

**MODÈLE DICHOTOMIQUE DE LA PERSÉVÉRANCE EN ÉTUDES
UNIVERSITAIRES DES BACHELIERS DE LA PROMOTION 1996 INSCRIT EN
CYCLE COURT**

***Akila SID:**

ENSTA, Dergana

Email : akila.sid@ensta.dz

Abdelouahab LATRECHE :

ENSSEA, Koléa

Email : Latreche.abdelouahab@enssea.edu.dz

Simplice Dossou-Gbété

LMA-PUPPA : Laboratoire des Mathématiques Appliquées de Pau

Email : Simplice.dossou-gbete@univ-pau.fr

Reçu le: 13/05/2025 **Accepté le :** 28/05/2025 **Publication en ligne le:** 03/06/2025

RESUME :

Le système d'orientation des étudiants en vigueur en Algérie et qui a été promulgué en 1989 répartit les bacheliers à la fin de leurs études secondaires dans les différents secteurs et établissements de l'enseignement supérieur algérien, sur la base de la moyenne obtenue au baccalauréat, la série du baccalauréat et le genre. Notre objectif dans cette étude est de mesurer l'efficacité de ce système ainsi que ses règles en vigueur, à l'aide de méthodes statistiques appropriées dont les méthodes de régression paramétrique et semi-paramétrique adéquates à cette situation, notamment le modèle logit, et le modèle de régression par arbre binaire conditionnel.

Les résultats obtenus nous permettront de choisir le modèle le plus performant en matière de pouvoir explicatif, par l'application des tests de dépistages de la courbe ROC, et enfin répondre aux questions : Est-ce que ce système d'orientation a été efficace en matière de rendement universitaire ? Et y-a-t-il d'autres critères qu'il faut prendre en considération afin de l'améliorer ?

Mots-clés: Le système d'orientation, la persévérance dans les études, réponse binaire, modèle paramétrique, modèle non-paramétrique.

ABSTRACT:

The student placement system currently in effect in Algeria, which was established in 1989, allocates high school graduates to various sectors and institutions of Algerian higher education based on their final high school average, the stream of their baccalaureate, and their gender. The objective of this

*Akila SID

study is to assess the effectiveness of this system and its governing rules using appropriate statistical methods, including suitable parametric and semi-parametric regression techniques for this context—specifically, the logit model and the conditional binary tree regression model.

The results obtained will allow us to identify the most effective model in terms of explanatory power, through the application of ROC curve diagnostic tests, and ultimately answer the following questions : Has this placement system been effective in terms of academic performance? And are there other criteria that should be considered to improve it?

Keywords : placement system, academic perseverance, binary response, parametric model, non-parametric model.

I. INTRODUCTION

Le système universitaire algérien s'appui à chaque entrée universitaire sur une procédure d'orientation des nouveaux bacheliers selon certains critères. La question qui se pose est ce que ces procédures d'orientation sont efficaces ou non ? et y-a-t-il d'autres critères que les autorités publiques concernées doivent prendre en considération ? Pour répondre à cette question nous avons étudié un échantillon de 472 bacheliers de la promotion 1996 inscrit en cycle court, poursuit durant quatre années d'études universitaires.

Notre objectif est d'établir un modèle statistique prédictif du succès des étudiants au trois ans d'études, en fonction des caractéristiques démographiques, sociales et le niveau des études secondaire pour évaluer les chances d'aboutir au diplôme attendu. Dans un premier temps nous tentons d'étudier la corrélation entre le résultat d'un étudiant inscrit en cycle court et ses caractéristiques démographiques, ses conditions de vie et son niveau à l'entrée universitaire. Dans ce travail nous allons établir deux approches de régression pour évaluer les chances d'un étudiant d'obtenir son diplôme au bout des trois ans prévus, à cet effet à partir de la connaissance de ses caractéristiques démographiques, les conditions de subsistances en cours d'études, ses activités extra-scolaires éventuelles.

Nous utiliserons pour améliorer la probabilité de succès au trois ans d'études une régression par arbre binaire conditionnel, qui nous donnera les variables les plus corrélées au fait de terminer ses études à temps ou non, ainsi que la régression logistique considérée théoriquement la mieux adaptée à ce phénomène, où nous illustrons leurs estimations, nous interpréterons leurs résultats, ensuite nous établirons les courbes ROC pour les deux régressions, nous les comparons pour conclure le modèle le plus performant, sur lequel porteront nos conclusions.

II. PRESENTATION DES VARIABLES DEPENDANTES ET INDEPENDANTES:

II.1. L'origine des données :

Ces données sont issues de l'enquête réalisée en 1999/2000 au sein de l'unité de recherche en évaluation de la formation universitaire et de la recherche scientifique. Sise à l'Institut National de la Planification et de la Statistique, auprès d'un échantillon de 8000 bacheliers

de l'année 1996, qui a été tiré d'une population mère d'environ 80000 bacheliers répartis sur tout le territoire national, avec un taux de sondage égal à 1/10.

