Determination of factors influencing the achievement of the first-year university students using data mining methods

J.F. Superby
Production and Operations Management Department,
Catholic University of Mons,
chaussée de Binche 151, 7000 Mons,
Belgium
superby@fucam.ac.be

J.-P. Vandamme
N. Meskens
Production and Operations Management Department,
Catholic University of Mons,
chaussée de Binche 151, 7000 Mons,
Belgium
{vandamme, meskens}@fucam.ac.be

Abstract. Academic failure among first-year university students has long fuelled a large number of debates. Many educational psychologists have tried to understand and then explain it. Many statisticians have tried to foresee it.

Our research aims to classify, as early in the academic year as possible, students into three groups: the 'low-risk' students, who have a high probability of succeeding, the 'medium-risk' students, who may succeed thanks to the measures taken by the university, and the 'high-risk' students, who have a high probability of failing (or dropping out).

This article describes our methodology and provides the most significant variables correlated to academic success among all the questions asked to 533 first-year university students during the month of November of academic year 2003-04. Finally, it presents the results of the application of discriminant analysis, neural networks, random forests and decision trees aimed at predicting those students' academic success.

Keywords: Decision trees – Random forests - Neural networks – Discriminant analysis – Education – Prediction

INTRODUCTION

Although the Belgium Constitution divides Belgium into three communities, Flemish, French-speaking and German speaking, the structure of higher education is similar in all three communities. A Bachelor's Degree is awarded after two or three years of study. A Master's Degree is awarded after a further two or three years of study. A PhD can be acquired after a further two or three years of personal research and the production of a thesis, or theses, which the candidate must defend in public.

There are nine universities in the French Community of Belgium that are grouped together in the form of university academies: Académie Universitaire Louvain, Académie Universitaire Wallonie-Bruxelles, Académie Universitaire Wallonie-Europe. These universities are financed by the government and the received amount is calculated on the basis of the number of students registered in each university. The academic year in Belgium begins in September, consists of three periods of four months and three sessions of exams. The first of these sessions takes place in January and the two others in June and September.

Universities are faced more and more frequently with saturated and highly competitive markets. In this perspective, the student, i.e. the university's essential resource, is at the centre of the university's preoccupations and initiatives. The University must thus take its students' needs into account more than ever before: Who are they? How to attract them? How to keep them as long as possible in the university system without reducing the quality of their studies? What is the cost-benefit ratio of a student who succeeds in his/her first year? What is the cost-benefit ratio for a student who succeeds thanks to a number of measures taken by the university? etc.

After analyzing the results of first-year students in the Belgian French-speaking universities, we found that about 60% of theses students fail or drop out. Droesbeke et al. (2001) observed stability in terms of success rates, repeat rates and drop-out rates over a period of 10 years. They show that the success rate of secondary school-leaving first-year students is close to 41%, while their repeat rate is approximately 26% and their drop-out rate is 33%. Given these figures, appropriate action should be taken to reduce the worrying economic, social and human cost involved in such a high level of failure in the first year at university. For many years, most Belgian universities have provided supplementary activities to the normal first-year program (computer-assisted

teaching, tutorials, etc.) in order to fill in the gaps for 'failing' students, particularly after the first examination period of the year, in Belgium it corresponds to the January session.

Our main objective is to classify students into three groups: 'low-risk' students, with a high probability of succeeding; 'medium-risk' students, who may succeed thanks to the measures taken by the university; and 'high-risk' students, who have a high probability of failing (or dropping out). We will thus need to create a database in which every student is described according to a range of criteria or characteristics such as their age, their parents' level of education, their perception of the university environment, etc. To determine the factors to be taken into account we will use a model adapted from that of Philippe Parmentier (1994). In other words the idea is to determine if it is possible to predict a decision variable using the explanatory variables which we retained in the model.

At the beginning of the academic year 2003-2004 we distributed a questionnaire in three Belgian universities. Questionnaires were again distributed last year. As a result we will be able to establish correspondences and divergences between the predictive models obtained in different institutions.

Our sample contains 533 students registered in Belgian universities, of whom 151 have passed an entry examination. Each student is described using 375 variables included in the questionnaire. A selection of variables are thus necessary before any statistical or mathematical treatment. The decision variable used for the construction of our models is a variable of three modalities, built a posteriori, grouping students according to their academic performance.

