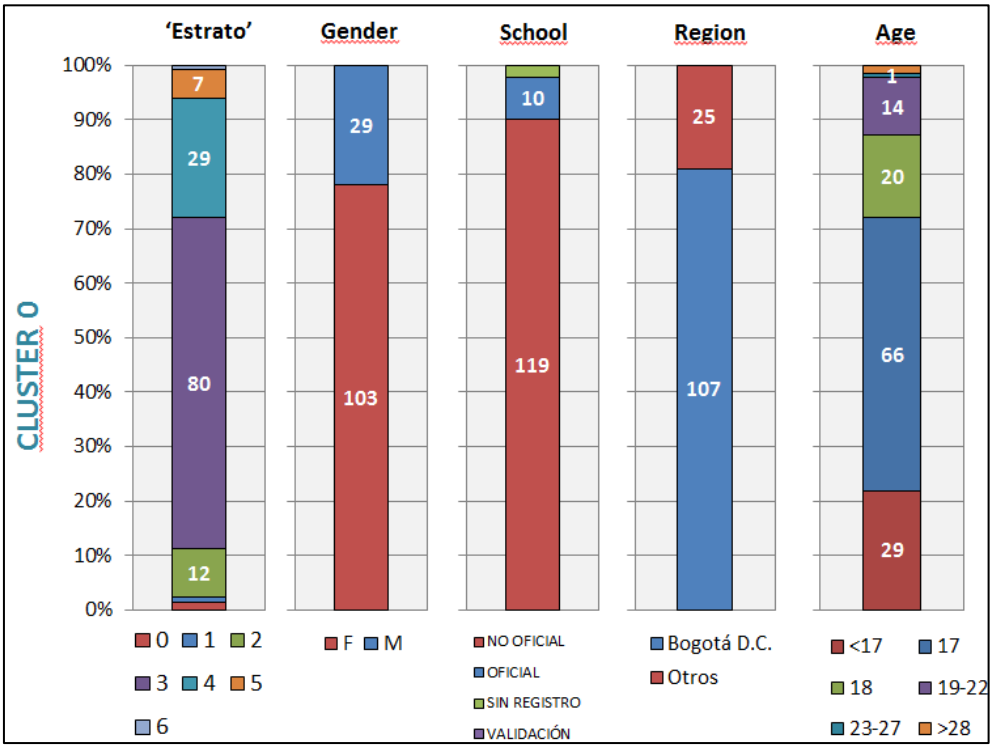


Figure 3-4: Cluster 1 – description of variables – Agricultural Engineering.

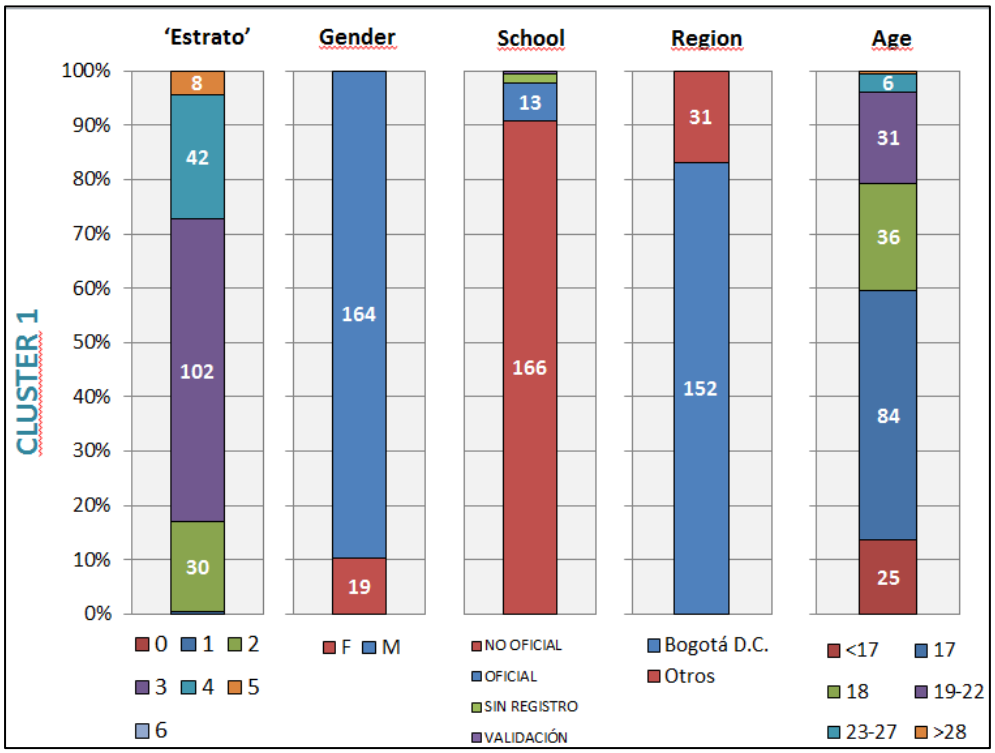


Figure 3-5: Cluster 2 – description of variables – Agricultural Engineering.

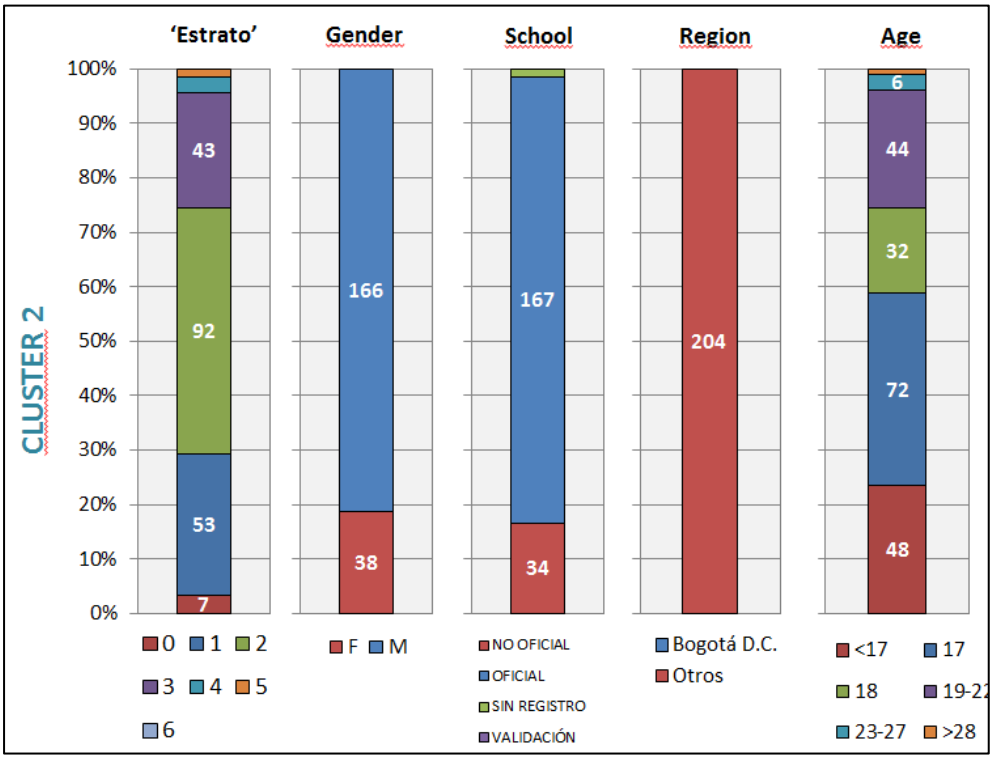


Figure 3-6: Cluster 3 – description of variables – Agricultural Engineering.

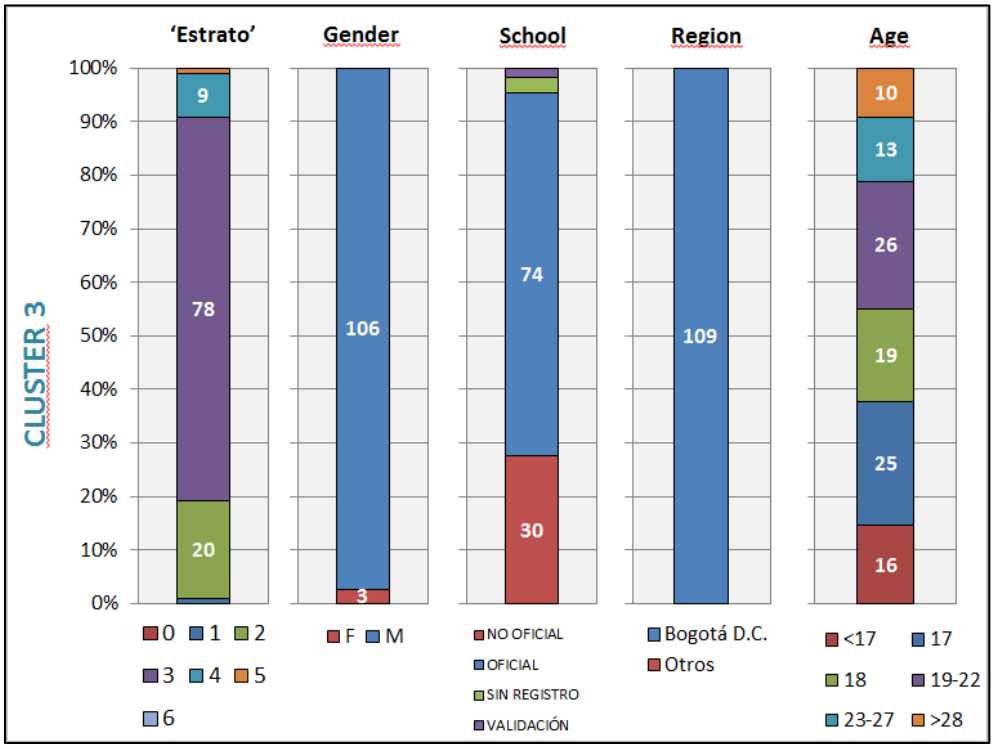


Figure 3-7: Cluster 4 – description of variables – Agricultural Engineering.

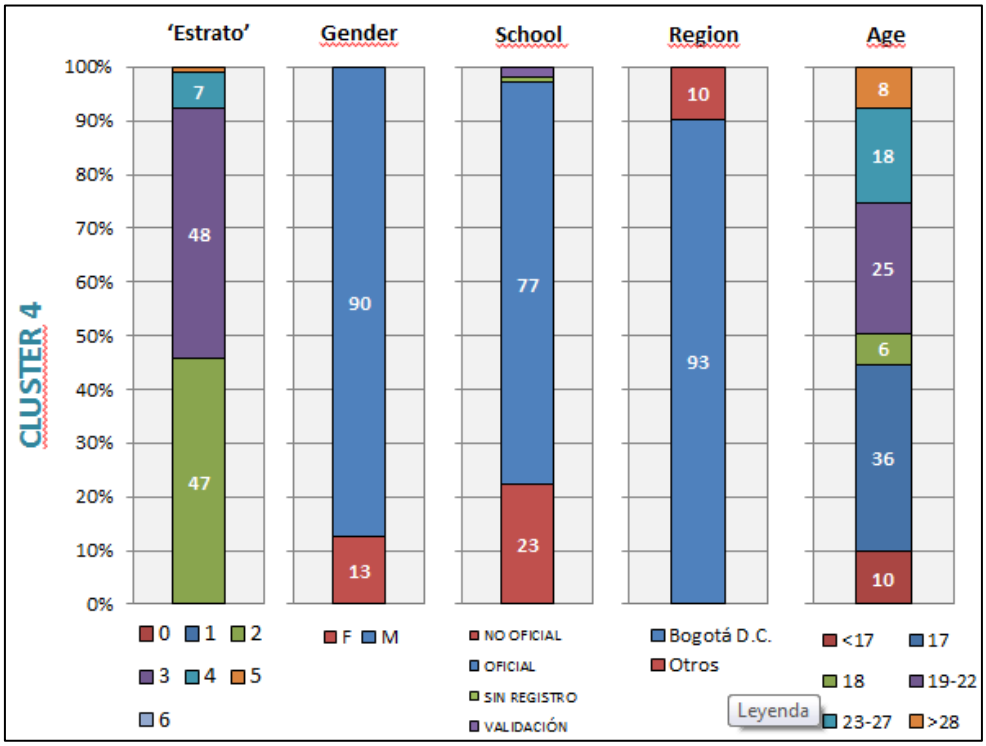


Figure 3-8: Cluster 5 – description of variables – Agricultural Engineering.

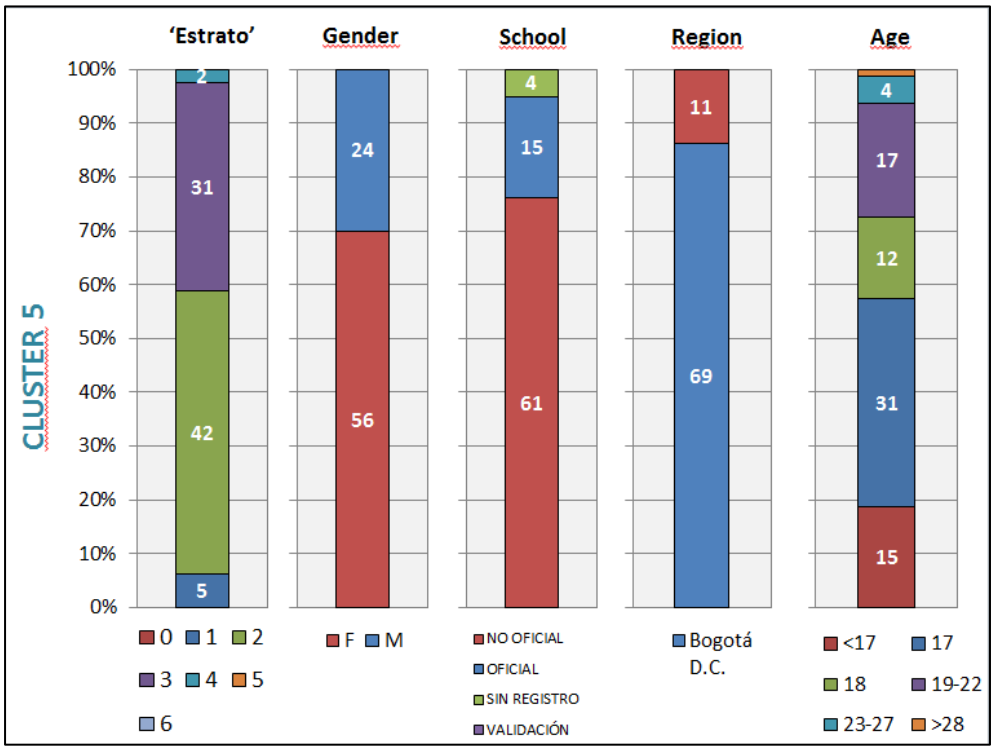


Figure 3-9: Cluster 6 – description of variables – Agricultural Engineering.