La méthode utilisée est la méthode des quotas, dont l'objectif essentiel est de donner à l'échantillon la même structure que la population mère. La méthode de tirage utilisée est le tirage pas à pas. Les réponses ont atteint 3089 questionnaires, donc un taux de réponse d'environ 40%.

II.2. La variable réponse :

La variable à expliquer est notée y, tel que :

$$y_i = \begin{cases} 1 & \text{si } \textit{diplômé après les trois ans} \\ 0 & \textit{sinon} \end{cases} \quad i = 1, \bar{n}$$

II.3. Les variables exogènes :

Tableau n° 1 : Description statistique de la variable indicatrice du résultat et des caractéristiques descriptives des étudiants

Types de descripteur	Dénomination	Modalités	Proportion	Int. de confiance
Le résultat au bout des trois ans Socio-démographique	Diplômé	Oui		
		Non		
	Genre	Masculin	0.502	[0.456 , 0.548]
		Féminin	0.498	[0.452 , 0.544]
	Age de l'étudiant	[16 , 18[	0.0437	[0.027 , 0.067]
		[18 , 19[	0.310	[0.268 , 0.355]
		[19 , 20[	0.320	[0.277 , 0.364]
		[20 , 21[	0.223	[0.186 , 0.264]
		[21 , 41[	0.103	[0.076 , 0.134]

Académique	Situation matrimoniale	Célibataire	0.934	[0.907 , 0.954]
		Marié	0.066	[0.046 , 0.093]
	Nombre de frères et sœurs	[0 , 2[	0.004	[0.000 , 0.015]
		[2 , 4[	0.077	[0.054 , 0.104]
		[4 , 6[	0.254	[0.215 , 0.296]
		[6 , 8[	0.294	[0.253 , 0.338]
		[8 , 10[	0.230	[0.193 , 0.271]
		10 ou plus	0.140	[0.110 , 0.175]
	Série du baccalauréat	Sciences exactes	0.080	[0.057 , 0.109]
		Sciences de la nature et de la vie	0.373	[0.330 , 0.420]
		Lettres	0.108	[0.081 , 0.140]
		Sciences économiques	0.256	[0.217 , 0.298]
		Technologie	0.182	[0.148 , 0.220]
	Moyenne au baccalauréat	[6 , 9[	0.084	[0.061 , 0.114]
		[9 , 12[	0.896	[0.865 , 0.922]
		[12 , 13[	0.020	[0.008 , 0.036]
Socio-économique	Ressources financières	Travail rémunéré	0.133	[0.103 , 0.169]
		Aide familiale	0.517	[0.470 , 0.565]
		Bourse d'Etat	0.207	[0.170 , 0.248]

Modèle Dichotomique De La Persévérance En Etudes Universitaires Des Bacheliers De La Promotion 1996 Inscrit En Cycle Court

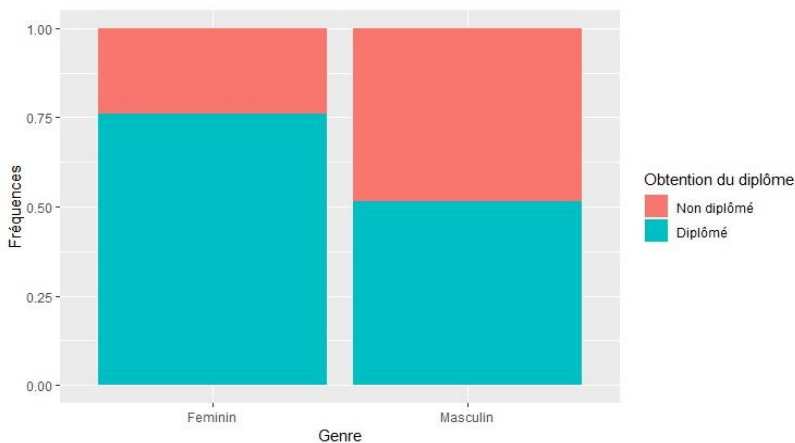
Informations sur les études		Deux ressourcesfinancières et plus	0.142	[0.111 , 0.179]
	Type d'hébergement	Cité universitaire	0.618	[0.572 , 0.662]
		Chez les parents	0.315	[0.273 , 0.360]
		Autre	0.670	[0.046 , 0.093]
	Recherche d'emploi	Oui	0.370	[0.323 , 0.416]
		Non	0.631	[0.584 , 0.676]
	Réunion d'information avec conseiller d'orientation	Jamais	0.740	[0.697 , 0.780]
		Au moins une fois	0.260	[0.220 , 0.302]
	Réunion d'information avec enseignants du supérieur	Jamais	0.886	[0.853 , 0.913]
		Une fois	0.073	[0.051 , 0.100]
		Plusieurs fois	0.030	[0.017 , 0.050]
		Autre	0.010	[0.003 , 0.025]
	Recueil d'information auprès d'un établissement	Oui	0.796	[0.755 , 0.830]
		Non	0.204	[0.169 , 0.247]
	Lecture de placard publicitaire portant	Oui	0.188	[0.153 , 0.230]