Classification into three groups has progressively less interest as the academic year advances. It is not beneficial to wait until April or even February to guide students who have a real need for support. Having collected sufficient and varied information via the questionnaires, our objective will be to establish a statistical model which will make it possible to predict academic success as early as possible in the academic year and thus provide for an optimal distribution of teaching resources in order to curb academic failure.

First of all, we will present the methodology that we have adopted. Then, we will describe the data. We will present the different results obtained by different data mining methods. Finally, we will compare their performance with a linear discriminant analysis.

METHODOLOGY

Many studies (Ardila, 2001; Boxus, 1993; Busato et al., 1999, 2000; Chidolue, 1996; Furnham et al., 1999; Gallagher, 1996; Garton et al., 2002; King, 2000; Minnaert and Janssen, 1999; Parmentier, 1994.) were undertaken in order to try to explain the academic performance or to predict the success or the failure; they highlighted a series of explanatory factors associated to the student.

We first consulted this abundant literature on psychology and education in order to establish a list of factors that this professionals believe to be causes (or indices) of success and failure in the first academic year. Then, we targeted a set of factors to be taken into account based on a model used by Parmentier (1994) and adapted. He shows that the intermediate and final academic performance of the students are influenced by three sets of factors, in interaction with each other, of which the first groups structural or stable factors while the other two are composed of process or changing factors. The first of these sets groups everything relating to the personal history of the student (his identity, his socio-family past, his academic past, etc.). The second can be interpreted as the expression of the involvement of the student in his studies or of his behavior in relation to his studies (participation in optional activities, meetings with his professors to ask questions or to obtain feedback on periodic examinations, etc.). The last set of factors groups all the student's perceptions (the way in which he perceives the academic context, his professors, courses, etc.).

Secondly, we created a questionnaire allowing us to collect a large amount of interesting information on a certain number of students. In November 2003, we distributed this questionnaire to post-secondary first-year students in three universities of the French Community of Belgium. In November 2004, we again distributed the questionnaires to the same Belgian universities and also to a French university. However, results concerning this last year are not yet available. Figures presented here concern only data relating to the academic year 2003-2004, i.e. on 227 students registered in first year at the Catholic University of Mons (FUCaM), 151 at the Faculté Polytechnique de Mons (FPMs) and 155 at the Faculté Universitaire des Sciences Agronomiques de Gembloux (FSAGx). These students comprise 227 students in management science or political science, and 155 bio-engineering students who have all completed their secondary studies (the only requirement for entering to these faculties in Belgium) and 151 civil engineering students (FPMs) who have passed an entry examination in order to be accepted on their course.

The completed questionnaires led to the construction of the database in which each student is described according to a certain number of criteria or attributes such as age, education level of his/her parents, the student's perceptions of the university world that surrounds him/her, etc. As we wish to target students who need to be helped, it is necessary to extract information from the database that allows us to identify their profile. This will be done using data mining and/or statistical methods.

As a result we will be able to give to these students priority allocation of the limited resources available for teaching support (tutorial by an older student, private tutorial by a professor, etc.). Before analyzing the collected data, we should note that a model obtaining good rates of classification in internal validation is no interest to us and that only predictive power relating to new individuals is truly significant. This is why we have never worked on more than 70 per cent of the students, keeping the remaining 30 per cent for the validation phase.

DATA

The questionnaire comprised 42 questions or question-series, almost all of them closed, from which we extracted 148 variables, most of which were either binary or coded into 5 response categories, although some were also coded as percentages. Hence each student who completed the survey would be represented by 375 variables in the database.

