Comparaison de méthodes sur un modèle de prédiction de la réussite universitaire [1]

Jean-Philippe Vandamme*, Nadine Meskens*, Abdelhakim Artiba**

*Facultés Universitaires Catholiques de Mons (FUCaM)
151 Chaussée de Binche, 7000 Mons – Belgique
Tel : +32(0)65323211 - Fax : +32(0)65315691
vandamme@fucam.ac.be, meskens@fucam.ac.be

** École de technologie Supérieure (ETS)
1100, rue Notre-Dame Ouest, Montréal (Québec) H3C 1K3 – Canada
Tel : +32(0)65321542 - Fax : +32(0)65361746
abdelhakim.artiba@etsmtl.ca

Résumé

Depuis longtemps, l'échec scolaire en première année universitaire alimente bon nombre de débats. De nombreux psychopédagogues ont tenté de le comprendre puis de l'expliquer. De nombreux statisticiens ont quant à eux essayé de le prévoir. Nos recherches visent à établir un modèle permettant de déterminer le plus tôt possible dans l'année le groupe des étudiants de première année sur qui il faut cibler en priorité les ressources pédagogiques dont on dispose afin d'améliorer le taux de réussite. Pour cela, nous avons transposé sous forme de questionnaire les hypothèses posées dans de nombreux modèles théoriques. Ensuite, après avoir récolté via ce questionnaire des données suffisamment nombreuses et diverses, l'objectif a été d'extraire de l'information via des méthodes statistiques ou de data mining pur et ainsi permettre la classification des étudiants en trois classes les plus homogènes possibles. Cet article décrit la méthodologie adoptée, les variables qui ont été analysées et les méthodes qui ont été utilisées et comparées. Avec la mise en parallèle des résultats fournis par les diverses méthodes (analyses discriminantes, régressions, ensembles approximatifs, arbres de décision, etc.), il est possible de mettre en lumière leurs différences de performance. En effet, certaines méthodes se sont montrées plus efficaces en terme de taux de prédictions correctes réalisées, là où d'autres ont surtout été intéressantes pour leur capacité à mettre en évidence les facteurs prédictifs de la réussite universitaire.

Mots clés

Arbres de décision – Ensembles approximatifs – Analyses discriminantes – Éducation – Prédiction

[1] Cette recherche est financée par le "Programme Pôles d'attraction interuniversitaires - Etat belge - Services fédéraux des affaires scientifiques techniques et culturelles".

Comparaison de méthodes sur un modèle de prédiction de la réussite universitaire

Résumé

Depuis longtemps, l'échec scolaire en première année universitaire alimente bon nombre de débats. De nombreux psychopédagogues ont tenté de le comprendre puis de l'expliquer. De nombreux statisticiens ont quant à eux essayé de le prévoir. Nos recherches visent à établir un modèle permettant de déterminer le plus tôt possible dans l'année le groupe des étudiants de première année sur qui il faut cibler en priorité les ressources pédagogiques dont on dispose afin d'améliorer le taux de réussite. Pour cela, nous avons transposé sous forme de questionnaire les hypothèses posées dans de nombreux modèles théoriques. Ensuite, après avoir récolté via ce questionnaire des données suffisamment nombreuses et diverses, l'objectif a été d'extraire de l'information via des méthodes statistiques ou de data mining pur et ainsi permettre la classification des étudiants en trois classes les plus homogènes possibles. Cet article décrit la méthodologie adoptée, les variables qui ont été analysées et les méthodes qui ont été utilisées et comparées. Avec la mise en parallèle des résultats fournis par les diverses méthodes (analyses discriminantes, régressions, ensembles approximatifs, arbres de décision, etc.), il est possible de mettre en lumière leurs différences de performance. En effet, certaines méthodes se sont montrées plus efficaces en terme de taux de prédictions correctes réalisées, là où d'autres ont surtout été intéressantes pour leur capacité à mettre en évidence les facteurs prédictifs de la réussite universitaire.

1 Introduction

Depuis longtemps, l'échec scolaire en première année universitaire, qui concerne trois étudiants sur cinq en Belgique, préoccupe pour diverses raisons évidentes les enseignants, les parents, les décideurs politiques, les gestionnaires d'institutions universitaires... Dès lors, il n'est pas étonnant de voir psychopédagogues et statisticiens unir leurs forces pour construire des modèles permettant d'identifier le plus rapidement possible dans l'année les étudiants qui auront ou non des difficultés pour réussir leur première candidature (BAC 1).