Orientation administrative	conditions d'inscription universitaire	Non	0.811	[0.770 , 0.847]
	Renseignement auprès de proches	Oui	0.797	[0.756 , 0.833]
		Non	0.203	[0.167 , 0.243]
	Affectation initiale en cycle d'études			[0.830 , 0.893]
		Court	0.864	
		Long	0.136	[0.107 , 0.170]
	Conformité de l'affectation initiale en cycle d'études	Oui	0.330	[0.288 , 0.375]
	au vœu	Non	0.670	[0.625 , 0.711]

II.4. Analyse des corrélations entre le résultat des études et les caractéristiques descriptives de l'étudiant :

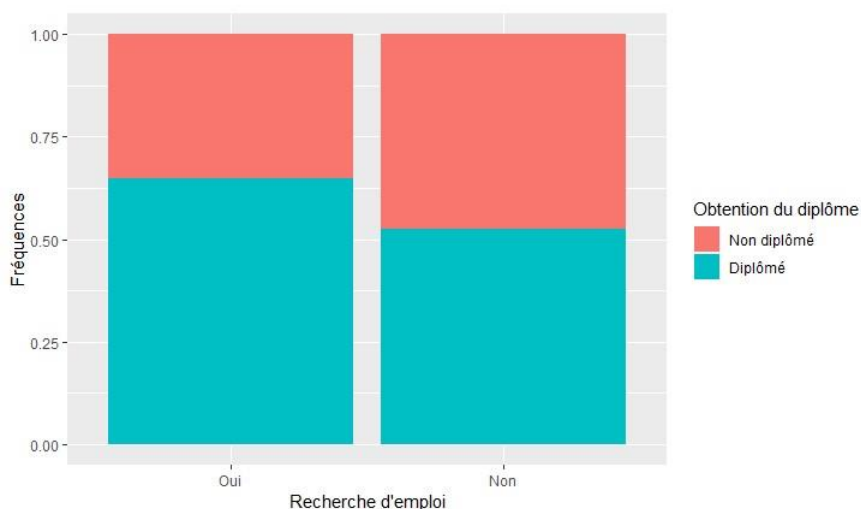
Parmi les caractéristiques retenues dans cette enquête pour décrire les bacheliers seules deux d'entre elles présente une corrélation avec la réussite au bout de trois ans : le genre et la recherche d'un emploi salarié:

Figure 1 : Répartition des étudiants selon l'obtention du diplôme et le genre



La valeur de la statistique du chi2 associée à la table croisant la variable genre et l'obtention du diplôme au bout de trois ans est 22.95 et la p-valeur associée est égale à 1.6510-6. il en résulte une corrélation significative entre les deux variables ce qui implique que la proportion d'étudiantes qui aboutit au diplôme au bout de trois ans est significativement différente de celle des étudiants. 59.47 % des étudiantes obtiennent leur diplôme au bout de trois ans tandis que chez les étudiants cette proportion chute à 40.53 %.

Figure 2 : Répartition des étudiants selon l'obtention du diplôme et la recherche d'emploi



Le test du chi-deux étant significatif conduit à conclure l'existence d'une corrélation significative entre la recherche d'un emploi salarié et l'obtention du diplôme en trois ans.

($\chi^2=8.77$, p-valeur=0.003). La proportion de succès au bout de trois ans est de 67,97% chez les étudiants qui ont choisi de consacrer la totalité de leurs temps aux études. Cette proportion passe à 32,03 %, soit une différence de 35,97 %, chez les étudiants qui ont une activité professionnelle.

La valeur de la statistique du chi2 est inférieure à 14.41 lorsque les autres prédicteurs sont croisés avec la variable réponse soit une p-valeur supérieur à 0.12 il en résulte que ces covariables ne présentent pas de corrélations significatives avec la variable réponse. On peut donc s'attendre que ces covariables ne contribuent pas à la formulation d'une fonction de prévision de résultat d'un étudiant au bout des trois années d'études nécessaires à la conclusion de sa formation.