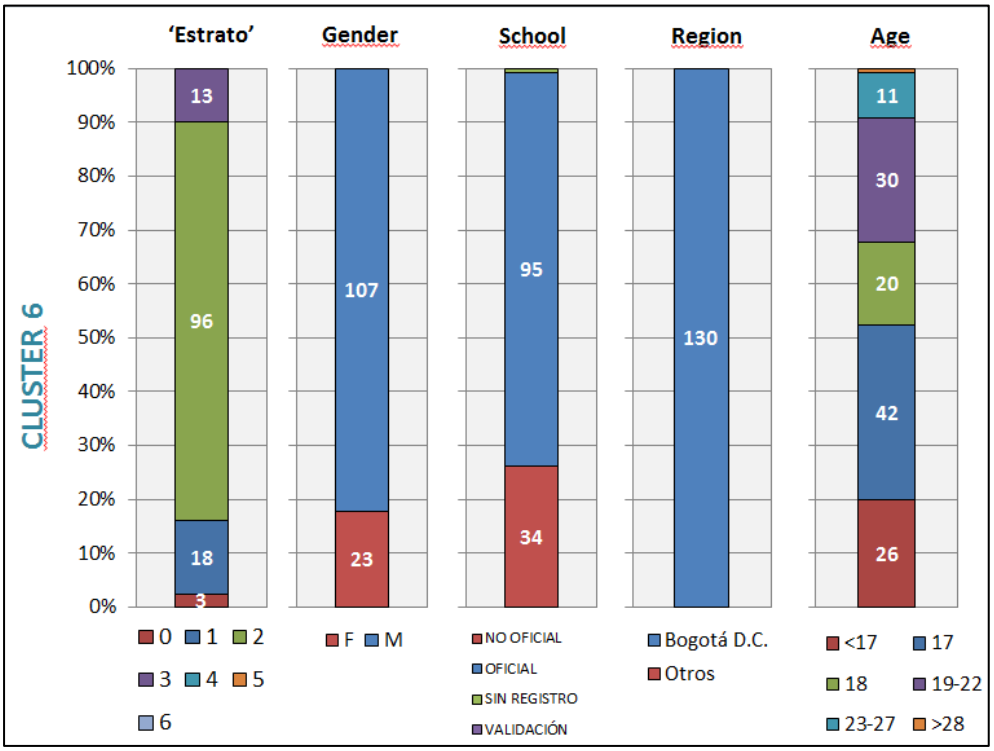
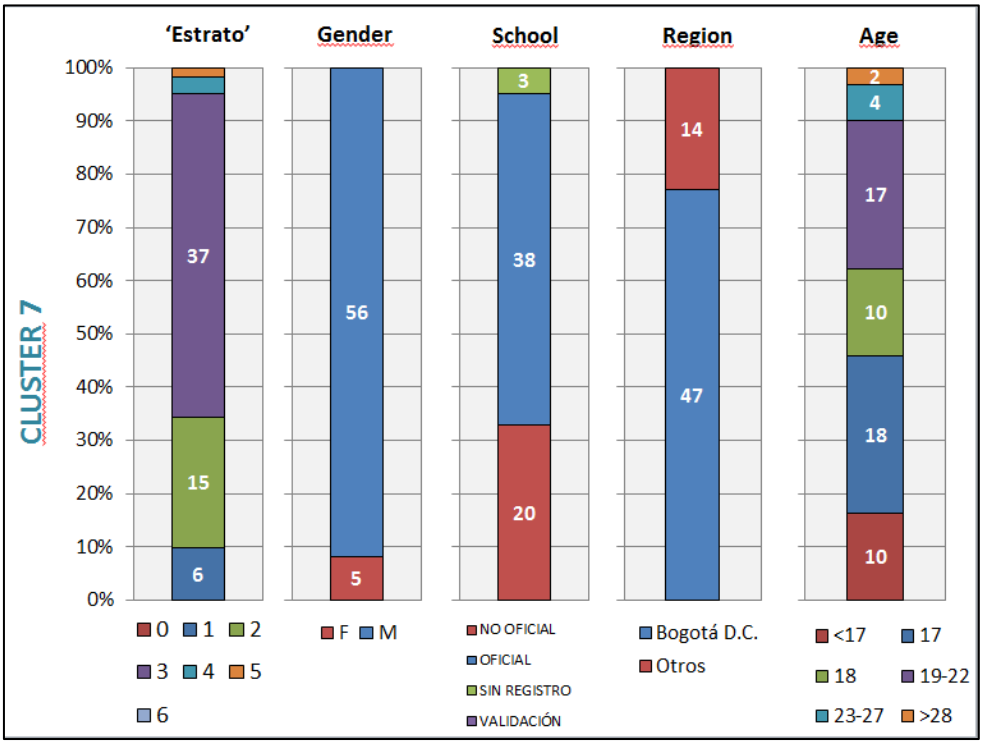


Figure 3-10: Cluster 7 – description of variables – Agricultural Engineering.



3.4.2 Computer and Systems Engineering Clustering

Computer and Systems Engineer students formed similar clusters, however, since there are no large differences in gender, and option for the program, the results tended to have clusters less differentiated. Another similarity between all groups was the performance in the Math component of the admission test. Here are some of the characteristics of these clusters. Results are presented in Figures 3-11 to Fig. 3-18

Cluster 0: A cluster formed by men from public schools from out of town and lower-medium 'estrato'. It also has the lowest results in the test scores. Figure 3-11.

Cluster 1: Students from private schools from medium 'estrato'. Figure 3-12.

Cluster 2: The main characteristic of this cluster is the residency, since all of the students come from out of Bogotá, from public schools and lower 'estrato'. The average scores were superior to those of other clusters, except from image analysis where they are ranked second. Figure 3-13.

Cluster 3: younger students from private schools, mostly in Bogotá and high 'estrato'. Figure 3-14.

Cluster 4: This cluster had significantly better results in the image analysis component, where students' average score was more than one standard deviation higher than Math average score and almost two the other three components, which are the lowest results in the entire sample. Demographically, it is formed by students from lower-medium 'estrato', public school and older students (around 45% are older than 18). Figure 3-15.

Cluster 5: a young population from public school and lower 'estrato'. Figure 3-16.

Cluster 6: a young population from public school and medium 'estrato'. Figure 3-17.

Cluster 7: Similar to cluster 2 but with a youngest population. Figure 3-18.

Figure 3-11: Cluster 0 – description of variables – Computer and Systems Engineering.

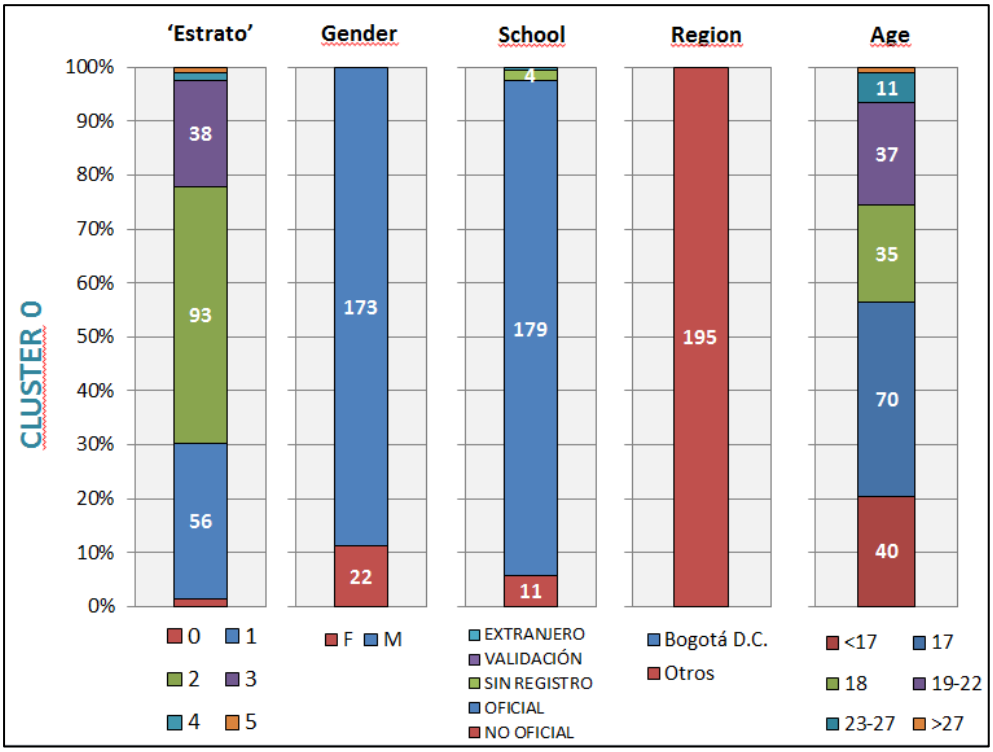


Figure 3-12: Cluster 1 – description of variables – Computer and Systems Engineering.

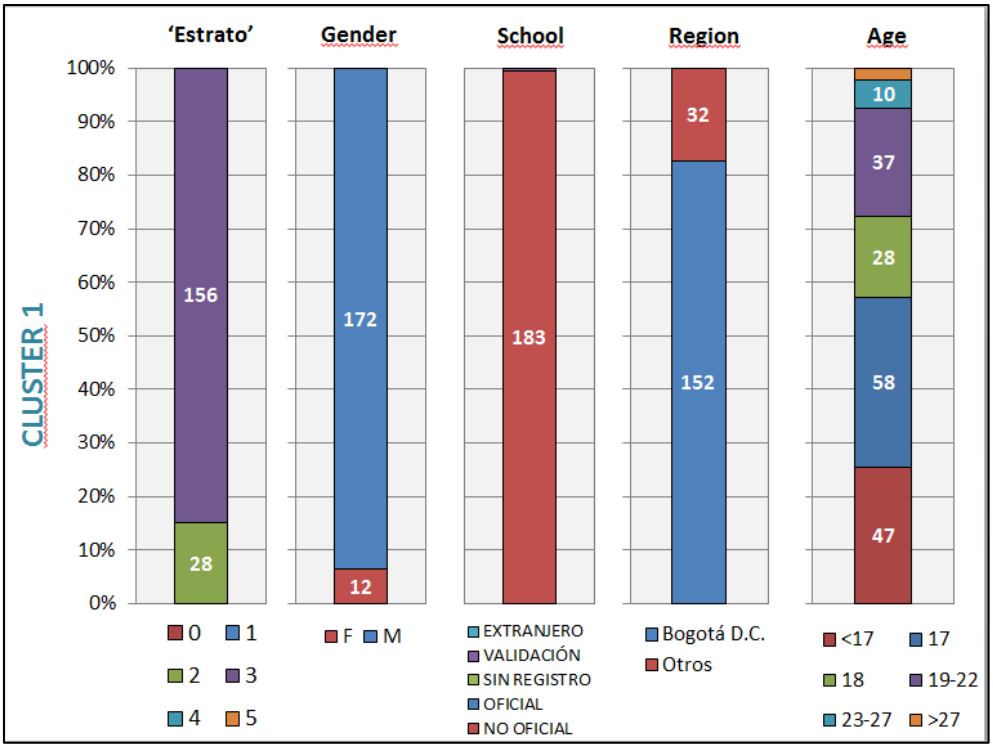


Figure 3-13: Cluster 2 – description of variables – Computer and Systems Engineering.

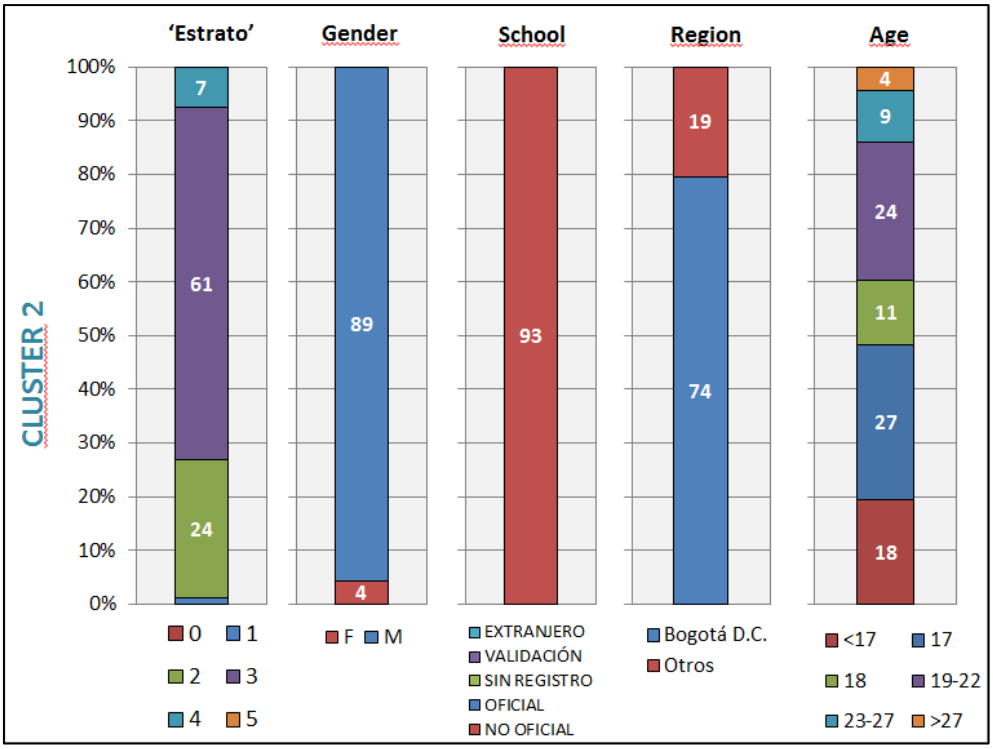


Figure 3-14: Cluster 3 – description of variables – Computer and Systems Engineering.

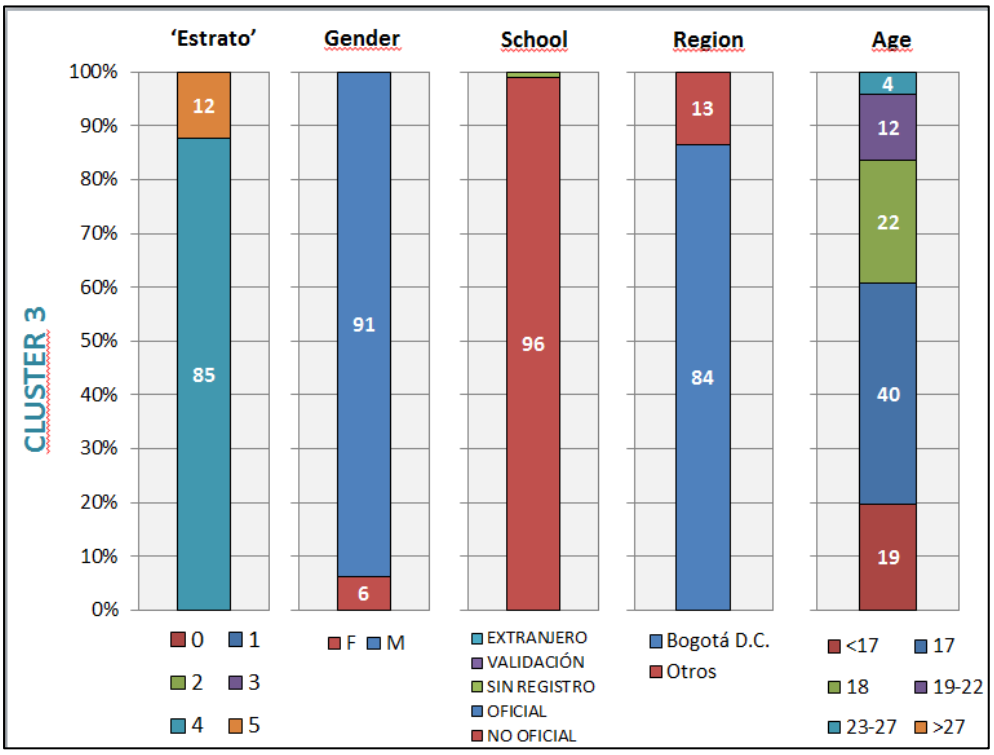


Figure 3-15: Cluster 4 – description of variables – Computer and Systems Engineering.