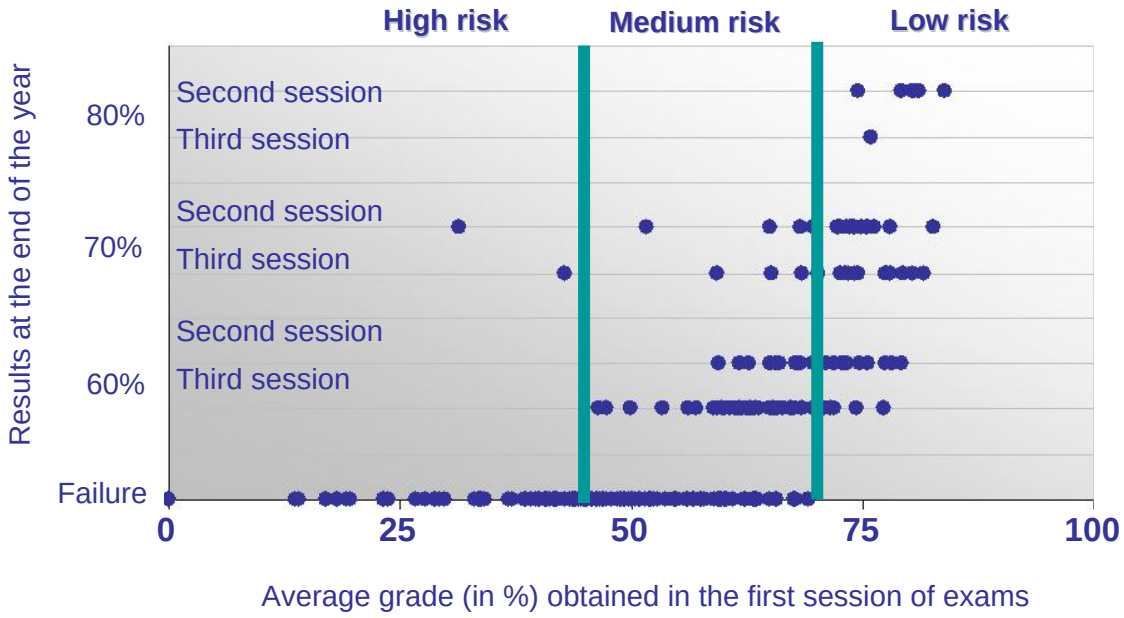

Fig. 1. Construction of the decision variable (example)

To this set of variables we will add a particular variable which will be used as a decision variable (variable to be explained). If it is academic success that we wish to explain, this variable would not be available until September following the administration of the questionnaire, since it is necessary to wait to find out if the student does or does not move on to the next academic year. Our objective is to classify students, before the first session of exams into three groups according to their probability of success. This will allow to identify the students who require aid, and to propose to them specific actions. One variable with three categories ('low risk', 'medium risk', 'high risk') was built a posteriori. This variable had to be not only the reflection of the total results of the students but also of their capacity to develop during the year. A graph (Figure 1) relating the average of the marks obtained by a student during the January session with their academic rank at the end of the year made it possible to distinguish clearly between two contrasted groups of students: those who obtained a total mark of less than 45% in the January session, all failed at the end of the year (all except for two students) and those who, having obtained a total mark in January of more than 70%, all passed at the end of the year. We thus constructed the decision variable to reflect the zone (left, central or right) occupied by the student.

At the level of the variables themselves, a preliminary study has shown us how harmful variables which are not correlated with the decision variable can be to prediction in our field of application. An analysis of the correlations between the explanatory variables and our decision variable was thus carried out. The rest of this chapter is devoted to a description, according to the classification of Parmentier (1994), of variables that are shown to be the most correlated. The value of the correlation coefficient between each explanatory variable and the decision variable is indicated between brackets. For the continuous variables, the value of the correlation coefficient is followed by one, two or three stars to indicate if the test of significance of this coefficient were shown to be significant (α=0,05), very significant (α=0,01) or highly significant (α=0,001). In the case of variables having only some categories, the coefficient of correlation is less relevant than the result of a Chi-square test carried out to measure the degree of dependence between the explanatory variable and the decision

variable. For these variables, we will thus indicate the p-value of this Chi-square test with the value of the coefficient of correlation.

Personal history of the student

It comes as no surprise that those variables which relate to the scholastic history of the student and to his socio-familial background have the highest correlation coefficients. Thus, the average in the last year of secondary education (rhetoric) of the students (rhô = 0.337 ***) is the variable which has shown to be the most correlated among all those which we have tested. The number of hours of mathematics in the last year of secondary education (rhô = 0.313 ***) is also correlated in a highly significant way with university success. In the same way, not having to finance one's own studies (rhô = 0.162; p-value = 0.027), not having followed courses in economic sciences or social sciences in secondary education (rhô = 0.157; p-value = 0.030), not being older than average, probably representative of school failure in the past (rhô = 0.152) or even not smoking (rhô = 0.177; p-value = 0.006) are all factors that significantly influence university success. Conversely, the sex of the student, the highest educational level obtained by his/her parents, parental occupation, the number of brothers or sisters (whether older or younger or having brothers or sisters already in higher education), married or separated parents are not factors significantly correlated with success.