Afin de proposer une démarche visant à construire un modèle de prédiction de la réussite universitaire qui permettrait de cibler les étudiants ayant le plus besoin de soutien pédagogique, nous présenterons ici, outre le contexte dans lequel nous nous situons, la méthodologie que nous avons adoptée. Nous décrirons les données dont nous disposons, nous présenterons les différents résultats obtenus par les méthodes des arbres de décision et des ensembles approximatifs et nous comparerons leurs performances avec celle de l'analyse discriminante linéaire.

2 Contexte

Lorsque nous analysons les résultats des étudiants de première candidature dans les universités francophones de Belgique, nous constatons qu'environ 60 % des étudiants de première génération échouent ou abandonnent en première année. (Droesbeke et al. 2001) ont observé que les taux de réussite, de redoublement et d'abandon sont relativement stables depuis plus de 10 ans. Ils ont établi que le taux de réussite des entrants de première candidature provenant de l'enseignement secondaire avoisine les 41%, le taux de redoublement est de l'ordre de 26% et le taux d'abandon est de 33%. Ces chiffres doivent susciter réflexion et conduire à diverses actions susceptibles de réduire le coût économique, social et humain préoccupant qu'entraîne ce taux élevé d'échec en première année. C'est pourquoi, depuis quelques années, la plupart des universités belges offrent des activités supplémentaires au programme requis de première année (enseignement assisté par ordinateur, monitorat,...) en vue de remédier aux lacunes constatées auprès d'étudiants "en situation d'échec" notamment après la session des examens de janvier.

Notre objectif final est de pouvoir classer les étudiants en trois groupes : le groupe des étudiants ayant une forte probabilité de réussir l'année ("low risk"), le groupe des étudiants qui peuvent éventuellement réussir moyennant des actions à mener ("medium risk") et le groupe des étudiants ayant une forte probabilité d'échouer (ou d'abandonner) ("high risk"). Cette classification a de moins en moins d'intérêt au fur et à mesure que l'année scolaire avance. Il n'est pas intéressant de devoir attendre février voire avril pour arriver à cibler correctement les étudiants qui ont vraiment besoin de mesures d'accompagnement. Notre objectif est donc bien de proposer une méthode permettant d'identifier les étudiants « à risque » le plus tôt possible dans l'année.

3 Méthodologie

Afin d'atteindre notre objectif, il faut donc en tout premier lieu que nous déterminions quels sont les critères qui vont réellement nous permettre de réaliser des prédictions sur la réussite des étudiants. Une fois que nous aurons arrêté la liste des facteurs susceptibles de conduire au classement de ces étudiants en plusieurs groupes, il nous faudra établir un questionnaire permettant de relever cet ensemble de facteurs pour chacun des étudiants. Les questionnaires complétés conduiront alors à la construction de la base de données qui nous servira à réaliser effectivement nos prédictions. En effet, nous avons besoin d'un tableau de données où chaque étudiant sera décrit selon un certain nombre de critères ou d'attributs tels que son âge, le niveau d'éducation de ses parents, ses perceptions par rapport au monde universitaire qui l'entoure, etc. Nous avons aussi besoin de méthodes statistiques ou mathématiques (data mining) afin de traiter cette base de données, d'en extraire de l'information et de la rendre utilisable pour cibler efficacement les étudiants qui ont le plus besoin d'être aidés, ceux à qui il faut consacrer en priorité les ressources limitées dont on dispose pour faire de l'accompagnement pédagogique (tutorat par un étudiant plus âgé, monitorat par un professeur en particulier, etc.). Et c'est seulement la conjonction des données et des méthodes qui conduiront au résultat attendu.