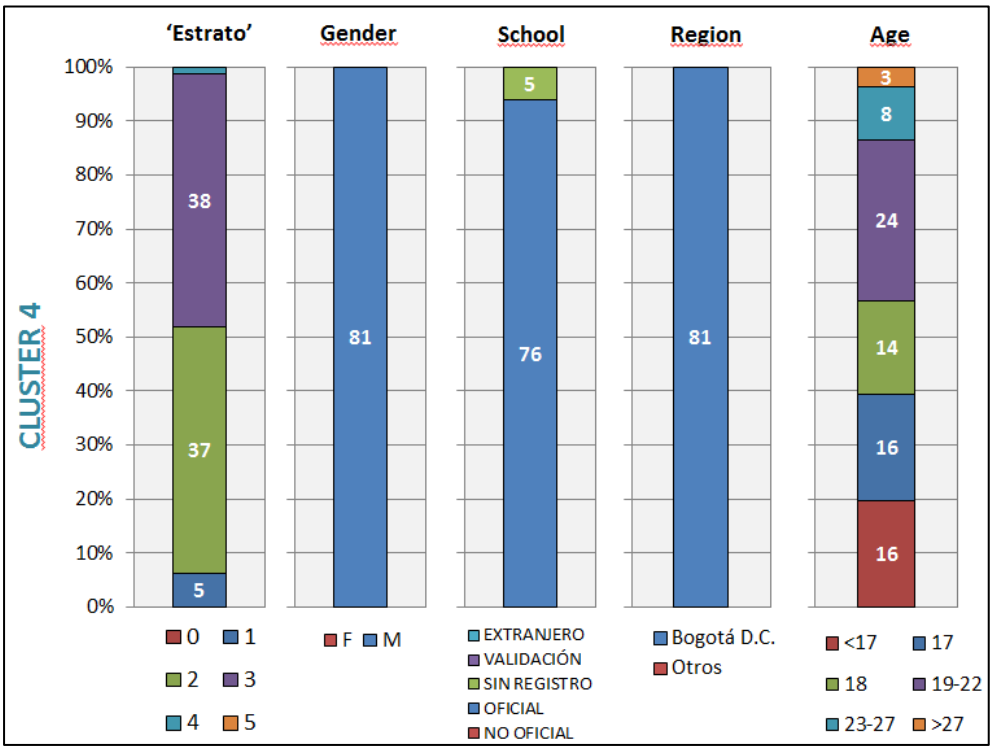


Figure 3-16: Cluster 5 – description of variables – Computer and Systems Engineering.

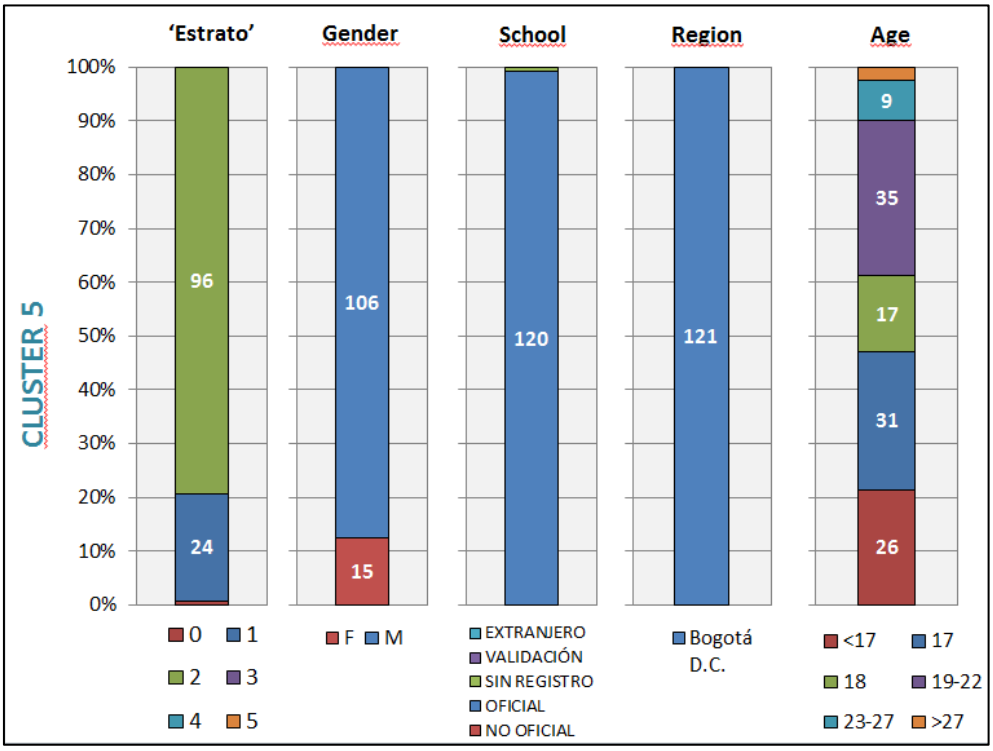


Figure 3-17: Cluster 6 – description of variables – Computer and Systems Engineering.

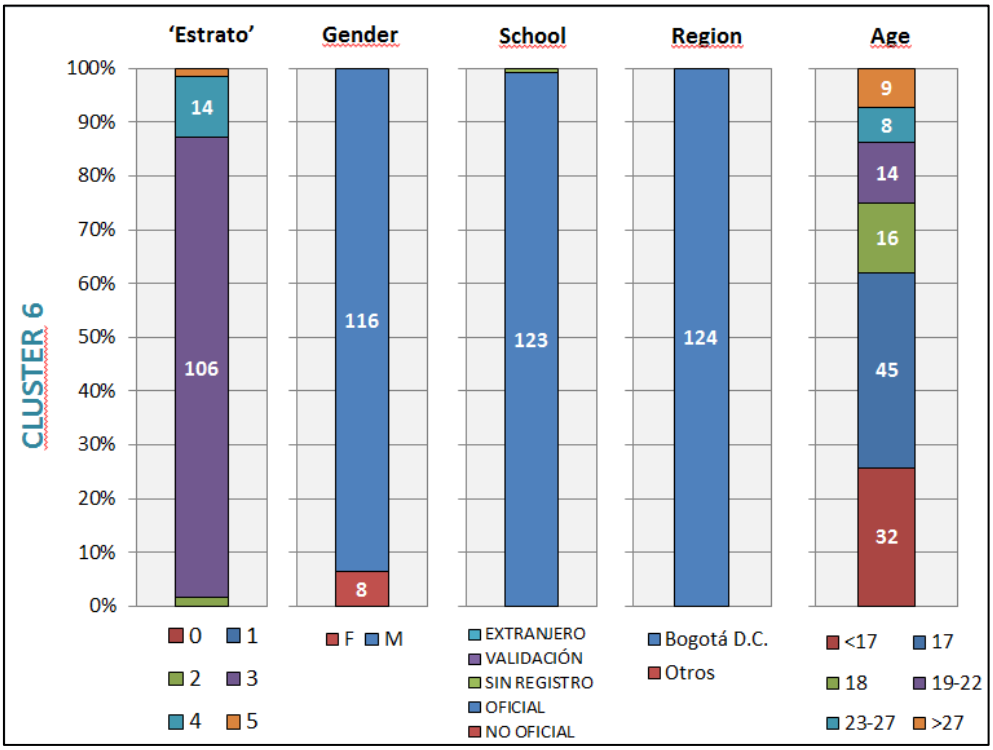
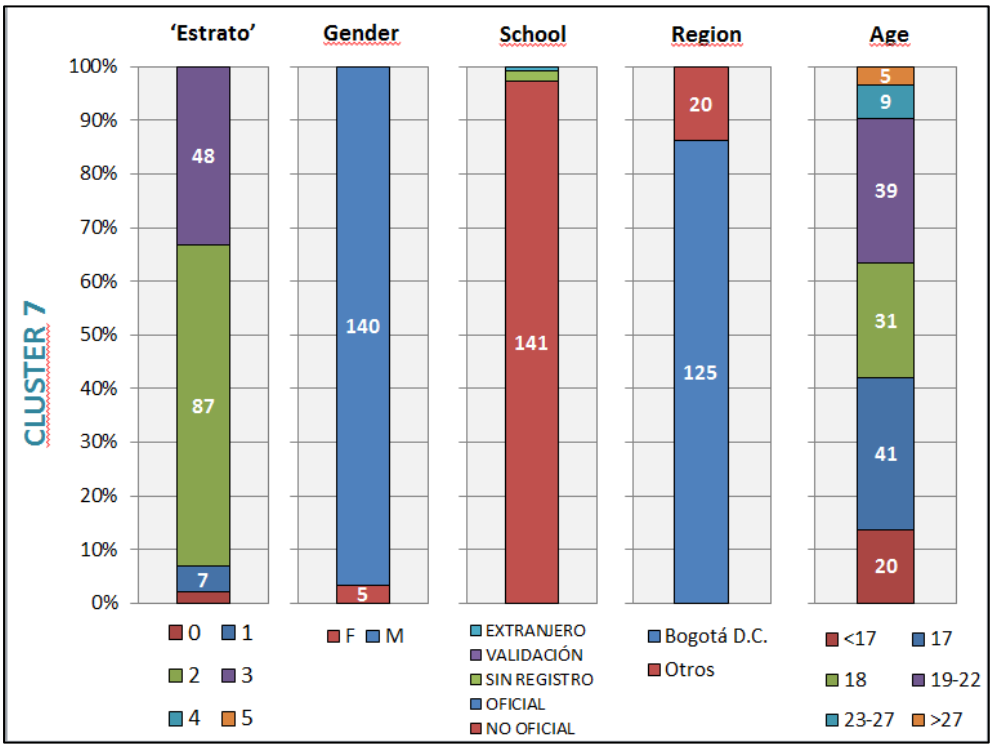


Figure 3-18: Cluster 7 – description of variables – Computer and Systems Engineering.



3.4.3 Both programs Clustering

Finally, K-Means was applied to the complete data set, using both programs at the same time. In this case, the number of clusters, K, was set to 6. It is interesting to see how the characteristics of the students differs between the programs, this can be appreciated in the clustering results where three clusters have majority of CE students and the other three of AE students, one is completely composed of AE. Given the characteristics of the population, most of the clusters have a majority of men from Bogotá; other characteristics have a better distribution and are explained here. The clusters are shown in Figure 3-19 to 3-24.

Cluster 0: AE students, mostly enrolled in the program as a second option, the rest are of third option. They are men from private schools in Bogotá and a medium-high 'estrato'. The students of this cluster didn't choose the program as the first option. 37% of these students lost the academic status due to low academic performance. Figure 3-19.

Cluster 1: Mostly women who studied in a private school, and low-medium 'estrato', AE students with a very small amount of CE, the enrollment option is mostly third-option. 22% of these students lost the academic status due to low academic performance. Figure 3-20.

Cluster 2: Public School students with low-medium 'estrato'. Mostly CE students. 25% of these students lost the academic status due to low academic performance. 86% of the students chose the program as first option. Figure 3-21.

Cluster 3: The largest group, composed of private school students. Although, the proportion is not too high, there is a high number, compared to other groups, of high 'estrato'. 98% of the students chose the program as first option. 26% of these students lost the academic status due to low academic performance. Figure 3-22.

Cluster 4: Public school. Low to medium 'estrato'. A majority of CE students, although it has a large participation of AE students. 31% of these students lost the academic status due to low academic performance. Figure 3-23.

Cluster 5: The students from this group come from out of the city, studied in a public school and most of them are AE students. Almost 85% are from 'estrato' 2 or lower. 51% of the students chose the program as second option and only 30% as first option. 40% of these students lost the academic status due to low academic performance. Figure 3-24.

Figure 3-19: Cluster 0 – description of variables – Both programs.

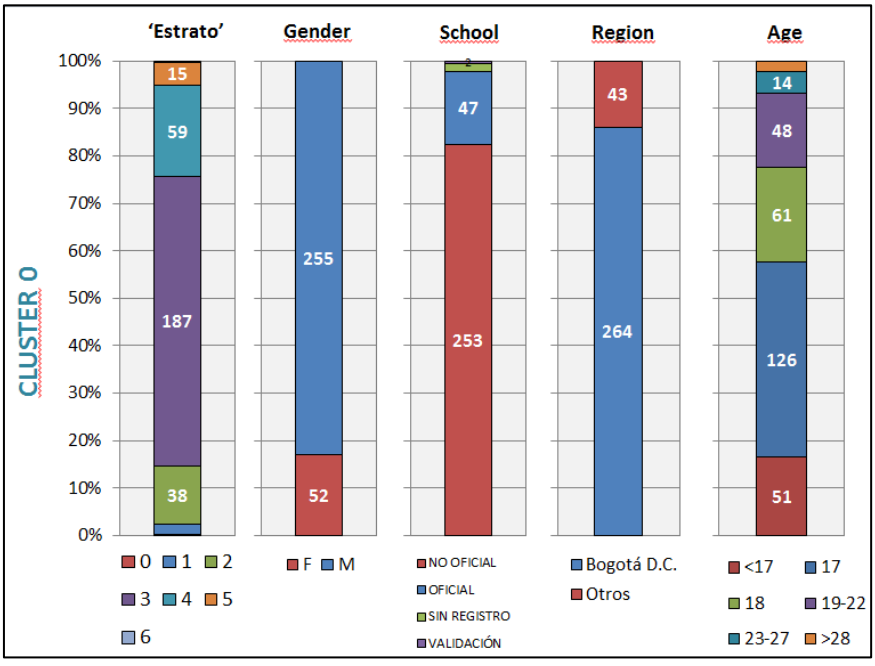


Figure 3-20: Cluster 1 – description of variables – Both programs.

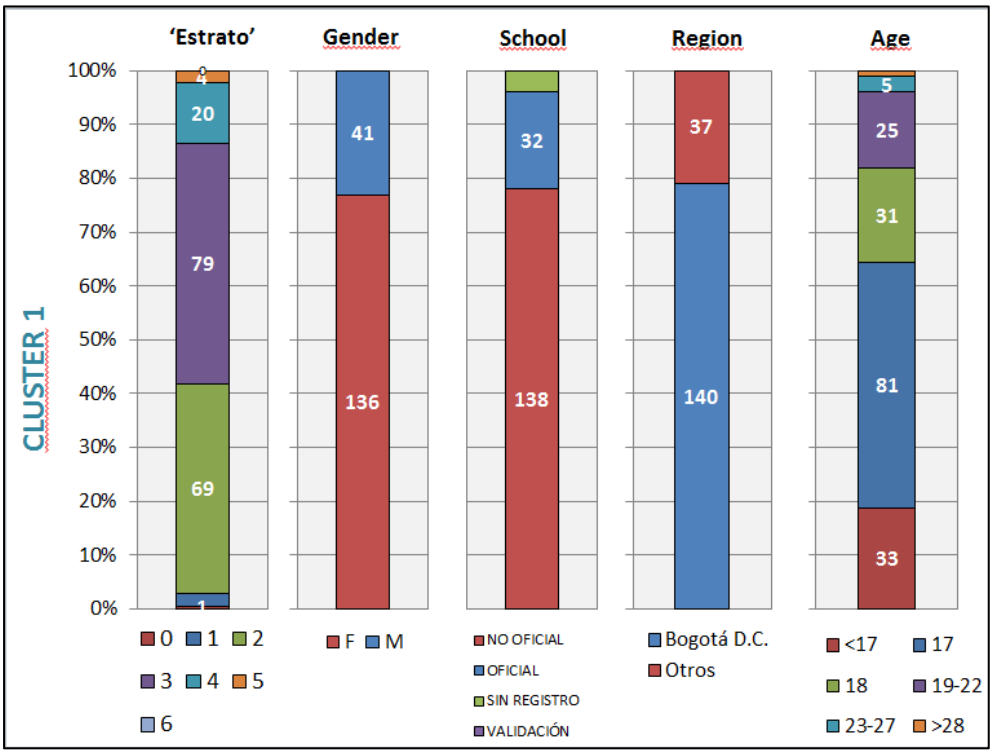


Figure 3-21: Cluster 2 – description of variables – Both programs.