Implication of student behavior

The number of hours which the student admits to attending class is highly correlated with academic success (rhô = 0.250 ***). The less a student mentions regularly missing classes, the higher his likelihood of succeeding (rhô = 0.164 ; p-value = 0.017). It is a bad idea for a student to miss even those classes that are the least frequented by fellow students. (rhô = 0.134 ; p-value = 0.0004).

It is advisable that a student thoroughly understands the material he is studying (rhô = 0.165; p-value = 0.002) and not simply dwell on what interests him (rhô = 0.159; p-value = 0.0006). Here the theory of Entwistle (Entwistle, 1988) on various ways of how students study may be regarded as relevant.

Finally, let us note that students who understand that the course requires regular homework are also those who succeed (rhô = 0.143 ***).

All these factors are very strongly related to success, unlike variables relating to students' extra-curricular activities. Thus, whether they participate in student-organized activity, whether they have undergone student initiation ceremonies (or hazing), whether they spend time pursuing hobbies or with family are not variables significantly correlated with success. Before concluding this section on student behaviors and the implication of these for their studies, we note that we adapted the scale of Laurent and Kapferer (Laurent, 1986), well-known in marketing, to this field. Thus, if one defines involvement as a non-observable state of motivation, excitation or interest, created by an object or a specific situation and involving behaviors (Rothschild, 1984), according to Laurent and Kapferer, all treatments of this involvement in social psychology or in marketing are treatments of one or more variables which they identify as being the causes of the involvement. By adapting these involvement factors to the world of the university, we obtain a series of 16 questions which closely reflect the initial scale suggested by Laurent and Kapferer. This series of questions thus gives us another way to measure the involvement of the students and the resulting variable is also highly significant correlated with university performance.

Perceptions of the student

This last group of parameters, the perceptions of the student, often seem to be more subjective than tangible. However, we also find highly significant variables here. Most important is the confidence that the student has in his or her own abilities. In fact, a student who have a strong confidence in his/her capacities is more persistent, more productive and more motivated by his academic studies (Zimmerman, 1992). Indeed, the higher the student rates his or her own chances of success, the greater the probability that he or she actually succeeds (rhô = 0,326***). In the same way, it is better not to find the course too difficult from the beginning of the year (rhô = 0.150; p-value = 0.030) or to think that one was badly prepared for higher education (rhô = 0.182; p-value = 0.017). Students who, in November, felt that they had chosen well when enrolling at their university (rhô = 0.182; p-value = 0.0002), those who did not overestimate the study time necessary for success (rhô = 0.159; p-value = 0.012) and those who preferred group work to working alone (rhô = 0.232; p-value < 0.0001) are those most likely to succeed a few months later. On the other hand, the variables that are most significant in explaining success or failure do not relate to perception of the environment nor, to a large extent, to the perception of the academic context.

It should be emphasized that all these measurements were carried out on all 533 students of our sample and thus constitute average values for the three universities considered. However, large differences exist between the three sub samples corresponding to the three universities. Thus, the variable most correlated on average occupies

third, sixth and nineteenth position in the classification of the variables most correlated in the sub samples. The variable constructed to reflect the student's average in the last year of secondary schooling is the most correlated with our decision variable in two of the three institutions, but it is classified 167th in the third. There are many such examples. They are confirmed by a test of Chi-square between the decision variable and a categorical variable modeling membership of one or the other university. This test returns a Chi-square of 82 with 6 degrees of freedom that corresponds to a p-value of 10^{-15}, so that there is no doubt of a relationship between university success and the fact of belonging to one or other of our three sub samples. For reasons of confidentiality, however, we are not able to provide details of the results obtained for each institution.