En premier lieu, nous avons donc parcouru l'abondante littérature ayant trait à la psychopédagogie afin d'établir une liste de facteurs que les professionnels du domaine estiment être des causes ou des indices de réussite ou d'échec en première année universitaire. Nous avons alors ciblé une série de facteurs à prendre en considération et d'hypothèses à tester en nous basant sur un modèle adapté de (Parmentier 1994) où les performances académiques intermédiaires et finales des étudiants sont influencées par trois ensembles de facteurs en interactions les uns avec les autres. Le premier de ces ensembles reprend tout ce qui concerne l'histoire personnelle de l'étudiant (son identité, son passé socio-familial, son passé scolaire, etc.). Le deuxième peut s'interpréter comme l'expression de l'implication de l'étudiant dans ses études ou de son comportement face à celles-ci (participation à des activités facultatives, rencontre avec ses professeurs pour poser des questions ou obtenir un feedback d'un examen partiel, etc.). Le dernier ensemble de facteurs regroupe toutes les perceptions de l'étudiant (la manière dont il perçoit le contexte académique, ses professeurs, les cours, etc.).

Dans un second temps, nous avons constitué un questionnaire permettant de relever chez un certain nombre d'étudiants un maximum d'informations intéressantes. En novembre 2003, nous avons distribué ce questionnaire à des étudiants en première année post-secondaire dans deux institutions marocaines et trois universités de la Communauté Française de Belgique. En novembre 2004, nous avons élargi vers la France. Cela dit, tous les résultats ne sont pas encore disponibles. Les chiffres présentés ici ne portent encore que sur les étudiants inscrits

en première année aux Facultés Universitaires Catholiques de Mons (FUCaM) en 2003-2004, c'est-à-dire un peu plus de 200 étudiants en sciences de gestion ou en sciences politiques, donc ayant terminé leurs études secondaires, seule condition pour accéder à ce type d'études en Belgique.

Notons encore avant d'analyser les données que nous avons récoltées qu'un modèle obtenant de bons taux de classement en validation interne ne nous intéresse bien entendu pas et que seul le pouvoir prédictif sur de nouveaux individus est véritablement significatif. C'est pourquoi nous n'avons jamais travaillé que sur 70 pourcents des étudiants, gardant ainsi les 30 autres pourcents pour la phase de validation.

4 Données

Tous les facteurs dont on supposait une influence sur les performances académiques ont fait l'objet de plusieurs questions dans le questionnaire que nous avons distribué. Ce questionnaire comportait ainsi 42 questions ou séries de questions, presque exclusivement fermées, desquelles nous avons extrait 148 variables souvent binaires ou à 5 modalités mais parfois aussi codées sous forme de pourcentages. À partir de ces 148 variables, nous en avons créé 227 autres, principalement par re-codification ou par combinaison. Au total, chaque étudiant qui a participé à l'enquête est donc représenté par une ligne dans notre base de données, soit par 375 variables.

Actuellement, nous ne disposons encore que des résultats finaux des étudiants de première année inscrits en 2003-04 aux FUCaM. Or, au mois de novembre 2003, 227 étudiants avaient accepté de se prêter au jeu : c'est donc sur 227 questionnaires que se basent les résultats chiffrés qui vont suivre. Au sein de ces étudiants, il est bon de noter dès à présent que 117 ont réussi pour 110 qui ont échoué.