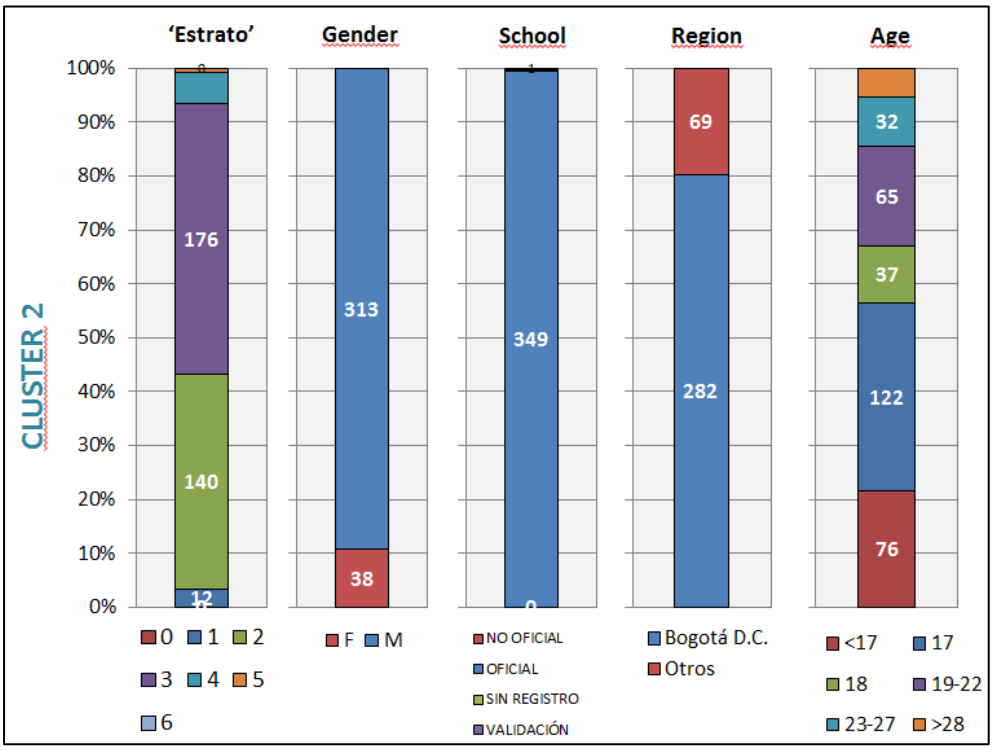


Figure 3-22: Cluster 3 – description of variables – Both programs.

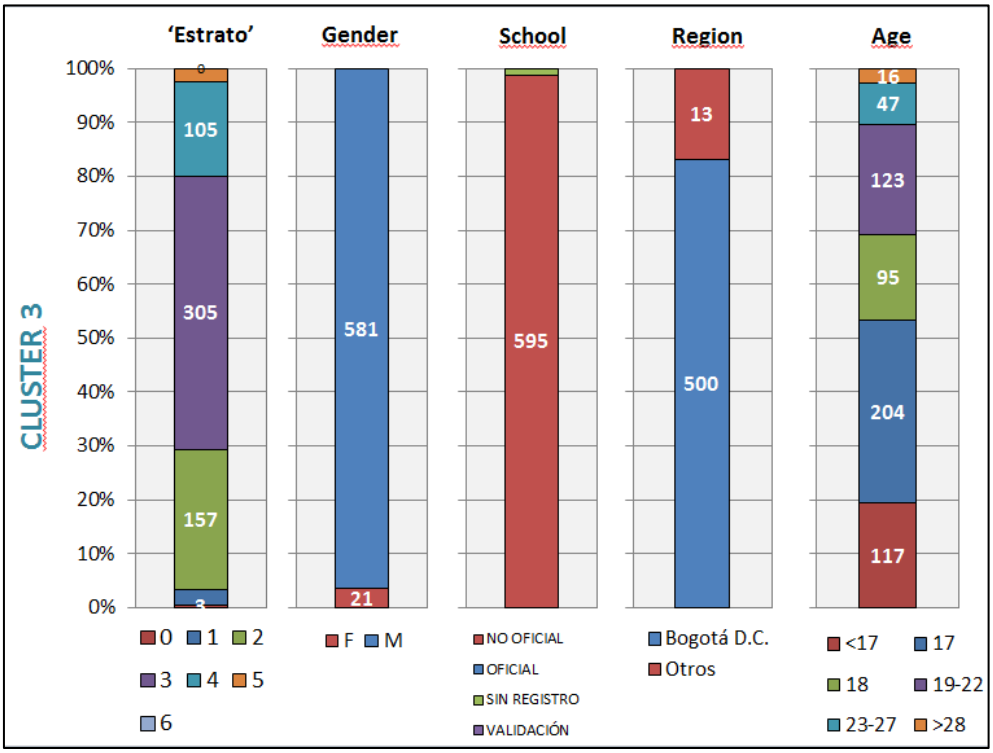


Figure 3-23: Cluster 4 – description of variables – Both programs.

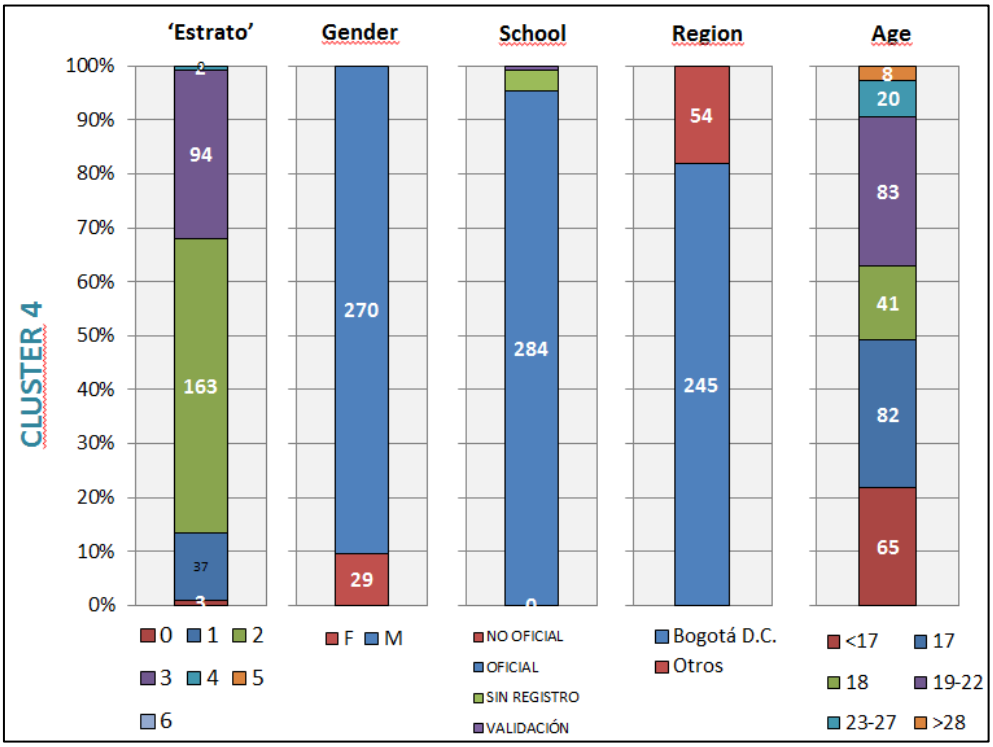
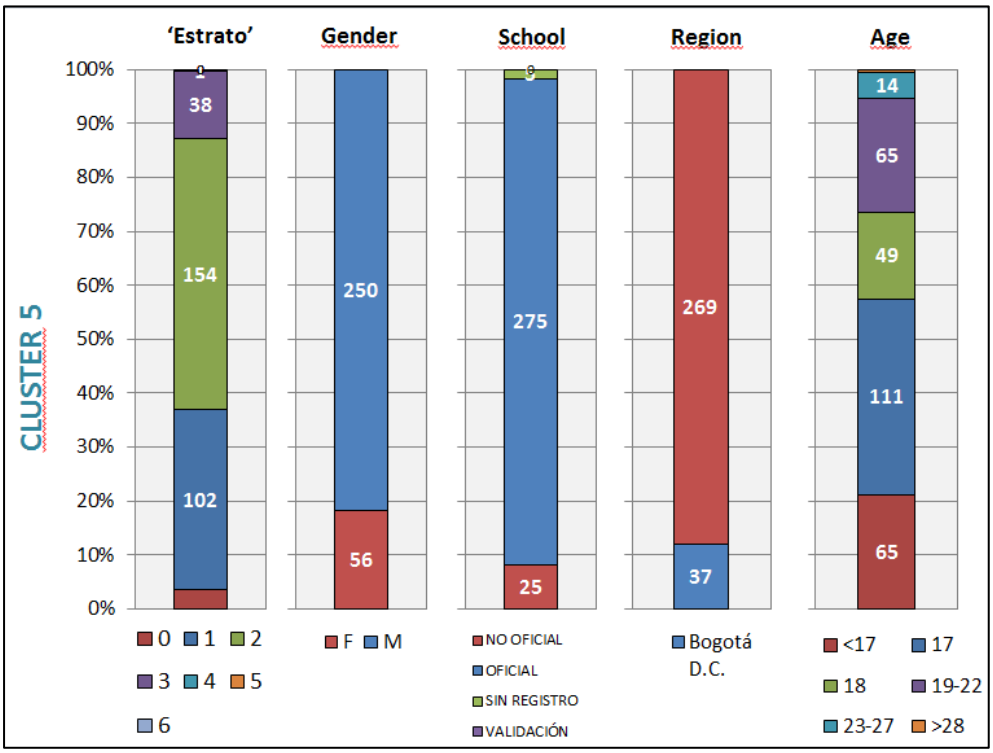
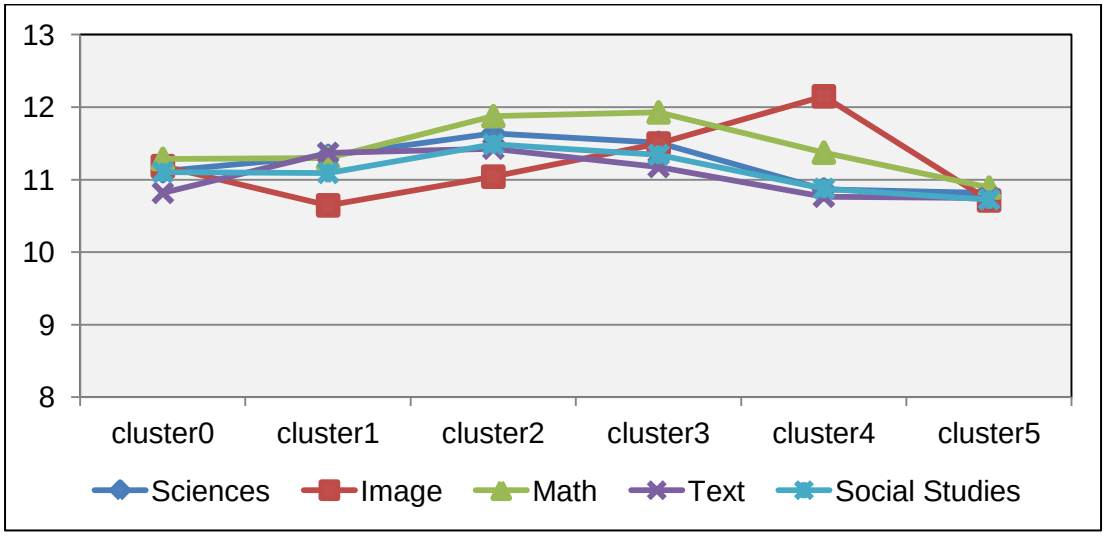


Figure 3-24: Cluster 5 – description of variables – Both programs.



Regarding the academic test results, there is no cluster with an overall performance above the others, the behavior varied depending on the component, especially on Image and Math since there were not major differences among the results of Sciences, Text analysis and Social Studies between the clusters. This can be seen in Figure 3-25.

Figure 3-25: K-Means – Admission test results per cluster and component.

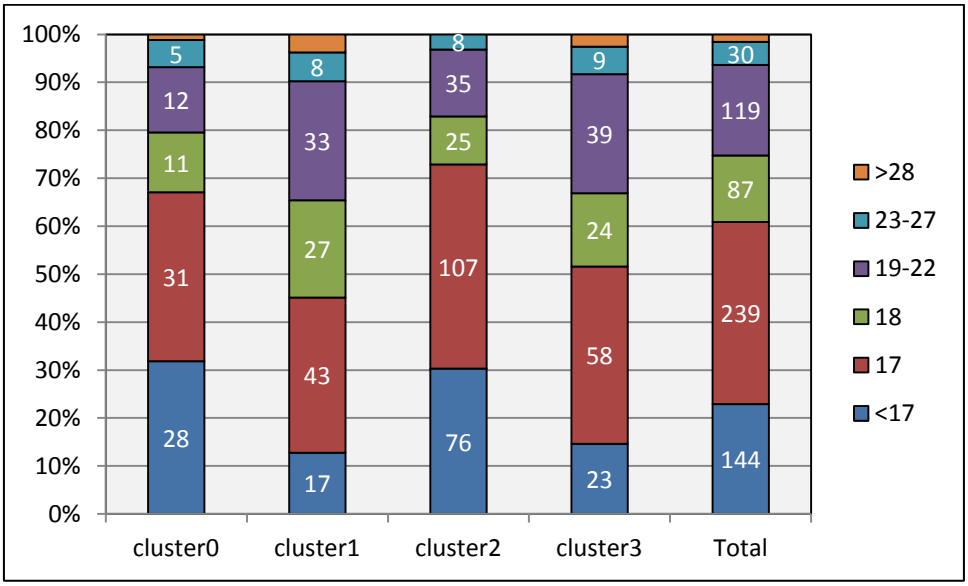


3.4.4 Clustering of students at fourth enrollment

A clustering was formed with the students who had a fourth enrollment in any of the two programs in order to compare the initial population to those who continue their studies after three academic periods. The process was repeated one more time and the number of clusters, K, was set to 4. It is important to notice that the reasons for a student not having an enrollment are not limited to low academic performance and it includes the voluntary retirements; however, it provides a characterization of the before mentioned students.

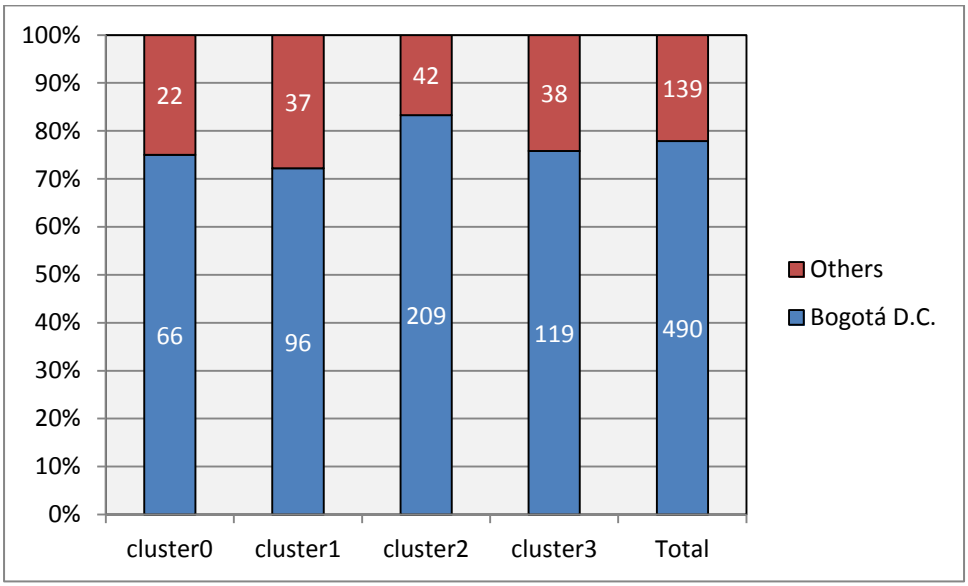
At this point, the clusters are more similar. For instance, there are not big differences regarding gender (between 84-88% of male students); however, as can be seen in Figure 3-26 most of the students who survived joined the program at a younger age. Clusters 0 and 2 are similar in that way.