Summary

One variable in five proved to be correlated (of which more than one third were very strongly correlated) with university performance. The most correlated concerned to attendance at courses, previous academic experience mainly concerned with mathematics and study skills and the estimated chance of success. Significantly influential factors were found in each of the three groups of variables and we note that if many things are decided before entry to university (structural factors), nothing is yet final and the process factors also explain a large part of academic performance.

RESULTS

The objective of this study is to determine if it is possible to predict the decision variable using the explanatory variables which we retained in the model (following suppression of the variables which did not show direct influence on the decision variable in term of correlation or independence such as measured by a Chi-square test) and which characterize the profile of the 533 first-year students university, and to do this before November. To do this, we used several methods: decision trees, random forests, neural networks and a linear discriminant analysis, and compared results obtained from each one of them. In order to validate the results obtained, we estimated our model using a 70% subset of the data file and used the remaining 30% as a validation set for the models obtained from the estimation subset.

Decision tree

A decision tree (Witten and Frank, 2005) is made up of a hierarchical structure of decision nodes which are linked with branches. To determine the root node and after that the next nodes, for each attribute we calculated which one will most exactly classify objects (here the students) according to the decision variable values. From each decision node, branches going out to each node correspond to the various possible answers to the test. For a continuous variable, the test will be in this form: 'if the value on the variable is higher than a threshold value then take the first connect, otherwise take the second'. For the categorical variables, one of the outgoing branches of the node is associated with each category of the variable. The final nodes are called leaf nodes and are linked during the training phase to one of the categories of the decision variable. To realize a prediction on a new individual, it is enough to make him traverse the graph to a leaf. The leaf which the new individual reaches will determine the predicted value of the decision variable.

We used the SAS/Enterprise Miner software to build such a decision tree. We have chosen to build our tree on the basis of Shannon's entropy and of the algorithm ID3 (Quinlan, 1979). We obtained a tree which presents the advantage of being particularly simple to interpret. The classification of students is only carried out on the basis of five variables. Thus, in decreasing order of importance, we find the variable relating to the weekly attendance of the student at courses, the feeling of having chosen well by registering at this university, and three less important variables relating to the reasons that led the student to register at university or to begin this type of study.

TABLE 1. Summary of the results of validation for the decision trees

		Predictions carried out by SAS/Enterprise		
		High Risk	**Medium Risk**	**Low Risk**
ACTUAL	**High Risk**	**48.65 %**	10.81 %	40.54 %
	Medium Risk	33.85 %	**18.46 %**	47.69 %
	Low Risk	22.41 %	17.24 %	**60.34 %**

On the other hand, as shown in table 1, the proportion of correct predictions in the model validation phase are not very good: only 48.65% of the students of class 1 were correctly classified by means of the elaborated tree; only 18.46% of the students of class 2 were actually classified into class 2 and 60.34% of the students of class 3 were correctly classified into class 3. We can see that for the extreme classes, the decision tree manages reasonably well, but predictions concerning students at 'medium risk' are rather eccentric (although this is the most densely populated class with 40% of the students, compared to 27% in the class 'high risk' and 33% in the class 'low risk'). We obtain an overall rate of correct classification of only 40.63%.

Random Forests

Random forests are one of the most successful ensemble methods which exhibits performance on the level of boosting. The method is fast, robust to noise, does not overfit. (Robnik-Sikonja, 2004).

The random forests (Breiman, 2001) is a method that tries to build a set of individual classification trees, using each one of them a small number of variables. The idea is to build several sets of decision rules and to extract a maximum of information from the training set.

Concretely, over a population of n individuals described by p variables, we are going to make a double random selection: sample n individuals with replacement among n and sample q variables without replacement among p (classically, q is equivalent to the square root of p). With this double random selection, we will have retained a subset of q variables to describe n individuals (i.e. the same number of individuals from the beginning, but among them are presented several times while others have been left aside). Over this game of modified data we are going to build a decision tree according to the CART algorithm.

We have chosen to build 800 trees by this way, each of them leading to an ensemble of rules, and like this, to a prediction to a new individual. This new individual will belong to the class which more number of forest trees had affected.

We used R software to build this method. Among all the variables, we found that the variables concerning to the estimated chances of success, the number of hours of mathematics, sciences and literature in the last year of secondary education, his/her age, among others variables, like the more important in the results.