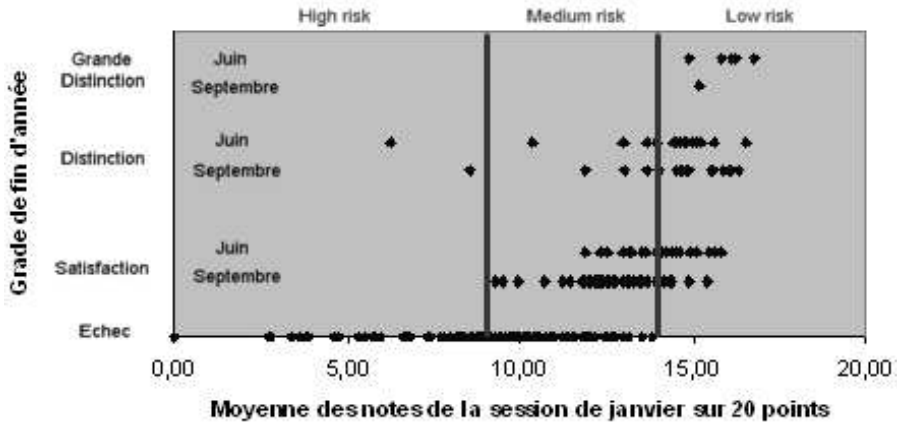

FIG. 1 - *Construction de la variable de décision.*

A notre table de données, il faut à présent ajouter une variable un peu particulière : celle qui servira de variable de décision (dirait-on en data mining), de variable à expliquer (dirait-on en régression). Si on cherche à expliquer la réussite universitaire, on ne dispose de cette variable qu'au mois de septembre qui suit l'administration du questionnaire puisqu'il faut attendre pour savoir si l'étudiant est ou non admis dans l'année supérieure. Notre objectif étant de proposer au cours du premier semestre une découpe en trois groupes d'étudiants selon leur probabilité de réussite, la variable binaire réussite/échec n'est pas la plus appropriée. Nous avons donc construit a posteriori une variable à trois modalités se voulant

l'image de ce que nous avons appelé les étudiants "low, medium ou high risk". Cette variable devait être non seulement le reflet des résultats globaux des étudiants mais aussi de leur capacité à évoluer aux cours de l'année. Nous avons donc construit un graphique (FIG. 1) mettant en rapport la moyenne des notes obtenues par un étudiant lors de la session de janvier avec son grade académique en fin d'année. Ce graphique met clairement en évidence deux groupes extrêmes d'étudiants dont la note de janvier était telle que les surprises ne pouvaient être que très rares. Nous avons donc constitué la valeur de la variable de décision en fonction de la zone (gauche, centrale ou droite) dans laquelle l'étudiant se trouvait.

Au niveau des variables elles-mêmes à présent, une étude préliminaire réalisée par (Vandamme et al. 2004) nous a montré à quel point les variables non corrélées à la variable de décision étaient néfastes à la réalisation de prédiction dans notre domaine d'application. Nous avons donc décidé de ne garder comme variables dans nos modèles que celles qui montreront une corrélation significative avec la variable binaire « réussite de l'année académique ». La suite de ce chapitre est consacré à la description, selon la classification de Parmentier (1994), des variables qui ont donc été retenues.

4.1 Histoire personnelle de l'étudiant

Sans trop de surprise, on retrouve des variables concernant le passé scolaire de l'étudiant et d'autres sur son passé socio-familial avec des coefficients de corrélation parmi les plus élevés. Ainsi, la moyenne en terminale (rhétorique) de l'étudiant (rhô = 0.282) ou le nombre d'heures de mathématiques en fin de secondaire (rhô = 0.278) constituent des variables corrélées de manière hautement significative à la réussite universitaire. De même, être de langue maternelle française (rhô = 0.238), avoir des parents dont l'activité professionnelle est importante (rhô = entre 0.242 et 0.201), avoir choisi son université parce que le programme des cours semblait intéressant (rhô = 0.261), trouver ses parents plus encourageants que décourageants (rhô = 0.280), ne pas financer soi-même ses études (rhô = 0.228) ou encore ne pas fumer (rhô = 0.195) sont tous des facteurs influençant très significativement (voire hautement significativement) la réussite universitaire. Au contraire, le sexe de l'étudiant(e), le temps écoulé depuis la fin de ses études secondaires, le niveau du diplôme le plus élevé obtenu par ses parents, le nombre de frères ou de sœurs (plus âgés ou plus jeunes ou ayant déjà entrepris des études universitaires), des parents mariés ou séparés ne sont pas des facteurs significativement corrélés à la réussite. Enfin, même si on constate souvent un impact important de l'âge d'entrée à l'université, nous n'avons, ici, mesuré qu'une corrélation faiblement significative (rhô = 0.167) avec la performance académique. Cette corrélation étant encore plus faible lorsqu'il s'agit d'attentes de l'étudiant ou de raisons de choix d'université ou de type d'étude.

4.2 Comportement d'implication de l'étudiant

« Va aux cours en étant bien reposé, tu comprendras mieux ! » Ma maman avait raison... La proportion d'heures de cours à laquelle l'étudiant dit participer est la variable la plus corrélée à la réussite universitaire (rhô = 0.373). Au moins l'étudiant cite de cours auxquels il est régulièrement absent, et au plus il a de chance de réussir (rhô = 0.296). Même sécher le cours le plus séché par ses condisciples n'est pas une bonne idée (rhô = 0.229).