Figure 3-26: K-Means – Clusters at 4th enrollment – Age.



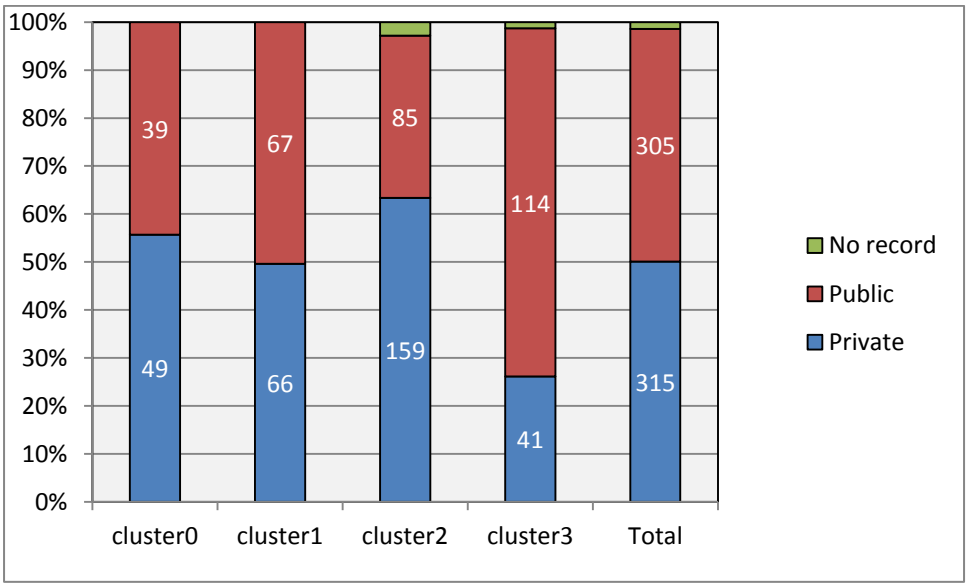
The region of origin used to differentiate one of the clusters, but at this academic period it is not that influential. Region of origin within clusters is more homogeneous, it varies between 72-83% of students coming from Bogotá. This is presented in Figure 3-27. Overall, the proportion of students coming from out of Bogotá decreased from 29% to 22%.

Figure 3-27: Clusters at 4th enrollment – Region of origin.



Something similar occurred to the type of school. The proportions are similar to those at the admission but only clusters 2 and 3 presents differences on the proportion compared to the average as can be seen in Figure 3-28.

Figure 3-28: Clusters at 4th enrollment – Type of school.



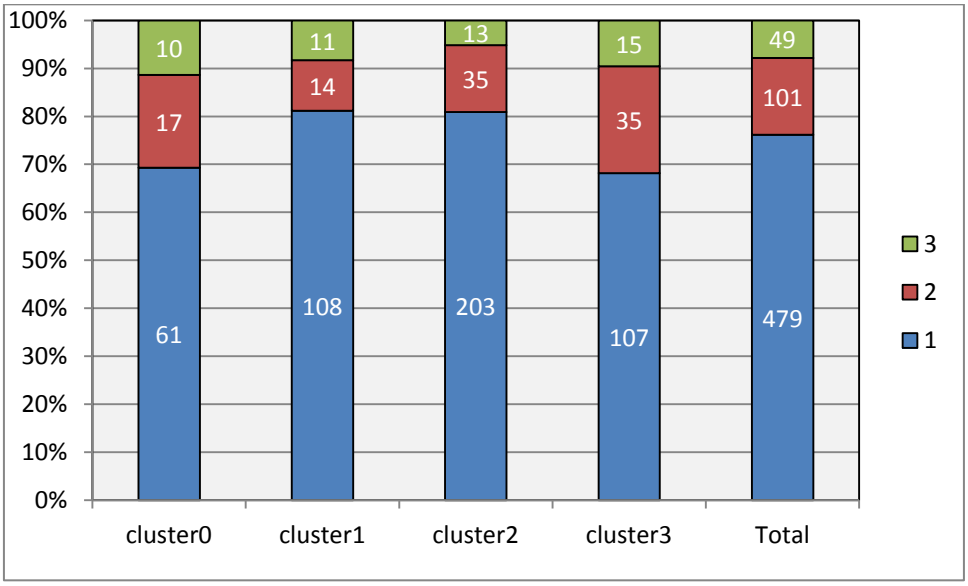
Regarding to the option for enrollment, the proportion of students who chose the program as their option increased, opposite to those who chose their program as a second or third

option as presented in Table 3-1. Figure 3-29 presents the clusters related to the option for enrollment attribute.

Table 3-1: Proportion of students according to their option for enrollment at admission and the 4th enrollment.

	1	2	3
m4	76%	16%	8%
Admission	61%	26%	14%

Figure 3-29: Clusters at 4th enrollment – Option for enrollment.



The averages of the admission test results were similar to those at the admission where sciences, text analysis and social studies presented similar values and the Math result being of the largest. One interesting difference is the decrease in the image analysis component results which doesn't stand out anymore, Text analysis average score also decreased compared to that from at the admission. Figure 3-30 presents these results.

Table 3-2: Average admission test results at admission and the 4th enrollment.

	Sciences	Image	Math	Text	Social Studies	Total
m4	11,41	11,38	11,78	11,14	11,26	696,61
Admission	11,26	11,28	11,53	11,06	11,15	680,19

Figure 3-30: K-Means – Admission test results per cluster and component at the 4th enrollment.

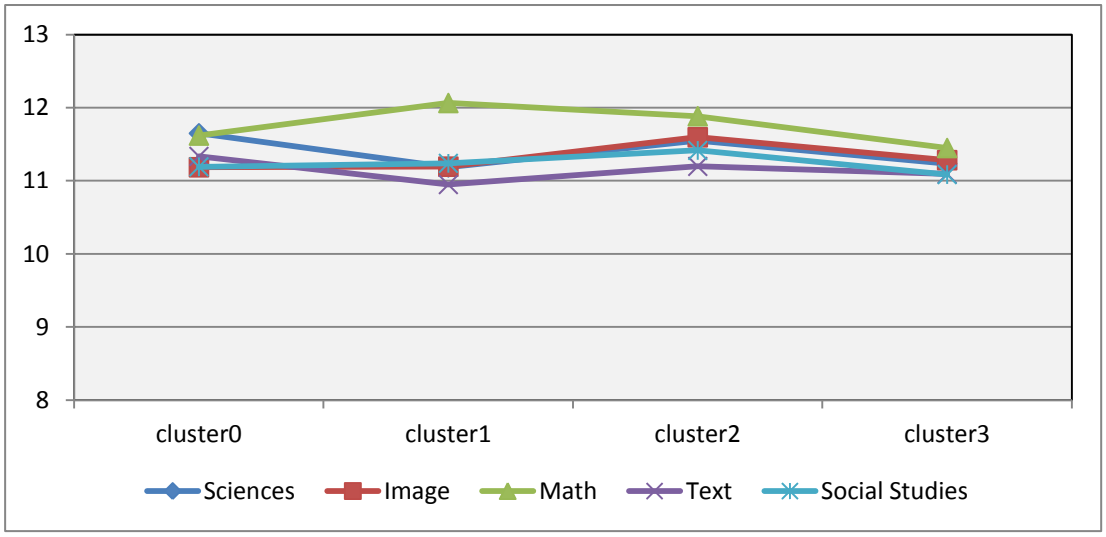


Table 3-4 shows the correspondence between the clustering at admissions using both programs and the current clustering which considers only students with four enrollments. It can be seen how the previous cluster2, cluster3, and cluster4 kept more students. These clusters have in common the larger presence of students whose option for enrollment in a particular program was their first. Cluster3 was composed of private schools students, from medium-high estrato it corresponds mainly to the new cluster2. The admissions' cluster2, of public schools and low-medium estrato, especially Computer and systems engineering students is evenly distributed within the new clusters.

Table 3-4: Average admission test results at admission and the 4th enrollment.

Clusters at Admission, Both programs	Cluster at 4th enrollment				not enrolled	Total	% of dropout
	cluster0	cluster1	cluster2	cluster3			
cluster0	11	8	25	12	251	307	82%
cluster1	10	9	15	11	132	177	75%
cluster2	23	32	43	43	210	351	60%
cluster3	31	53	120	28	370	602	61%
cluster4	5	16	41	43	194	299	65%
cluster5	8	15	7	20	256	306	84%

3.4.5 Clusters and loss academic status

The three initial models were also analyzed according to the loss of academic status to examine if there were any relationship between the cluster and the event of an academic blocking in the first semesters. Figures 3-31 to 3-33 present a visualization of these results.

Figure 3-31: Loss of Academic Status (BLQ.Acad) per cluster – Computer and systems engineering.

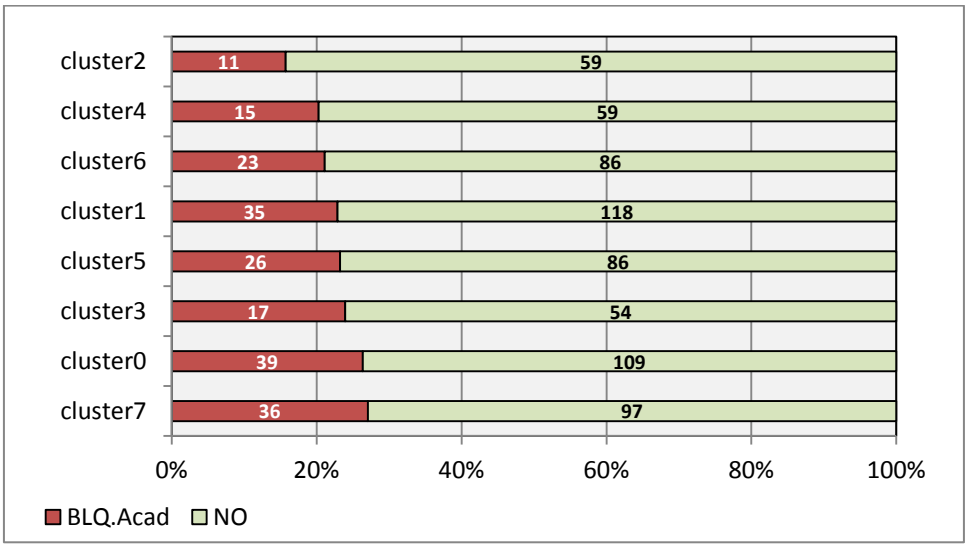


Figure 3-32: Loss of Academic Status (BLQ.Acad) per cluster – Agricultural engineering.

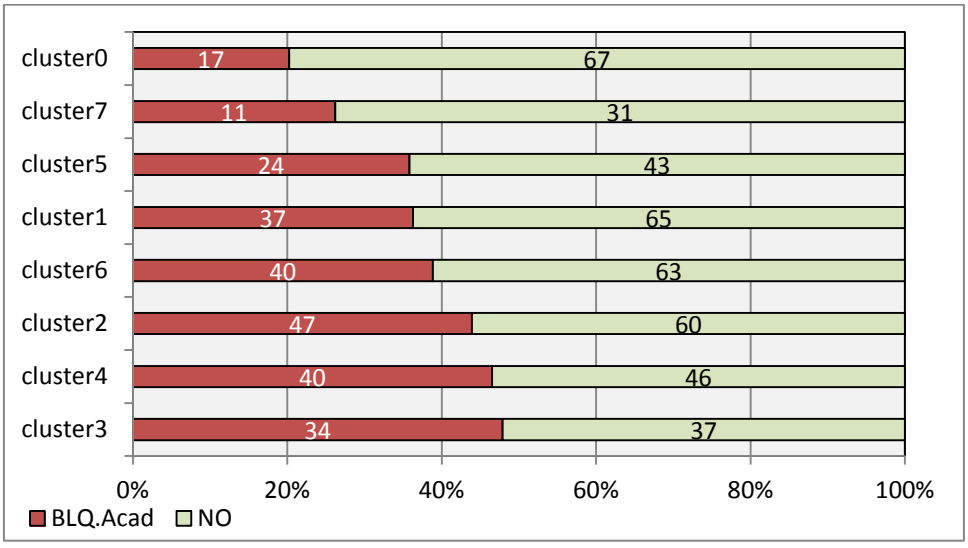
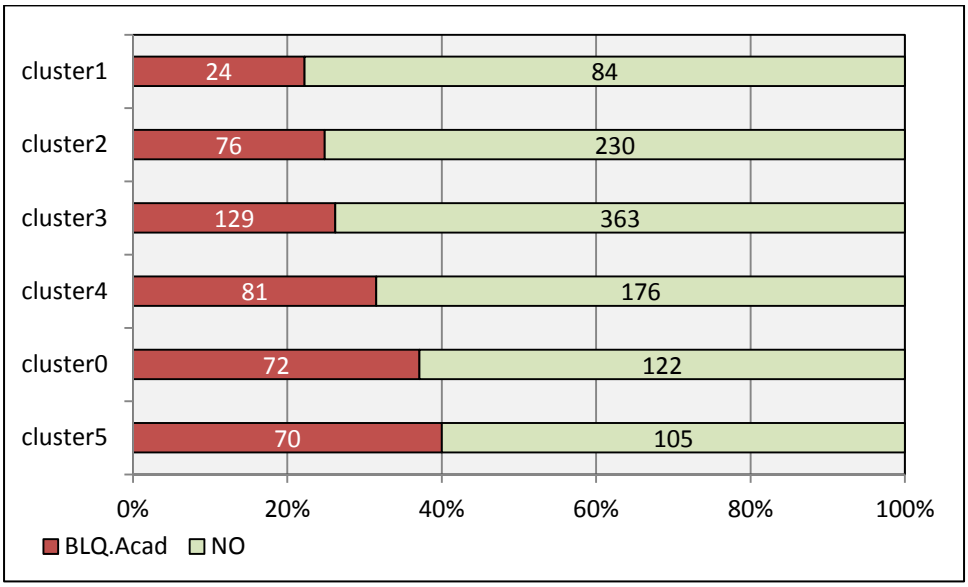


Figure 3-33: Loss of Academic Status (BLQ.Acad) per cluster – Both programs.



Further examination of the relationship among clusters and the loss of academic status included a chi-square independence test. The null hypothesis is that there are no differences between the different clusters regarding the loss of academic status. The relation was not significant when examining the CE data, but it was on the other two cases, AE and data with both programs as can be seen in Table 3-5.

Table 3-5: Chi-square independence test results.