III. MODELISATION DU PHENOMENE :

III.1. Régression par arbre binaire conditionnel :

Nous nous intéressons aux modèles de régression décrivant la distribution conditionnelle de la variable dépendante Y , étant donné m covariables (Variables indépendantes ou explicatives), notées par $X = (X_1, \dots, X_m)$, à travers un partitionnement récursif structuré en arborescence. La variable à expliquer Y est dichotomique. Le vecteur de covariables m -dimensionnel $X = (X_1, \dots, X_m)$. Nous supposons que la distribution conditionnelle $D(Y | X)$ dépend d'une fonction f des covariables, dans ce travail f est la loi logistique :

$$D(Y|X) = D(Y|X_1, \dots, X_m) = D(Y/f(X_1), \dots, f(X_m))$$

Où nous nous limitons aux relations de régression basées sur les partitions, c-à-d :

Les nœuds $B_1, \dots, B_r$ divisant l'espace covariable, noté : $F = F_1 \times \dots \times F_m = \bigcup_{k=1}^r B_k$.

Un modèle de régression de la relation doit être ajusté en fonction de nos données, c-à-d un échantillon aléatoire de n observations indépendantes et identiquement distribuées, avec éventuellement certaines covariables X_{ji} manquantes,

$$L_n = \{(Y_i, X_{1i}, \dots, X_{mi}); i = 1, \dots, n\}.$$

Un algorithme générique pour le partitionnement binaire récursif de notre échantillon L_n peut être formulé en utilisant des pondérations de nombre entier, noté : $w = (w_1, \dots, w_n)$. Chaque nœud d'un arbre est caractérisé par un vecteur de poids d'individu ayant des éléments non nuls lorsque les observations correspondantes sont des éléments du nœud et sont nuls dans le cas contraire. L'algorithme générique suivant implémente le partitionnement binaire récursif : Pour la prédiction, nous nous intéressons généralement à une estimation de l'espérance de la variable dépendante $E(Y | X = x)$ dans chaque nœud; une estimation peut être obtenue par

$$\hat{E}(Y | X = x) = \left(\sum_{i=1}^n w_i(x) \right)^{-1} \sum_{i=1}^n w_i(x) Y_i,$$

tel que $w_i(x) = w_i$ quand x est un élément du même nœud terminal que la i^{eme} observation, et $w_i(x) = 0$ sinon.

III.2. Algorithme de régression par arbre binaire conditionnel :

Etape 01 : Critère d'arrêt

Tester l'hypothèse nulle globale H_0 d'indépendance entre Y et toutes les X_j avec

$$H_0 = \bigcap_{j=1}^m H_0^j : D(Y/X_j) = D(Y)$$

Si H_0 n'est pas rejetée $\Rightarrow$ Arrêt

Etape 02 : Sélectionnons la variable X_{j^*} ayant la plus forte dépendance

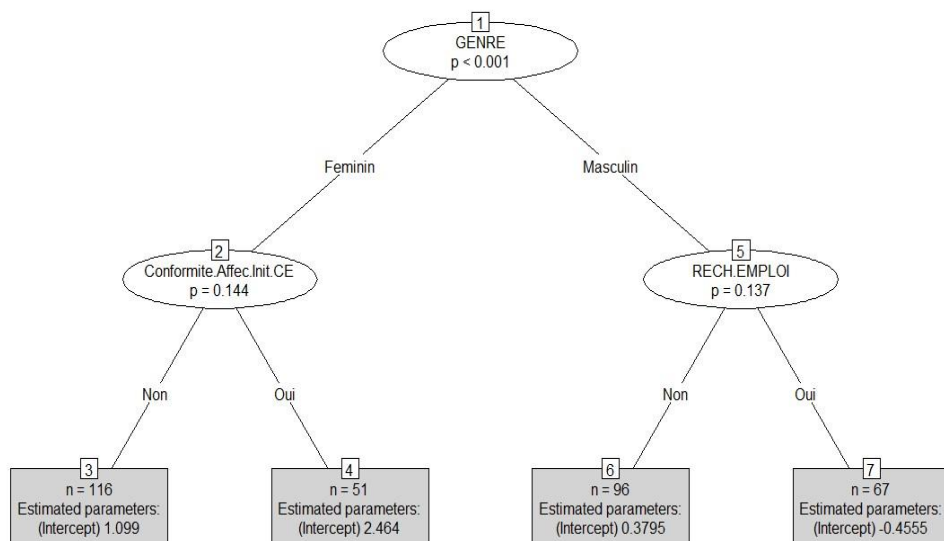
Etape 03 : rechercher le meilleur point de partage pour X_{j^*} et partitionner les données

Etape 04 : répétez les étapes 1, 2 et 3 pour les deux nouvelles partitions

III.3. Estimation de l'arbre binaire conditionnel :

Nous avons appliqué la fonction `glmtree` du package `Party` sous le logiciel R, qui donne des divisions récursives binaires conditionnelles, sur la base d'une fonction de lien logit. En utilisant le critère d'information d'Akaike l'arbre optimal est obtenue avec un seuil de signification $\alpha = 0.05$, et une taille minimale dans chaque noeud ls égale à 40 observations.