Étudier en essayant de comprendre en profondeur (rhô = 0.241) et non pas simplement m'attarder sur ce qui m'intéresse le plus (rhô = 0.282), voilà encore un bon conseil, et voilà

aussi que la théorie d'Entwistle (1988) sur les différentes façons de concevoir l'étude chez les étudiants est une nouvelle fois à considérer comme pertinente.

Enfin, notons que le temps consacré aux loisirs d'étudiants ne doit pas être trop important (rhô = -0.204) au contraire de celui passé à dormir qui ne doit pas être réduit à sa portion congrue (rhô = 0.208).

Si tous ces facteurs sont au moins très significativement corrélés avec la réussite, il n'en va pas de même de la manière dont les étudiants gèrent le temps qui leur reste après les cours. Ainsi, le fait de se rendre ou non aux activités organisées par les étudiants, d'avoir ou non été baptisé (bizuté), de participer à des conférences ou des activités organisées par l'université en dehors des heures de cours ne sont pas des variables significativement corrélés à la réussite.

Avant de clore cette partie consacrée aux comportements et à l'implication des étudiants dans leurs études, signalons que nous avons adapté l'échelle de Laurent et Kapferer (1986), bien connue en marketing, au domaine qui nous occupe. Ainsi, si on définit l'implication comme un état non observable de motivation, d'excitation ou d'intérêt, créé par un objet ou une situation spécifique et entraînant des comportements (Rothschild, 1984), selon Laurent et Kapferer, toutes les manipulations de l'implication en psychologie sociale ou en marketing sont des manipulations d'une ou plusieurs des variables qu'ils ont identifiées comme étant les causes de l'implication. En reprenant ces causes et en les adaptant au monde de l'université, on obtient une série de 16 questions, copies conformes de l'échelle initiale proposée par Laurent et Kapferer. Cette série de questions nous donne donc un autre moyen de mesurer l'implication des étudiants et la variable qui en résulte est elle aussi significativement reliée aux performances universitaires.

4.3 Perceptions de l'étudiant

Dernier ensemble de paramètres, les perceptions de l'étudiant semblent souvent relever plus du subjectif que du tangible. Pourtant, ici aussi, on retrouve des variables hautement significatives. Au tout premier rang, la confiance qu'il a en ses capacités. En effet, au plus il se donne de chances de réussir son année et au plus il en a effectivement (rhô = 0.351). Dans le même ordre d'idées, il vaut mieux ne pas déjà trouver les cours trop difficiles dès le début de l'année (rhô = 0.242) ni estimer qu'on a été mal préparé aux études universitaires (rhô = 0.301). Les étudiants qui ont l'impression d'avoir fait un bon choix en s'inscrivant dans leur université (rhô = 0.234) et ceux qui ont trouvé que leurs professeurs étaient disponibles (rhô = 0.191) sont ceux qui ont le mieux réussi quelques mois plus tard. Par contre, la perception de l'environnement et dans une large mesure la perception du contexte académique ne regroupent pas les variables les plus significatives pour expliquer la réussite ou l'échec.

4.4 En résumé

Une variable sur cinq s'est avérée être corrélée (dont plus d'un tiers très fortement) à la performance universitaire. Les plus corrélées concernaient la présence aux cours, les chances de réussite estimées et la manière d'étudier. On a bel et bien retrouvé des facteurs significativement influents dans chacun des trois groupes de variables et on peut constater que si pas mal de choses se décident déjà avant l'entrée à l'université (facteurs structurels), rien n'est encore définitif et les facteurs processuels renferment aussi une grande part de l'explication des performances académiques.

5 Résultats

L'objectif de ce chapitre est donc de déterminer s'il est possible de réaliser des prédictions sur la variable de décision au moyen des 83 variables explicatives que nous avons retenues (celles qui parmi les variables de l'enquête sont hautement corrélées à la variable binaire « réussite de l'année académique ») et qui caractérisent le profil de 227 étudiants de première année universitaire au mois de novembre (échantillon FUCaM 2003-04). Pour ce faire, nous allons présenter les résultats fournis par les méthodes des arbres de décision, des ensembles approximatifs et par une analyse discriminante.