Program	Chi-square	p-value
CE	4.7973	0.6847
AE	21.2498	**0.003417**
Both	23.6811	**0.00025**

As stated above, chi-square independence test results indicate the presence of an association between clusters and the loss of academic status in two of the clustering models, that is, that the proportion of loss of academic status is different among the clusters; however it is not known for which particular clusters the rate of loss of academic status differs. In order to evaluate which pair of clusters has different rates, the chi-square test is applied to every pair of clusters using the Bonferroni correction, in which the significance level is adjusted according to the number of comparisons, i.e. 0.05/number of comparisons; in the case of AE there are 8 clusters and 28 possible comparisons;

therefore, the p-value must be greater than 0.002 to be statistically significant. When data from both programs is used there are six clusters, 15 comparisons and an adjusted p-value of 0.003. Results are presented in Tables 3-6 and 3-7.

Table 3-6: Chi-Square independence test with Bonferroni correction. Clustering model using Agricultural Engineering

	cluster3	cluster4	cluster2	cluster6	cluster1	cluster5	cluster7	cluster0
cluster3	*							
cluster4	0,9910	*						
cluster2	0,7143	0,8310	*					
cluster6	0,3025	0,3597	0,5429	*				
cluster1	0,1706	0,2030	0,3239	0,8148	*			
cluster5	0,2067	0,2441	0,3681	0,8147	1,0000	*		
cluster7	0,0377	0,0441	0,0702	0,2096	0,3309	0,4024	*	
cluster0	0,0005	0,0005	0,0010	0,0096	0,0254	0,0506	0,5959	*

It is noticed in the results that loss of academic status rates are not significantly different between the clusters, and only cluster 0 presents differences to clusters 3, 4 and 2. A similar situation can be seen for the clustering model using data from both programs where cluster 5 is the only one that presents significant differences to clusters 1, 2, and 3.

Table 3-7: Chi-Square independence test with Bonferroni correction. Clustering model using both programs

	cluster5	cluster0	cluster4	cluster3	cluster2	cluster1
cluster5	*					
cluster0	0,5693	*				
cluster4	0,0695	0,2140	*			
cluster3	0,0006	0,0048	0,1254	*		
cluster2	0,0005	0,0034	0,0783	0,6637	*	
cluster1	0,0020	0,0077	0,0734	0,3881	0,5853	*

3.5 Summary

Clustering algorithms were applied to analyze a population of students of the Universidad Nacional de Colombia. It is interesting to see how the initial characteristics of a student in the University allow us to define profiles or characteristic groups. Further examination included a statistical significance test to examine the association of these clusters with the

event that a student loses his academic status. According to the results, there was not a significant association for the Computer and Systems Engineering program, but there was on the Agricultural Engineering.

A configuration of the full data set, including both programs was included. The number of clusters decreased to six and clusters presented an organization based on the programs, although this attribute was not considered in the clustering process.

The clustering was repeated by using only data of the students at their fourth enrollment. The number of clusters decreased as well as the variability within them. Most of the students at this particular point chose their program as the first option, All of the test scores went higher except for those of Image and Text analysis.

4. Predicting loss of academic status

The second phase of the Data Mining model is presented. The classifier uses two different algorithms, C4.5, a decision tree, and Naïve Bayes, a Bayesian classifier to predict the loss of academic status at different academic periods, and using different datasets.

4.1 Data preparation

In this phase, it was necessary to integrate all the datasets into a mining view to perform the classification task: admissions data; and those from the SIA, enrollment, grades, and loss of academic status. This process was done in several steps: first, two views were created to summarize the information of the grades and loss of academic status data. Both of them have the information per student per academic period; the marks view includes the number of credits, percentage of approved credits, and the average grade, both, in general, and also specific to the typology of the course, i.e. professional, foundation and optional electives. There is also information regarding the performance in two leveling courses, basic Math, and literacy. The loss of academic status view, on the other hand summarizes the types of blocking of the academic history, i.e. Academic, non-academic or others. A previous step filtered out the records where the blocking of the academic history was canceled due to administrative reasons, for instance, to modify the grades of the students.

The basic student information, i.e. socioeconomic, demographic, and previous academic data, coming from admissions and SIA was joined into one table. In addition, there are new attributes corresponding to the academic period when the student joins the University, one receives the values of A and B depending on the semester where the student was admitted to, i.e. A for the first semester of the year and B for the second; and the second takes the values of ‘EQUAL’ if the student joins the University in the same

period at which he applied and ‘Not.EQUAL in other case. This student data was then joined to the academic information described above. As a result there was a date set with the students and their grades and blockings per academic period. Only the records from students who had information in the enrollments and grades data sets were kept.

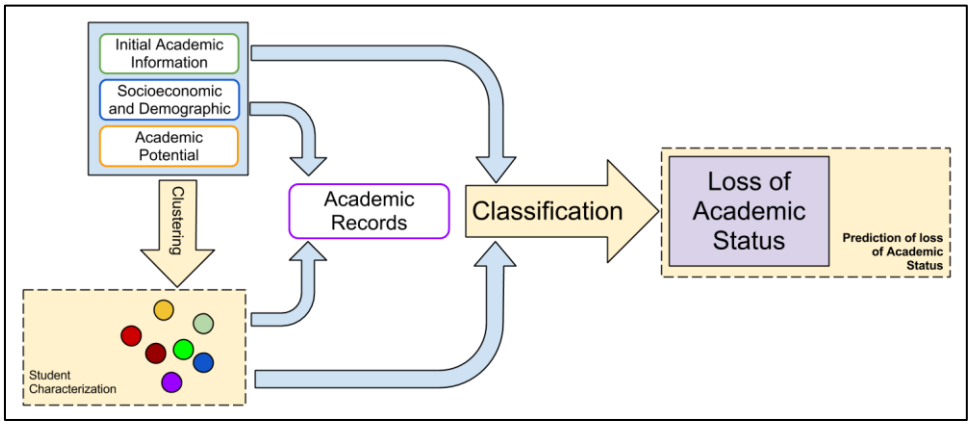
Finally, the mining view is a table with one record per student and the aggregates of the academic results in a given period with the label being the loss of academic status due to academic performance.

4.2 Classification model

The classification model uses the results of the student profiling from the previous phase in order to predict the students’ loss of academic status in the first semesters. For the classification models, two widely used techniques are used, Decision Trees and a Bayesian Classifier; these were selected based on the results of previous work and the need for a predictive model that is descriptive at the same time, so that a better understanding of the event of loss of academic status can be acquired. The implementation of the model was done using Rapid Miner with the operators NaiveBayes and Weka’s W-J48 respectively.

There are different configurations regarding the data used in the model, as can be seen in Figure 4-1. On the one hand, the data from the mining view is used with and without the academic records; on the other hand the clusters from the previous phase are used in replacement of the initial data.

Figure 4-1: Classification model.



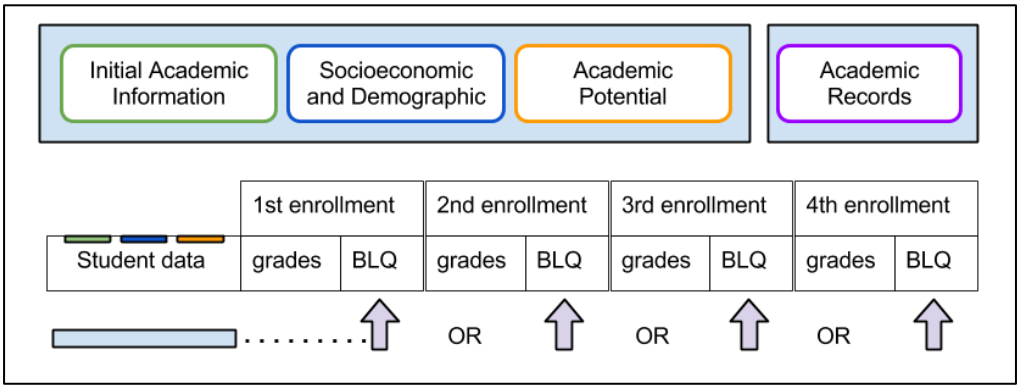
4.2.1 Classification Sub models

Different models were trained and tested, first a prediction of the loss of academic status regardless of the enrollment at which occurs; second, a prediction at a given enrollment is performed based on the initial information, the data gathered during the admission process; then, using the information known before the academic period starts, it includes the grades of the previous academic period when available. The different models are explained below.

- Predicting loss of academic status

The most general case, in which the interest is to predict the occurrence of the loss of academic status at any time in the first four academic periods based on the initial information, or entrance data. Figure 4-2 shows this configuration.

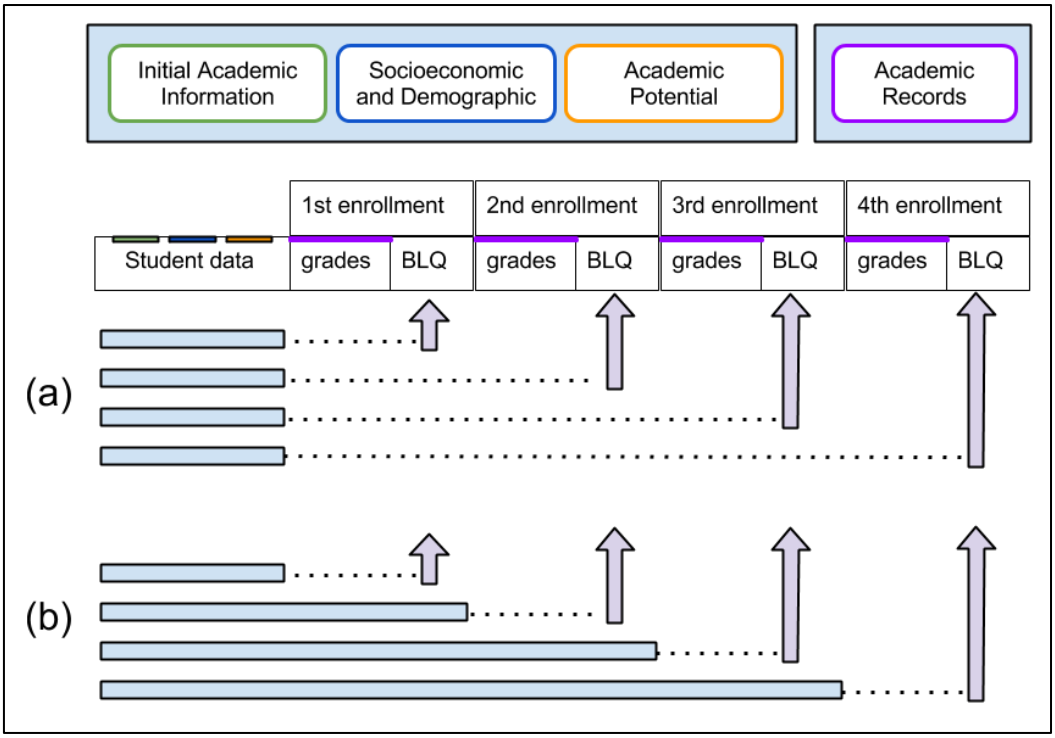
Figure 4-2: Experiments. Predicting loss of academic status. Arrows indicate the prediction; the boxes indicate the data used to train the model



- Predicting loss of academic status at a given semester

First, initial data is used to train a model to predict the loss of academic status at a particular academic period. The model is then complemented by adding academic information to the entrance data. The event of loss of academic status in a given period uses the academic information, grades and previous blocks, all available before the current period. For instance, to make a prediction in the third semester, data from the first two are used with the initial data. A visual representation of these configurations is presented in Figure 4-3.

Figure 4-3: Experiments. Predicting loss of academic status at a given semester. Arrows indicate the prediction; the boxes indicate the data used to train the model



The number of variables and records is presented in table 4-1

Table 4-1: Number of record and variables.

			# records		
To predict at enrollment	Prog	# variables	**Total**	BLQ	No.BLQ
1	AE	33	662	185	477
	CE		870	99	771
	Total		**1532**	**284**	**1248**
2	AE	62	431	48	383
	CE		679	64	615
	Total		**1110**	**112**	**998**
3	AE	91	298	18	280
	CE		527	32	495
	Total		**825**	**50**	**775**
4	AE	120	216	4	212
	CE		413	21	392
	Total		**629**	**25**	**604**

4.3 Experimental design and evaluation

For the experiments setup, 10-fold cross validation was used to train the model, in this, the data set is divided into ten equally distributed groups; the model was learned from nine of them, corresponding to the training set, then the model is evaluated on the tenth group, corresponding to a validation set. The process is repeated ten times so that every group is used for learning and testing.

The model is then applied to a previously unseen records corresponding to the 2012-03 academic period, the test set, in order to test the model in a more realistic way because all possible known data is used to train the model for the current semester. Additionally, two special characteristics were considered, the academic reform and the imbalance between the two classes:

- The academic reform

In 2009, the University went through an academic reform that had a clear impact in terms of loss of academic status, especially for those in their first-year of studies. A new model was trained by using only data from students who join the University in the academic period of 2009-01 or after, regardless of the period when they applied.

- Unbalanced dataset

To overcome the imbalance of the dataset, a cost-sensitive technique was included in the model; the metaCost algorithm [42] provides weights that represent the cost of classifying a record correctly or incorrectly, depending on the type of error that is more accepted. In this model, an error of classifying a student as No Risk when he is at risk is more critical than classifying a non-risk student as being at risk. Because of that, the following weights are considered in the model (Table 4-2).

The weights in (a) are the same for both types of errors; configuration in (b) and (c) consider the different acceptance regarding classification errors, (b) has a cost of misclassifying a student who is at risk as three times the error of misclassifying a non-at risk student as he were; finally, (c) presents a cost of misclassifying a student at risk but also considers a reward for classifying the BLQ class correctly.

Table 4-2: Weights in the cost-sensitive model.

		TRUE	
		No	BLQ
Predicted	No	0	1
	BLQ	1	0

(a)

	TRUE	
	No	BLQ
No	0	3
BLQ	1	0

(b)

	TRUE	
	No	BLQ
No	0	3
BLQ	1	-2

(c)

The performance of a classification model depends on the number of records of the validation set (academic period of 2012-03) correctly classified. These counts are commonly represented by a confusion matrix, a table that presents the number of records correctly and incorrectly classified.

In this case, the values correspond to:

- **Number of True Positives (TP)**: The records correctly classified in the positive class (BLQ.Acad).
- **Number of True Negatives (TN):** The records correctly classified in the negative class.
- **Number of False Positives (FP):** The records incorrectly classified in the positive class. i.e. BLQ.Acad was incorrectly predicted.
- **Number of False Negatives (FN):** The records incorrectly classified in the negative class.

These values have a relation with the cost sensitive model mentioned above. In this work there is a cost to False Negatives and a reward to True Positives. The values are also used to construct the following measures:

- **Precision (P):** the fraction of instances classified as positive (TP + FP) that are correctly classified (TP).
- **True Positive Rate or Sensitivity (TPR):** The fraction of the instances of the positive class that are correctly classified. TPR = TP/(TP + FN)
- **True Negative Rate or Specificity (TNR):** The fraction of the instances of the negative class that are correctly classified. TNR = TN/(TN + FP)
- **Balanced Accuracy:** The average of the TPR and TNR. Bal. Acc. = (TPR) + TNR) / 2.

Balanced accuracy is the arithmetic mean of the accuracy of both classes. It is used instead of the regular accuracy to prevent the bias that is caused by the unbalanced dataset. Consider a dataset where the positive class is only 10% of the instances, a classifier that labels every instance as the negative class will have an accuracy of 90% but a balanced accuracy of only 50%.