Figure 03 : Régression par arbre binaire conditionnel



III.3.1. Interprétation des résultats :

Nous avons estimé l'arbre binaire conditionnel optimal en utilisant le critère d'information d'Akaike (AIC), le résultat a été représenté par le graphique en dessus (voir figure n°3). Ce graphique révèle une première division par rapport au genre, c-à-d la variable la plus corrélée au fait d'être diplômé en troisième année est le genre. Il en résulte deux nœuds fils, le premier représenté par les filles qui subira une division supplémentaire par rapport à la conformité de l'affectation initiale en cycle d'études. Le second nœud fils représentant les garçons subit quant à lui une division par rapport à la recherche d'emploi.

III.3.2. Caractéristiques des nœuds estimés :

Nous listons dans ce qui suit les différents nœuds terminaux obtenus via l'arbre binaire conditionnel, ainsi que leurs caractéristiques :

- 1- Le nœud terminal n°3 de taille d'échantillon $n_3 = 116$, caractérisé par la distribution marginale :

$$D_3(Y/\text{Féminin}, \text{l'affectation initiale en cycle d'études n'est pas conforme au vœu})$$

a enregistré une probabilité d'être diplômée de 0.75.

- 2- Le nœud terminal n°4 de taille d'échantillon $n_4 = 51$, caractérisé par la distribution marginale :

$$D_4(Y/\text{Féminin}, \text{l'affectation initiale en cycle d'études est conforme au vœu})$$

a enregistré une probabilité d'être diplômée de 0.92.

- 3- Le nœud terminal n°6 de taille d'échantillon $n_6 = 96$, caractérisé par la distribution marginale :

$$D_6(Y/\text{Masculin}, \text{Ne recherche pas d'emploi})$$

a enregistré une probabilité d'être diplômé de 0.59.

1. Le nœud terminal n°7 de taille d'échantillon $n_7=67$, caractérisé par la distribution marginale :

$$D_7(Y/\text{Masculin}, \text{Recherche de L'emploi})$$

a enregistré une probabilité d'être diplômé de 0.39.

III.4. La régression logistique simple :

Ces modèles ont été initialement utilisés pour les études biologiques, mais ont eu un champ d'application très vaste. Dans notre étude nous utilisons ces modèles pour déterminer la probabilité d'être diplômé après trois ans d'études. Nous mettrons en relation le fait qu'un étudiant inscrit en cycle court termine ces études en troisième année ou non, exprimée par y

avec un certain nombre de variables explicatives, telles que le genre, la moyenne au Baccalauréat, la série du Baccalauréat, etc.

La variable endogène y_i supposée être le résultat d'un modèle latent qui porte sur une variable continue, notée y_i^* , cette dernière inobservable est décrite par le modèle linéaire standard suivant :

$$y_i^* = z_i\theta + \varepsilon_i$$

(Voir le livre -Alban Thomas-Econométrie des variables qualitatives-page 52-)

où ε_i est une perturbation d'espérance nulle, sans perte de généralité tant que le modèle latent contient un terme constant. Pour pouvoir estimer ce modèle par le maximum de vraisemblance, il nous faut écrire la loi de la variable observable conditionnellement aux variables explicatives. Cette variable observable est définie par :

$$y_i = \begin{cases} 1 & \text{si } z_i\theta + \varepsilon_i > 0 \\ 0 & \text{si } z_i\theta + \varepsilon_i < 0 \end{cases}$$

Le fait de prendre la valeur 0 comme seuil de référence n'a aucune incidence sur les estimations tant que le modèle comporte un terme constant, car on peut alors utiliser la variable latente $y_i^* - c$, où c'est le seuil de troncature constant.