TABLE 2. Summary of the results of validation for the Random Forest

		Predictions carried out by R		
		High Risk	**Medium Risk**	**Low Risk**
ACTUAL	**High Risk**	**22.92 %**	67.36 %	9.72 %
	Medium Risk	6.64 %	**68.72 %**	24.64 %
	Low Risk	3.37 %	41.57 %	**55.06 %**

From the analysis of the results of the random forest method, presented in Table 2, we find a total rate of correct classification of 51.78%. These results are a little better than those that we obtained by the decision trees.

Neural networks

Neural networks (Dreyfus, 2005) are tools frequently used for classification, estimation or prediction. The method tries to categorize by an iterative algorithm the working of the human brain. It comprises a skeleton of neurons connected to one another which breaks up into three zones: the entry layer, the hidden layers and the exit layer. The variables of the problem are acted upon and weighted by the entry layer, which then transmits this information to the hidden layers; these combine all the information into a single value which is passed to the exit neuron, and which acts as a kind of estimated value for the decision variable.

On the basis of our training set containing 70% of the individuals, we built a model by means of the neural networks procedure in SAS/Enterprise Miner. The retained model is a multi-layer perceptron with a hyperbolic tangent as activation function, with one hidden layer containing three neurons, and using one exit neuron to carry out the predictions on our decision variable. The application of a procedure of selection of variables upstream of the use of the neural networks allowed us to determine which variables will be used in the model and the number of entry neurons. These variables are 23 and they cover all three categories of factors defined by Parmentier. For example, the student's age, the average percentage of classes followed during one week, the number of hours of mathematics studied at secondary level or their average at the end of the last grade.

TABLE 3. Summary of the results of validation for the neural networks

		Predictions carried out by SAS/Enterprise		
		High Risk	**Medium Risk**	**Low Risk**
ACTUAL	**High Risk**	**45.95 %**	40.54 %	13.51 %
	Medium Risk	30.88 %	**47.06 %**	22.06 %
	Low Risk	00.00 %	38.18 %	**61.82 %**

As Table 3 shows, the rates of correct classification are not fantastic, even if they are slightly higher than those of Tables 1 and 2, the total percentage of correctly classified students reaching 51.88% for the neural networks.

Linear discriminant analysis

Let us remember that discriminant analysis is a method which makes it possible to classify an individual into one of *g* groups to which he/she may belong (McLachlan, 2004). To do this, the method determines an allocation rule based on *p* variables which characterize each individual to be classified. This allocation rule is defined as a function of *g* samples taken from each of the groups. However, several solutions exist to define the classification rule which makes it possible to assign each new individual to one of the defined *g* classes.

By way of comparison with more recent methods such as decision trees or neural networks, it is interesting to look at the results (Table 4) provided by a linear discriminant analysis with a preliminary selection of variables via a "stepwise" strategy (carried out with SAS software). The variables which were selected and retained for the construction of discriminant functions almost match those chosen by the neural networks.

TABLE 4. Summary of the results of validation for the linear discriminant analysis

		Predictions carried out by SAS/Enterprise		
		High Risk	**Medium Risk**	**Low Risk**
ACTUAL	**High Risk**	**45.95 %**	40.54 %	13.51 %
	Medium Risk	22.06 %	**57.35 %**	20.59 %
	Low Risk	1.82 %	30.91 %	**67.27 %**

From the analysis of the results presented in Table 4, which are 20 - 30% worse than those that would have been observed if we had been interested in a binary variable success/failure, we find a total rate of correct classification of 57.35% - this is the least bad result of the four methods.

CONCLUSIONS AND PERSPECTIVES

We noted that 20% of our variables showed significant correlation with academic success. These 20% of variables are found in each of the factor categories proposed by Philippe Parmentier in his model. The same is true for the variables used by each of the three methods of prediction that we compared in this research. The theoretical model on which we base our research consequently seems rather well adapted to what we expected.

With respect to the results obtained by the methods of prediction, we conclude that the rates of prediction obtained in validation are not remarkable. We note large disparities between the three universities from which our sample are taken – these are not beneficial for the predictive power of each of the four methods. However, discriminant analysis, and to a lesser extent neural networks and random forests, seem to be able to lead to interesting results, on the condition, however, that in the future we increase the size of our samples from each university, incorporating data from an additional academic year, for example.