4.3.1 Analysis of results

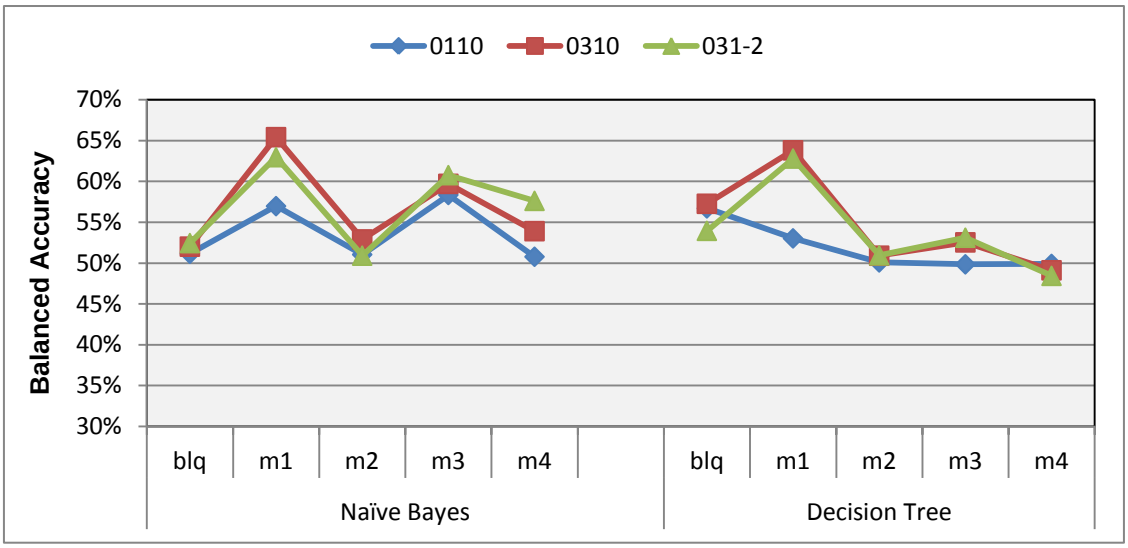
First experiments were intended to predict the loss of academic status in any of the first four academic periods using the initial data, but results were not satisfactory; the balanced accuracy ranged from 51 to 52% in the decision tree and between 54-57% in Naïve Bayes. After this, a selection of features was performed. The attributes were selected based on the Information Gain and Information Gain ratio values. The subset was formed based on the attributes for which any of the two measures mentioned above was in the top 10.

The resulting subset was: region, the six test scores, the type of school, age, type of application, estrato, gender, marital status, program, option for enrollment and PBM, a score to compute the tuition fee. A second subset, corresponding to the top 10 features was: age, the test scores of Math, Sciences, and total; marital status, type of application, estrato, program, option for enrollment and PBM, a score to compute the tuition fee. Using the second subset of attributes results increased to 60% using Naïve Bayes and 56% using the decision tree.

Although the performance increased, it was still too low, with a value close to 50%, this is similar to guessing.

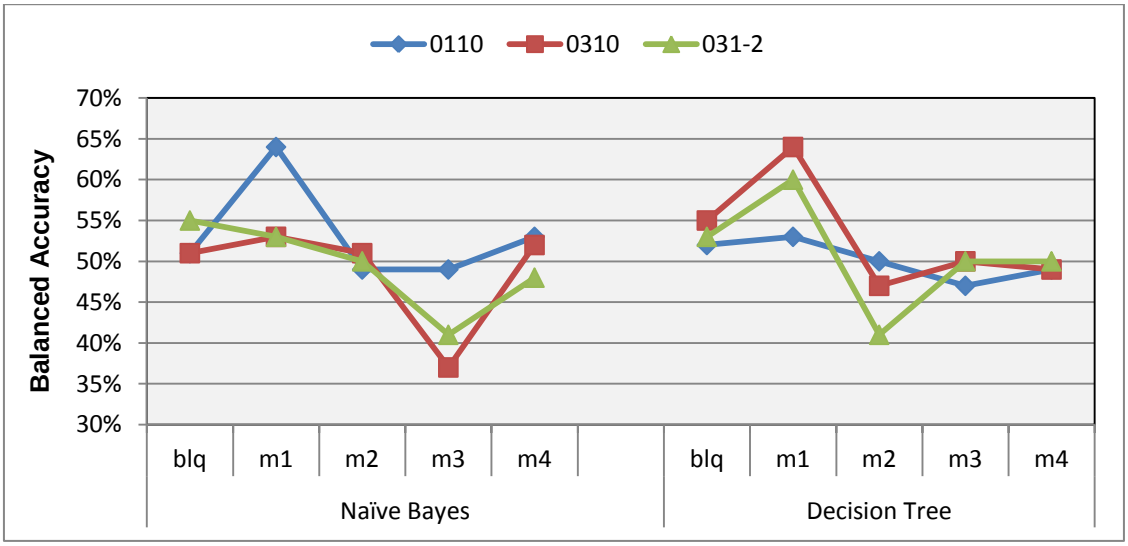
The next set of experiments was intended to predict, not only the event of loss of academic status, but also the semester in which occurs. First, only the initial data was used to predict an academic history blocking in the second, third and fourth enrollment. Results on the validation set are presented in Figure 4-4; it shows how initial data is particularly useful when predicting the event at first enrollment, after that, balanced accuracy tends to decrease over time for both algorithms.

Figure 4-4: Predicting loss of academic status at a given semester using entry data. Validation results



The algorithms were tested with the 2012-03 academic period (Figure 4-5). The results have a similar behavior when using the decision tree with an increase in performance in the prediction at the first enrollment and a posterior decrease. Naïve Bayes, on the other hand, showed an irregular behavior in the predictions at enrollment 3 and a shift in the performance of the different cost matrices used.

Figure 4-5: Predicting loss of academic status at a given semester using entry data. Test results (Academic period: 2012-03)



The next step was to include the academic records to the data. As it was described before, the academic data used are the grades and percentage of enrolled and approved credits in the previous academic period. Naïve Bayes had the best results, surpassing the 75% in balanced accuracy, up to 85% at the fourth enrollment on the test set. The decision tree didn't show much of an improvement, except for the second enrollment. This can be seen in Figures 4-6 and 4-7.

Figure 4-6: Predicting loss of academic status at a given semester using academic data

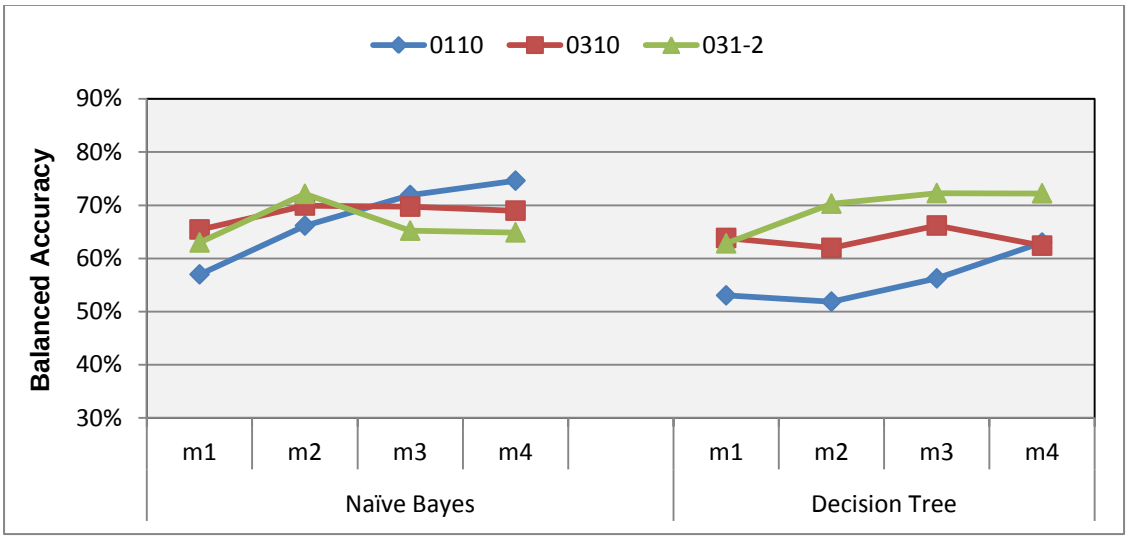
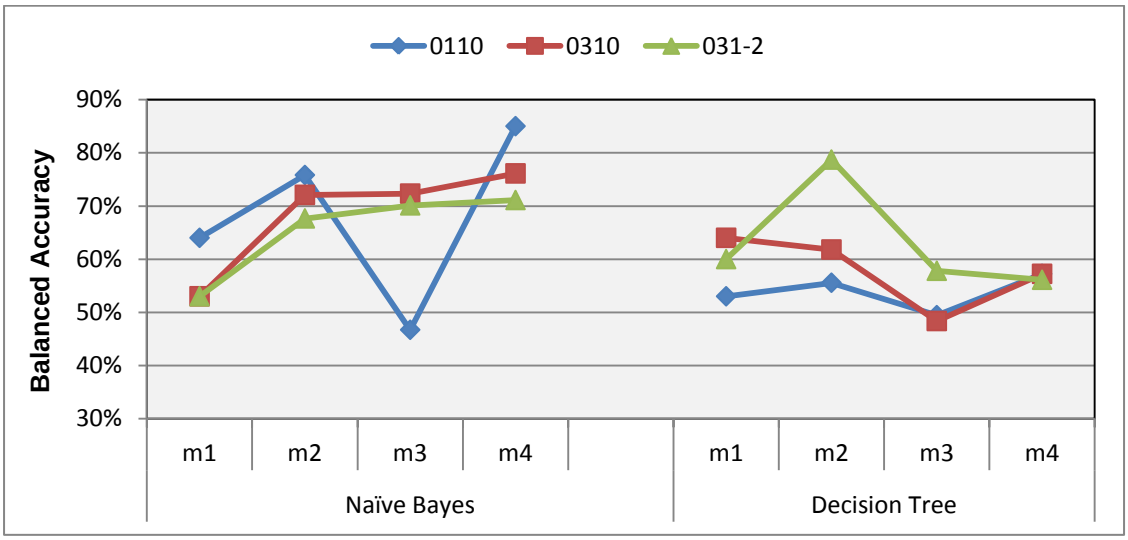


Figure 4-7: Predicting loss of academic status at a given semester using academic data. Test results (Academic period: 2012-03)



One of the set of attributes used above were the clustering results of the previous phase in order to evaluate if such characterization is able to summarize the variability present in the data. This new configuration didn't present a major change in the results especially on the training-validation set, although those were better when all the attributes were used. Figures 4-8 and 4-9 present a visual representation of the results using the training-validation and test sets respectively.

Figure 4-8: Predicting loss of academic status at a given semester using academic data. Test results (Academic period: 2012-03)

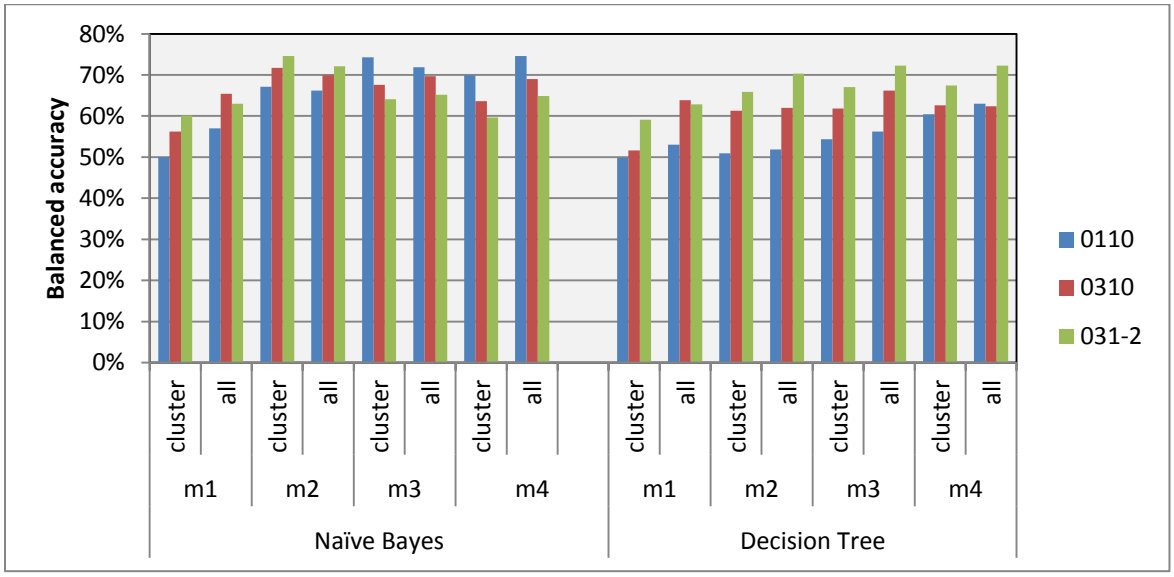
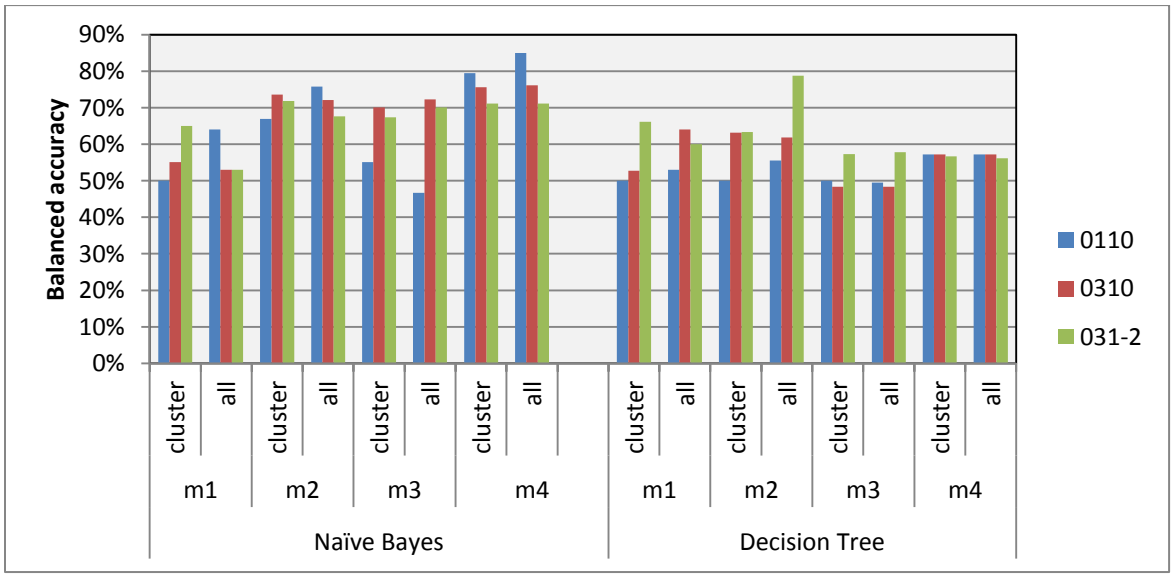
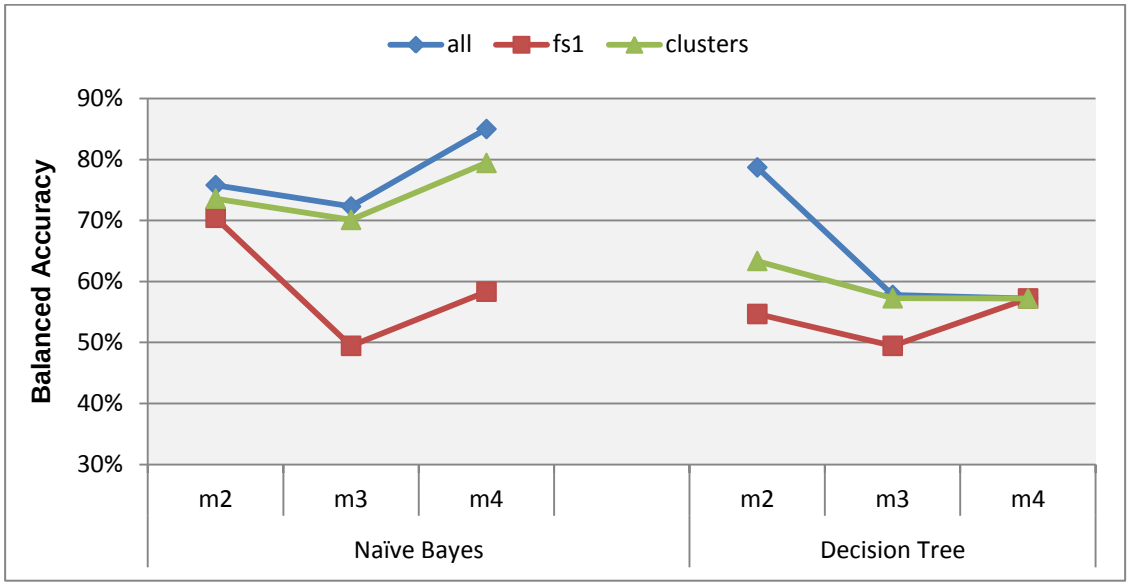


Figure 4-9: Predicting loss of academic status at a given semester using academic data. Test results (Academic period: 2012-03)



Considering only the best results regardless the cost-sensitive models, Naïve Bayes proved to be a superior classifier on this research, except from the prediction at the second enrollment when it reached 78.8% of balanced accuracy. Figure 4-10 presents the aforementioned results. The drop that both classifiers experience after the third semester can be explained by the increase in the imbalance. At this semester, only 4% of the records belong to the positive class. However, these results have to be taken carefully since the differences between training and test data are lower for the decision tree results, making them more consistent and making them more reliable when testing on new data.