III.4.1. Expression de La probabilité de l'évènement :

Notons que :

$$y \rightarrow B(P) \Rightarrow P(y = y_i) = P^{y_i}(1 - P)^{1-y_i}, i = \overline{1, n}, \text{ avec } P \in [0, 1],$$

Il en résulte que : y est une variable aléatoire discrète de dimension 2, suivant une loi de Bernoulli de paramètre P , avec $P \in [0, 1]$, et $Q = 1 - P$. Lorsque y_i prend la valeur 1, cela signifie que le bachelier a persévéré dans ses études, il a eu son diplôme à temps, et quand $y_i = 0$, signifie que le bachelier n'a pas été diplômé à temps, donc il a fait du retard ou il a abandonné, nous avons $P(y_i = 1) = P$. Notons :

$$y \rightarrow B(P)$$

Nous avons aussi La probabilité conjointe exprimant la probabilité de l'état d'un étudiant à la troisième année, sera :

$$P(y = y_i) = P^{y_i} \cdot (1 - P)^{(1-y_i)}, i = \overline{1, n}$$

La probabilité exprimant que l'étudiant soit diplômé en troisième année ou non, sera :

$$P^{y_i} \cdot (1 - P)^{(1-y_i)}, i = \overline{1, n}$$

Nous posons :

$$P_i(y_i = 1/z_i) = F(z_i\theta) = \frac{e^{z_i\theta}}{1 + e^{z_i\theta}}$$

Telle que F est la fonction de répartition des ε_i , une fonction monotone, deux fois dérivables à valeurs dans $[0,1]$

Notons par z_i le vecteur des variables explicatives exprimant les critères d'orientation à l'université tels que la série du baccalauréat, la moyenne au baccalauréat, le genre, ainsi que l'âge de l'étudiant, etc. :

$$z_i = (z_{i1}, z_{i2}, \dots, z_{ik}) \text{ avec } z_{i1} = 1$$

et θ le vecteur des paramètres à estimer, avec $j = \overline{1, k}$, nous aurons:

$$\theta = (\theta_1, \theta_2, \dots, \theta_k)'$$

Si F désigne la fonction de répartition de $X - z_i\theta$ conditionnellement à z et $z_i\theta$ est la valeur du paramètre de localisation de la loi nous obtiendrons que $Pr(X \leq z_i\theta) = F(X - z_i\theta)$, ceci veut dire que nous faisons l'hypothèse que la loi de X appartient à une famille paramétrique de lois de probabilité dont les éléments sont identiques par deux paramètres : μ le paramètre de localisation et σ le paramètre d'échelle. $\mu = 0$ et $\sigma = 1$ désigne l'élément générateur (ou loi de référence) de la famille. F peut donc être la fonction de répartition de la loi gaussienne centrée et réduite, la loi logistique standard, etc.

III.4.2. Estimation de la probabilité d'être diplômé ou non après trois ans:

Nous avons vu dans ce qui précède que la probabilité d'être diplômé après trois ans est donnée par l'équation suivante :

$$P_i = F(z_i\theta)$$

alors P_i est forcément fonction de θ . D'où l'expression :

$$\hat{P}_i = F(z_i\hat{\theta})$$

Pour i fixé, posons :

$$\hat{P}_{ir} = F(z_i\hat{\theta}_r)$$

Passons la limite quand r tend vers l'infini.

$$\hat{P}_i = \lim_{r \rightarrow \infty} \hat{P}_{ir} = \lim_{r \rightarrow \infty} F(z_i\hat{\theta}_r) = F(z_i\hat{\theta})$$

**Modèle Dichotomique De La Persévérance En Etudes Universitaires Des
Bacheliers De La Promotion 1996 Inscrit En Cycle Court**

III.5. Estimation du modèle :

III.6. Modèle introduisant toutes les variables explicatives :

III.6.1. Le modèle logit complet :

Tableau n°2 Le modèle logit complet

	Estimate	Std. Error	t value	Pr(> t)	Significativité
(Intercept)	1.0157	0.2961	3.43	0.0007	***
GENREMasculin	- 0.0525	0.0468	- 1.12	0.2629	
MOYb[9,12)	- 0.3818	0.1645	- 2.32	0.0210	*
MOYb[12,13)	- 0.5679	0.1859	- 3.06	0.0025	**
SMMarie	- 0.0220	0.0966	- 0.23	0.8201	
TYP.HEBERG.PEChez.parents	0.0080	0.1005	0.08	0.9362	
TYP.HEBERG.PECite.univ	0.0155	0.0968	0.16	0.8730	
RES.FIN.PEAide.fam	0.0362	0.0640	0.57	0.5722	
RES.FIN.PEBour.Etat	0.0304	0.0757	0.40	0.6877	
RES.FIN.PETrav.rem	0.0146	0.0727	0.20	0.8410	
NFS[10,88)	0.1371	0.2017	0.68	0.4974	
NFS[2,4)	0.1613	0.2037	0.79	0.4292	
NFS[4,6)	0.1989	0.1954	1.02	0.3095	
NFS[6,8)	0.1331	0.1956	0.68	0.4966	
NFS[8,10)	0.0844	0.1943	0.43	0.6644	
SERIEBACSc.Eco	0.1393	0.0721	1.93	0.0542	.