Will the factors influencing academic success be stable year by year within the same university? Is it possible to find facts that are common across the different universities studied and that make it possible to produce predictions like ours? Can a combination of different prediction methods lead to the improvement of the overall result? Will we find big differences by crossing the borders? Will influential factors be similar? So many questions today that still need to be answered over the coming months.

REFERENCES

Ardila, A. (2001). *''Predictors of university academic performance in Colombia''*, International Journal of Educational Research, vol. 35, pp. 411-417.

Boxus, E. (1993). *''Réussites en candidatures''*, Rapport du Conseil Interuniversitaire de la Communauté Française, Bruxelles.

Breiman, L. (2001). Random Forests. *Machine Learning*. Vol. 45, N°1, 5-32.

Busato, V.V., Prins, F.J., Elshout, J.J., Hamaker, C.. (1999). *''The relation between learning styles, the Big Five personality traits and achievement motivation in higher education''*, Personality and Individual Differences, vol. 26, pp. 129-140.

Busato, V.V., Prins, F.J., Elshout, J.J., Hamaker, C. (2000). *''Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education''*, Personality and Individual Differences, vol. 29, pp. 1057-1068.

Chidolue, M.E. (2001). *''The relationship between teacher characteristics, learning environment and student achievement and attitude''*, Studies in Educational Evaluation, vol. 22, 3, pp. 263-274, 1996.

Dreyfus, G. (2005). *"Neural Networks: Methodology and applications".* Springer-Verlag Berlin and Heidelberg GmbH & Co. K

Droesbeke, J.-J., Hecquet, I., Wattelar, C. (2001). *La population étudiante*. Paris : Ellipses.

Entwistle, N. (1988). Motivational factors in students' approaches to learning. In R.R. Schmeck (Ed.), *Learning strategies and learning styles*. 21-51. New York: Plenum.

Furnham, A., Jackson, C.J., Miller, T. (1999). *''Personality, learning style and work performance''*, Personality and Individual Differences, vol. 27, pp. 1113-1122.

Gallagher, D.J. (1996). *''Personality, Coping, and Objective Outcomes : extraversion, neuroticism, coping styles, and academic performance''*, Person. individ. diff., vol. 21, 3, pp. 421-429.

Garton, B.L., Dyer, J.E., King, B.O. (2002) *''Factors associated with the academic performance and retention of college agriculture students''*, NACTA Journal.

King, A.R. (2000). *''Relationships between CATI personality disorder variables and measures of academic performance''*, Personality and Individual Differences, vol. 29, pp. 177-190.

Minnaert, A., Janssen, P.J. (1999). *''The additive effect of regulatory activities on top of intelligence in relation to academic performance in higher education''*, Learning and Instruction, vol. 9, pp. 77-91.

McLachlan, Geoffrey, J. (2004). *Discriminant analysis and statistical pattern recognition.* NY: Wiley-Interscience. (Wiley Series in Probability and Statistics).

Laurent, G., Kapferer, J.N. (1986). Les profils d'implication. *Recherche et applications en marketing*, n°1, 41-57.

Parmentier, P. (1994). *La réussite des études universitaires: facteurs structurels et processuels de la performance académique en première année en médecine.* Ph.D. Faculté de Psychologie et des Sciences de l'Éducation, Catholic University of Louvain.

Quinlan, J.R. (1979). *Discovering rules by induction from large collections of examples*. Edinburgh: Ed. Expert Systems in the Micro Electronic Age, Edinburgh University Press.

Robnik-Sikonja, M. (2004). *"Improving Random Forests".* In J.F. Boulicaut et al. (eds) : Machine learning, ECML 2004 proceedings, Springer, Berlin.

Rothschild, M.L. (1984). Perspectives on Involvement: Current Problems and Future Directions. *Advances in Consumer Research*, Vol.11, 216-217.

Witten, I.H., Frank, E. (2005). *Data Mining: practical machine learning tools and techniques.* Morgan Kaufmann series in data management system.

Zimmerman, B.J., Bandura, A., Martinez-Pons, M. (1992). ''Self-motivation for academic attainment : the role of self-efficacy beliefs and personal goal setting'', American Educational Research Journal, vol. 29, 3, pp. 663-676.