5.1 Arbre de décision

Nous avons utilisé notre tableau de données avec le logiciel SAS/Enterprise Miner pour effectuer une analyse par les arbres de décision (Rakotomalala 1997). Nous avons choisi de construire notre arbre sur base de l'entropie de Shannon et de l'algorithme ID3 (Quinlan 1979) et nous avons obtenu un arbre qui présente l'avantage d'être particulièrement simple à interpréter. La classification des étudiants s'effectue sur base de trois variables uniquement : une dans chaque ensemble de facteurs décrits par (Parmentier 1994). Ainsi, par ordre décroissant d'importance, on retrouve une variable sur l'approche d'étude (Entwistle 1988) de l'étudiant, une autre sur la part des cours scientifiques dans son cursus de l'enseignement secondaire et la dernière modélise les chances de réussite de l'année en cours que l'étudiant se donne.

	Prédictions réalisées par SAS/Enterprise Miner		
	1 (high risk)	**2**	**3 (low risk)**
Réel = 1	**31 %**	69 %	0 %
Réel = 2	3 %	**97 %**	0 %
Réel = 3	0 %	100 %	**0 %**

TAB 1 – *Synthèse des résultats de la validation pour l'arbre de décision*

Par contre, comme le montre le tableau 1, les pourcentages de prédiction correcte en phase de validation ne sont pas très bons. Pour la classe 1, on voit que 31% des étudiants de cette classe ont été prédits correctement dans celle-ci tandis que 69% ont été prédits dans la classe 2. On constate qu'aucun étudiant de la classe 3 n'a été correctement prédit !

Cela nous amène à un taux global de bonne classification de seulement 57%.

5.2 Ensembles approximatifs

L'essentiel de nos variables étant des critères et non pas des attributs, il était évident que la méthode des ensembles approximatifs (Pawlak 1991) allait donner de meilleurs résultats en tenant compte des extensions proposées par (Greco et al. 2001). Nous avons donc utilisé le logiciel 4eMKa (Poznan University of Technology 2000). La qualité d'approximation obtenue sur base de nos données est de 97,0%. Nous avons trouvé une réduction de 10 critères conduisant à la même qualité d'approximation et c'est donc sur cette réduction que nous avons généré les règles qui nous ont servi à la validation externe. Parmi les 10 critères retenus pour cette analyse, nous retrouvons les trois variables qui avaient été utilisées lors de la création de l'arbre de décision mais aussi des indicateurs sur la manière dont l'étudiant dit

répartir son temps hebdomadaire, d'autres sur la perception du choix d'études qu'il a posé et une variable sur la façon dont il perçoit les cours dispensés en petits groupes. Notons enfin avant de regarder les résultats (TAB 2) qu'un nombre important de règles a été généré (2318 au total), ce qui a nécessité un filtrage des règles trop particulières (suppression de celles qui ne concernaient qu'un étudiant ou deux).

	Prédictions réalisées par 4EMKA		
	1	**2**	**3**
Réel = 1	**10**	16	3
Réel = 2	7	**35**	9
Réel = 3	0	9	**6**

TAB 2 - *Matrice de confusion fournie par le logiciel 4eMKa.*

Au final, on a donc appliqué 348 règles sur 68 objets (30% de 227) dont 18 objets appartiennent à la classe 1, 38 à la classe 2 et 12 à la classe 3. Parmi ces 68 individus, 4eMKa en a classé 26 correctement (38,2%), en a classé 14 incorrectement (20,6%) et s'est prononcé de manière ambiguë dans le reste des cas. C'est pourquoi dans les résultats (TAB 2), on retrouve le nombre d'objets que 4eMKa a affecté à chacune des classes de décision (chacun des concepts) mais on constate vite que la somme des nombres d'objets ainsi affectés est supérieure au nombre d'objets à affecter. C'est la mise en évidence des 28 objets que 4eMKa a affecté de manière ambiguë à plusieurs concepts. Quoi qu'il en soit, les taux de classification correcte ne sont pas non plus transcendants, même si l'on est très indulgent et que l'on calcule le taux d'objets correctement classés au sein des objets qui ont été classés (26 / 40 = 65%).