SERIEBACSc.Ex	0.2648	0.0901	2.94	0.0036	**
SERIEBACSc.N.V	0.1837	0.0559	3.29	0.0011	**
SERIEBACTechnologie	0.1258	0.1171	1.07	0.2834	
AGETU[18,19)	-	0.1040	-	0.2865	
	0.1111		1.07		
AGETU[19,20)	-	0.1038	-	0.6286	
	0.0503		0.48		
AGETU[20,21)	-	0.1112	-	0.4389	
	0.0862		0.78		
AGETU[21,41)	-	0.1232	-	0.1370	
	0.1837		1.49		
Reu.Inf.EnseignantSupPlusfois	0.1913	0.1300	1.47	0.1422	
Reu.Inf.EnseignantSupUnefois	-	0.0795	-	0.9214	
	0.0078		0.10		
Reu.Inf.ConseillerOrientationJamais	0.0083	0.0492	0.17	0.8668	
Receuil.Inf.EtabOui	-	0.0523	-	0.8212	
	0.0118		0.23		
LPlacardPubPCIOui	-	0.0532	-	0.3174	
	0.0533		1.00		
RenseignementAProchesOui	-	0.0566	-	0.5633	
	0.0328		0.58		
A ectation.Init.CELong	-	0.0466	-	0.0000	***
	0.5983		12.85		
Conformite.A ec.Init.CEOui	-	0.0439	-	0.3048	
	0.0451		1.03		
RECH.EMPLOIOui	-	0.0455	-	0.0143	*
	0.1120		2.46		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1275087)

Null deviance: 76.154 on 311 degrees of freedom Residual deviance: 35.702 on 280 degrees of freedom (201 observations deleted due to missingness) AIC: 275.07

Number of Fisher Scoring iterations: 2

A travers les résultats du modèle estimé nous avons constaté qu'une bonne partie des variables explicatives introduites n'est pas significative, par conséquent nous avons procédé à une sélection des plus pertinentes par la méthode Stepwise.

III.7. Modèles introduisant les variables sélectionnées par la Stepwise :

III.7.1. La régression ascendante pas à pas (stepwise) :

La question qui se pose généralement dans une régression quel est le plus petit nombre de variables explicatives X qui explique au mieux la variabilité de Y .

Une méthode courante est une régression itérative qui inclut d'abord dans le modèle la variable qui propose le critère d'information d'Akaike minimum. Ensuite, celle qui améliore le plus ce dernier et ainsi de suite.

Alternativement, toutes les variables sont entrées dans le modèle et les variables sont progressivement exclues, en fonction de celles qui contribuent le moins au modèle.

Il faut noter que la seconde variable qui entre dans le modèle n'est pas forcément celle qui présente, à elle seule, le second meilleur critère d'information d'Akaike avec Y . Sinon, la solution serait triviale. En effet, X_1 et X_2 peuvent être très corrélées, voire quasi redondantes. Dans ce cas la qualité du modèle ne sera pas améliorée. C'est donc la variable qui contribue le plus à réduire la variabilité résiduelle, du modèle en voie d'élaboration qui sera sélectionnée chaque étape.

III.7.2. Le modèle logit réduit :

Après avoir introduit toutes les variables dans le modèle précédent les plus pertinentes sont le genre, la moyenne au Baccalauréat, la série du Baccalauréat, l'affectation initiale en cycle d'études long et la recherche d'emploi pendant les études.

Tableau n°3 Résultats de l'estimation du modèle logit réduit

	Estimate	Std. Error	z value	Pr(> z)	Significativité
(Intercept)	3.2481	1.4743	2.20	0.0276	*
Affectation.Init.CELong	-3.4048	0.3728	-9.13	0.0000	***
SERIEBACSc.Eco	1.4613	0.5507	2.65	0.0080	**
SERIEBACSc.Ex	2.1485	0.6843	3.14	0.0017	**
SERIEBACSc.N.V	1.8093	0.4916	3.68	0.0002	***
SERIEBACTechnologie	0.8245	0.8536	0.97	0.3341	
RECH.EMPLOIOui	-1.1466	0.3772	-3.04	0.0024	**
MOYb[9,12)	-2.9664	1.3883	-2.14	0.0326	*
MOYb[12,13)	-4.6616	1.6241	-2.87	0.0041	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 425.11 on 311 degrees of freedom Residual deviance: 237.29 on 303 degrees of freedom (201 observations deleted due to missingness) AIC: 255.29