Figure 4-10: Predicting loss of academic status at a given semester using different attributes (Best results)



Models were also learned by using only data after the academic reform, i.e. records of students who joined in 2009-01 or after. This approach intends to compare the results and see the influence of this reform in the behavior of the students in terms of low academic performance. According to the results, the exclusion of the records before the academic reform, the academic period of 2009-01, decreased the results in almost every configuration; the loss of accuracy can be explained on the loss of data. This can be seen in Table 4-3.

Table 4-3: Differences in results. data after the academic reform – all records.

		NB		DT	
meta-cost	prediction	Diff in training set	Diff in test set	Diff in training set	Diff in test set
[0 1;1 0]	blq	0%	0%	-3%	1%
	m1	-4%	-13%	2%	1%
	m2	-6%	-2%	0%	-6%
	m3	-3%	34%	-6%	0%
	m4	-23%	-23%	-13%	-7%
[0 3;1 0]	blq	1%	0%	-6%	-5%
	m1	-9%	12%	-7%	-7%
	m2	-1%	3%	-3%	-12%
	m3	0%	8%	-14%	5%
	m4	4%	-2%	-13%	-8%
[0 3;1 -2]	blq	0%	-4%	-4%	-3%
	m1	-3%	12%	-3%	-4%
	m2	-4%	7%	-8%	-5%
	m3	-2%	-1%	-19%	-7%
	m4	5%	18%	-24%	-1%

4.4 Relevant Features

A systematic analysis was conducted to identify relevant factors related to the loss of academic status due to low academic performance according to the learned classification models. The interpretation depends on the selected model. The Bayesian classifier is interpreted based on the probabilities of the factors, or variables, and those with a higher probability are highlighted. When the attribute is continuous, a visual comparison of the density function is also taken into consideration. On a Decision Tree on the other hand, two approaches were followed: first, the features that are on the root of a tree are considered as more relevant; and second, the branches with more examples are also considered.

Results are described below.

- **Admission test results:** the academic potential shows an expected behavior considering the population under study, programs of the Faculty of Engineering. The components of Math and Science, along with the total score and the classification level for basic Math are the most relevant, poor performances are

more related to loss of academic status. On the other hand, for the prediction at fourth enrollment, high scores in the social sciences component are more related to this loss.

- **Age at enrollment:** a first thought could lead us to think that younger students are at more risk, and they are, in absolute terms; however, the age rank of 23-28 presents a higher risk.
- **Socioeconomic Status (Estrato):** there are to variables used to measure this, PBM and estrato, according to the results the estrato was more telling than the PBM.
- **Option for enrollment:** The models show that this feature is relevant when the loss of academic status is predicted at the first enrollment but not so much when the prediction is at a later enrollment. A further evaluation shows that there is a relationship between the option for enrollment and the loss of academic status at first enrollment and that this relationship disappears at a later enrollment.
- **Grades:** the grade average and the percentage of the approved credits are relevant features and there is a difference according to the typology of the courses, the performance at professional subjects is most telling than the performance at foundational subjects and that the absence of elective courses is more related to the loss of academic status. It is also important to notice that the grades become more relevant as time progresses, when trying to predict the loss of academic status at a later enrollment; under this scenario, these features gain even more importance than socioeconomic and demographic data.

4.5 Summary

Two algorithms, Naïve Bayes and a decision tree, were used to create classification models to predict the loss of academic status due to low of academic performance. Several models were learned to test the configuration. It includes the prediction of the academic history block at any time of the first two years, at a specific enrollment using only entry data and then including the academic information, i.e. grades and credits enrolled. Further experimentation was included in order to use different sets of attributes and the use of records after the academic reform. The models were tested with previously unseen records corresponding to the 2012-03 academic period.

Naïve Bayes results were better on the test set; however, there are differences between training and test data. The decision trees results were more consistent regarding that subject making it more reliable when testing on new data.

The classification results showed similar values to works reported in the literature using similar datasets. The accuracy of the classifiers improved when academic data was added; however, adding more academic data doesn't necessarily improve the classifier. It is important to notice, that early dropout researches suggest that retention is influenced by different factors involving the integration of the student to the University making the entry data insufficient for making predictions.

5. Conclusions and future work

5.1 Conclusions

Dropout and academic performance are topics that have been researched for a while. However, the use of Information Systems to keep the records of the students, and other sources of information, such as Learning Management Systems or mobile technology have led to different approaches driven by the data. It is important to notice the rise of three fields that use Information technologies and a data-driven approach to empower the actors involved in the Educational sector.

The application of clustering algorithms to analyze a population of students in the Universidad Nacional de Colombia allows identifying similar characteristics between groups. It is interesting to see how the initial characteristics of a student in the University allow us to define some profiles or characteristic groups. Further examination, included a statistical significance test to examine the association of these clusters with the event that a student losing his academic status. According to the results, there was not a significant association for the Computer and Systems Engineering program, but there was on the Agricultural Engineering and the two programs clustering.

It is discussed that the initial models using only data from the admission process might not be sufficient for a prediction beyond first enrollment. Therefore, a model that evolves through time is needed. The model presented in this work adds the academic information of previous academic previous in order to improve the model. Previous research by the Ministry of Education [23] focuses mostly on the students' initial characteristics.

The classifications results presented in this research are similar to those reported in the literature in problems that used similar datasets, but those were mostly reported on validation sets, data that were used to learn the model.

Bayes classifier performance improved when academic data from the first enrollment were added; however the performance decreased after the addition of the academic data of the second enrollment. This may be caused by the assumption of independence required by the algorithm. Naïve Bayes results were better on the test set; however, there are differences between training and test data. The decision trees results were more consistent regarding that subject making it more reliable when testing on new data.

The data used is gathered on each academic period, but the causes of low academic performance occur on a day to day basis. This leads to think that new, and possibly, non-traditional ways, for collecting information are needed.

5.2 Future work

This work is a starting point of Educational Data Mining research at the University and can be further developed in various ways.

To use different techniques to learn the classification models, or a combination of classifiers, i.e. ensemble classifiers or meta-classifiers, to improve the performance results.

To use more data: the research used only two Engineering programs for the model; however, the University Campus has 49 of them from 11 faculties. The inclusion of other programs in the model can bring a different perspective and allow the University to gain a better understanding of the dropout event.

To use more data, in terms of variety of data. There are good amounts of data related to the academic performance of the student, but there are few that consider other aspects of the University life. The model can benefit from the integration with other data sources, such as: ICFES, which has different background data, information of the family, academic and socio economic. The university is also implementing new initiatives, e.g. the Welfare Information System and the COMFIE program, which can collect data regarding other types of interactions between the student and the University. Besides the use of new data sources, it is important to consider new ways to gather information from the students and their interaction, taking advantage of social media and other communication tools.

This work focused on the loss of academic status due to low performance; however, the academic performance can also be studied at a different level, perhaps at the course level. The classification model could also include the non-academic loss of student status, or a new model could be built to reflect this situation.

References

[1] T. Hey and K. Tolle, The fourth paradigm data-intensive scientific discovery. Redmond, Wash.: Microsoft Research, 2009.

[2] S. Kelling, W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker, “Data-intensive Science: A New Paradigm for Biodiversity Studies,” BioScience, vol. 59, no. 7, pp. 613–620, Jul. 2009.

[3] R. Ferguson, “The State of Learning Analytics in 2012: A Review and Future Challenges,” Knowledge Media Institute, The Open University, UK, Technical Report KMI-12-01, 2012.

[4] Educational Data Mining Society, Available: http://www.educationaldatamining.org/

[5] R. S. J. d Baker and K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, Oct. 2009.

[6] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 40, no. 6, pp. 601 –618, Nov. 2010.

[7] P. J. Goldstein, and R. N. Katz, "Academic Analytics: The Uses of Management Information and Technology in Higher Education," EDUCAUSE Center for Applied Research (ECAR) Study, vol. 8. 2005

[8] J. Campbell and D. Oblinger, “Academic analytics,” EDUCAUSE Center for Applied Research, 2007.

[9] Available: http://www.solaresearch.org/mission/about/

[10] G. Siemens, “Structure and Logic of the Learning Analytics Field.” EdLab Seminars at Columbia University. Available at: http://www.learninganalytics.net/?p=187 (video) and http://www.slideshare.net/gsiemens/columbia-tc (slides)

[11] R. Baker, E. Duval, J. Stamper, D. Wiley, and S. Buckingham Shum, “Educational Data Mining meets Learning Analytics: Plenary Panel,” presented at the 2nd International Conference on Learning Analytics & Knowledge, Vancouver, BC, Canada, 2012.

[12] G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining: towards communication and collaboration,” in Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, New York, NY, USA, 2012, pp. 252–254.

[13] R. D. Reason, “Student Variables that Predict Retention: Recent Research and New Developments,” Journal of Student Affairs Research and Practice, vol. 46, no. 3, Jan. 2009.

[14] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “Preventing student dropout in distance learning systems using machine learning techniques,” in Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst., Oxford, U.K., 2003, pp. 3–5.

[15] J. F. Superby, J. P. Vandamme, and N. Meskens, “Determination of factors influencing the achievement of the first-year university students using data mining methods,” in Workshop on Educational Data Mining, Boston, USA, 2006, pp. 37–44.

[16] P. Parmentier, “La réussite des études universitaires: facteurs structurels es processuels de la performance académique en première année en médicine.,” PhD, Catholic University of Louvain, 1994.

[17] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, “Predicting Students Drop Out: A Case Study,” in Proceedings of the 2nd International Conference on Educational Data Mining, Cordoba, Spain, 2009, vol. 9, pp. 41–50.

[18] J. Bayer, H. Bydzovská, J. Géryk, T. Obsıvac, and L. Popelınskỳ, “Predicting drop-out from social behaviour of students,” in Proceedings of the 5th International Conference on Educational Data Mining-EDM 2012, Chania, Greece, 2012, pp. 103–109.

[19] S. Kotsiantis, “Educational data mining: a case study for predicting dropout-prone students,” International Journal of Knowledge Engineering and Soft Data Paradigms, vol. 1, no. 2, pp. 101–111, Jan. 2009.

[20] C. Marquez-Vera, C. Romero, and S. Ventura, “Predicting School Failure Using Data Mining,” in Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, 2011, pp. 271–276.

[21] S. R. Timarán Pereira, “Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos,” in Octava Conferencia Iberoamericana En Sistemas, Cibernética E Informática, International Institute Of Informatics And Systemics, Orlando, Florida, USA, 2009, pp. 146–150.

[22] L. L. Pinzón Cadena, “Aplicando minería de datos al marketing educativo,” NOTAS D MARKETING, vol. 1, pp. 45–61, Jun-2011.

[23] E. Castaño Vélez, D. Durán Muriel, J. Franco Gallego, S. Gallón Gómez, K. Gómez Portilla, C. Guzmán Ruiz, and J. Vásquez Velásquez, Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención, 1a ed. Bogotá: Ministerio de Educación Nacional, 2009.

[24] M. Pinto Segura, Cuestión de supervivencia graduación, deserción y rezago en la Universidad Nacional de Colombia. Bogotá: Universidad Nacional de Colombia, 2007.

[25] A. Rodríguez Rodríguez, Permanencia estudiantil en los postgrados de la Universidad Nacional de Colombia. Bogotá: Universidad Nacional de Colombia, 2010.

[26] C.-F. Tsai, C.-T. Tsai, C.-S. Hung, and P.-S. Hwang, “Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation,” AJET, vol. 27, no. 3, pp. 481–498, 2011.

[27] V. P. Bresfelean, M. Bresfelean, N. Ghisoiu, and C.-A. Comes, “Determining students’ academic failure profile founded on data mining methods,” in 30th International Conference on Information Technology Interfaces, 2008. ITI 2008, 2008, pp. 317–322.

[28] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, “Data mining algorithms to classify students,” in Proceedings of the 1st International Conference on Educational Data Mining, Montreal, Quebec, Canada, 2008, pp. 20–21.

[29] L. Talavera and E. Gaudioso, “Mining student data to characterize similar behavior groups in unstructured collaboration spaces,” in Workshop on Artificial Intelligence in CSCL. 16th European Conference on Artificial Intelligence, Valencia, Spain, 2004, pp. 17–23.

[30] M. I. Lopez, J. M. Luna, C. Romero, and S. Ventura, “Classification via clustering for predicting final marks based on student participation in forums,” in Proceedings of the 5th International Conference on Educational Data Mining, Chania, Greece, 2012, pp. 148–151.

[31] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” AI Magazine, vol. 17, no. 3, p. 37, Mar. 1996.

[32] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to data mining. Boston: Pearson Addison Wesley, 2005.

[33] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, “Top 10 algorithms in data mining,” Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, Dec. 2007.

[34] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” SIGMOD Rec., vol. 22, no. 2, pp. 207–216, Jun. 1993.

[35] M. J. Zaki, “Parallel and distributed association mining: a survey,” IEEE Concurrency, vol. 7, no. 4, pp. 14–25, 1999.

[36] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. 20th Int. Conf. Very Large Data Bases, 1994, vol. 1215, pp. 487–499.

[37] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo, “Finding interesting rules from large sets of discovered association rules,” in Proceedings of the third international conference on Information and knowledge management, New York, NY, USA, 1994, pp. 401–407.

[38] P.-N. Tan, V. Kumar, and J. Srivastava, “Selecting the right interestingness measure for association patterns,” in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, 2002, pp. 32–41.

[39] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[40] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, “YALE: rapid prototyping for complex data mining tasks,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, 2006, pp. 935–940.

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18, Nov. 2009.

[42] P. Domingos, “MetaCost: A General Method for Making Classifiers Cost-Sensitive,” in In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, 1999, pp. 155–164.