5.3 Analyse discriminante linéaire

A titre de comparaison avec ces méthodes plus récentes, il est intéressant de regarder les résultats (TAB 3) fourni par une analyse discriminante linéaire (Palm 1999) avec une sélection préalable de variables (réalisée avec le logiciel SAS). La méthode stepwise (0,15 comme seuil d'entrée de sortie) a tout d'abord conduit à la sélection de 8 variables : on y retrouve, comme précédemment, les trois variables nécessaires à la construction de l'arbre de décision, mais aussi 4 variables appartenant à la réduction générée par les ensembles approximatifs et, enfin, une variable exprimant la quantité de cigarettes fumées par l'étudiant.

	Prédictions réalisées par SAS		
	1	**2**	**3**
Réel = 1	**53,85 %**	46,15 %	0 %
Réel = 2	12,62 %	**73,79 %**	13,59 %
Réel = 3	0 %	50,0 %	**50,0 %**

TAB 3 – *Synthèse des résultats de la validation pour l'analyse discriminante linéaire.*

A l'analyse de ces résultats, la première chose marquante est l'absence d'erreur grossière commise par l'analyse discriminante : jamais un étudiant d'un groupe extrême n'a été versé à l'opposé de sa position réelle. Ensuite, les taux d'erreur ne sont pas très importants en comparaison des résultats précédents. Enfin, en analysant les prédictions au cas par cas, on a constaté que les erreurs ont systématiquement été commises pour des étudiants proches des

limites entre les zones "low et medium risk" ou "medium et high risk", c'est-à-dire où c'est le moins problématique.

6 Conclusions et perspectives

En se basant sur des ensembles de variables souvent proches l'un de l'autre, les trois méthodes que nous avons comparées fournissent des résultats sensiblement différents. Les arbres de décision rendent l'analyse très sommaire alors que les ensembles approximatifs la rendent totalement opaque par la production d'une masse de règles. Au niveau des performances, l'analyse discriminante donne de bien meilleurs résultats que les méthodes de fouilles de données utilisées ici. Évidemment, on est en droit de se demander si la petitesse de l'échantillon d'étudiants n'est pas la principale cause de cet écart de performance. Cela dit, lorsque dans un futur proche nous aurons plus de données, l'analyse discriminante en tirera sans doute profit elle aussi. Est-ce qu'en mélangeant les Facultés universitaires, on pourra ajouter de l'information utile pour la réalisation de prédiction ? Trouvera-t-on de grandes différences en franchissant les frontières ? Les facteurs influents seront-ils semblables ? D'autres méthodes (comme les réseaux de neurones par exemple) se montreront-elles plus efficaces ? Est-il possible de regrouper les prédictions de plusieurs méthodes pour obtenir un affinement des résultats ? Autant de questions aujourd'hui sans réponse mais qui pourraient en trouver une dans les mois qui viennent.

7 Bibliographie

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

Entwistle N. (1988), Motivational factors in student's approaches to learning, Shmeck R.R., Learning strategies and learning styles, Plenum press.

Greco S., Matarazzo B., Slowinski R. (2001), Rough sets theory for multicriteria decision analysis, European Journal of Operational Research 129, pp 1-47.

Laurent G., Kapferer J.N. (1986), Les profils d'implication. Recherche et applications en marketing, n°2, Paris.

Palm R. (1999), L'analyse discriminante décisionnelle: principes et application, Notes stat. Inform. Gembloux, Vol. 99, n°4.

Parmentier P. (1994), La réussite des études universitaires: facteurs structurels et processuels de la performance académique en première année en médecine (thèse), Faculté de Psychologie et des Sciences de l'Éducation, Université Catholique de Louvain.

Pawlak Z. (1991), Rough sets: Theoretical Aspects of Reasoning about Data, Kluwer academic publisher.

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

Rakotomalala R. (1997), Graphes d'induction (thèse), Université Claude Bernard, Lyon I.

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

Vandamme J.-Ph., Meskens N., Artiba A. (2004), La réussite universitaire : méthodes et outils d'analyse. The first international congress on quality management education and training systems, Rabat.

"4eMka System - a rule system for multicriteria decision support integrating dominance relation with rough approximation". (2000), Laboratory of Intelligent Decision Support Systems, Institute of Computing Science, Poznan University of Technology. http://www-idss.cs.put.poznan.pl/