Number of Fisher Scoring iterations: 5

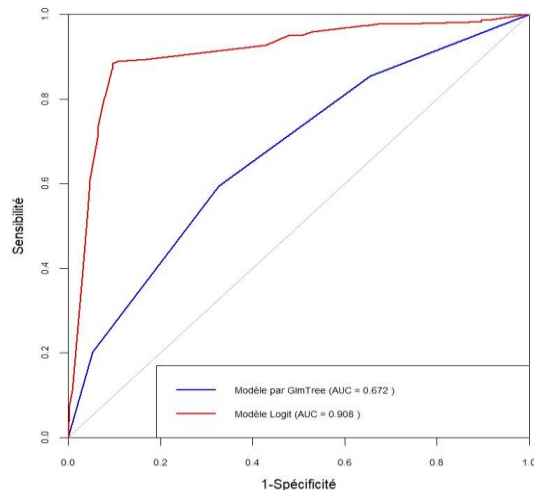
Donc pour le choix du modèle nous nous sommes basé sur le critère d'Akaike, c-à-d nous avons gardé le modèle dont le AIC est le plus faible (AIC=255.29).

III.7.3. Interprétation des résultats :

D'après le tableau n°3 les modalités qui jouent en défaveur de l'achèvement des études au bout des trois ans, sont : L'affectation initiale en cycle d'études "long", "la recherche d'emploi pendant les études" et les moyennes aux baccalauréat variant entre 9 et 13, toutes ces modalités ont des coefficients négatifs dans le modèle. Tandis que celles qui augmentent la probabilité d'être diplômé au bout des trois ans et enregistrant des coefficients positifs dans le modèle sont : les séries de baccalauréat "Sciences exactes", "Sciences de la Nature et de la Vie" et "Sciences économiques", ainsi que l'existence d'autres facteurs qui n'ont pas été introduit dans le modèle ce que révèle la constante.

III.8. Comparaison des deux modèles en utilisant les courbes ROC:

Figure n°4: Courbes ROC des modèles par arbre binaire conditionnel et Logit



Le graphique ci-dessus présente les courbes ROC (Receiver Operating Characteristic) pour deux modèles de classification appliqués aux mêmes données : un modèle GlmTree (en bleu), basé sur un arbre conditionnel avec modélisation logistique dans les nœuds et un modèle

Logit (en rouge), basé sur une régression logistique classique. La performance de chaque modèle est mesurée à l'aide de l'aire sous la courbe ROC (AUC) :

Le modèle Logit atteint une AUC de 0.908, ce qui reflète une excellente capacité de discrimination entre les deux classes (Diplômé au bout de trois ans vs Non diplômé).

Le modèle GlmTree obtient une AUC de 0.672, indiquant une performance modérée, certes meilleure que le hasard (AUC = 0.5), mais sensiblement inférieure à celle du modèle Logit.

La courbe du modèle Logit est visuellement plus proche du coin supérieur gauche du graphique (sensibilité élevée pour une faible proportion de faux positifs), ce qui confirme sa supériorité en termes de précision de classification.

Afin d'évaluer si la différence observée entre les deux modèles (GlmTree et Logit) est statistiquement significative, le test de DeLong a été appliqué. Ce test compare directement les aires sous les courbes ROC (AUC) des deux modèles sur les mêmes données. Résultat du test :

1. Différence de performance : ~0.2369
2. Statistique de test D : -8.686
3. Degrés de liberté : 727.14
4. p-value : $< 2.2 \times 10^{-16}$

La p-value extrêmement faible (< 0.001) indique que la différence entre les deux courbes ROC est hautement significative. Nous pouvons donc rejeter l'hypothèse nulle selon laquelle les deux modèles auraient des performances identiques. En conclusion, le modèle Logit est significativement meilleur que le modèle GlmTree en termes de capacité à discriminer entre les deux classes.

3.8.1- Estimation de la probabilité d'être diplômé ou non après trois ans:

Nous choisissons donc le modèle Logit pour nos prévisions, nous avons vu dans ce qui précède que la probabilité d'être diplômé après trois ans est donnée par l'équation suivante :

$$P_i = F(z_i\theta)$$

alors $\hat{P}_i$ est forcément fonction de $\hat{\theta}$. D'où l'expression :

$$\hat{P}_i = F(z_i\hat{\theta})$$

$$\Rightarrow \hat{P}_i(y_i = 1/z_i) = F(z_i\hat{\theta}) = \frac{e^{z_i\hat{\theta}}}{1 + e^{z_i\hat{\theta}}} = 1 - \frac{1}{1 + e^{z_i\hat{\theta}}}$$

Application:

D'après le modèle estimé nous avons:

$$\Rightarrow \hat{P}_i(y_i = 1/z_i) = F(z_i\hat{\theta}) = \frac{e^{z_i\hat{\theta}}}{1 + e^{z_i\hat{\theta}}} = 1 - \frac{1}{1 + e^{z_i\hat{\theta}}}$$

$$\Rightarrow \hat{P}_i(y_i = 1/z_i) = \frac{e^{(3.25-3.40.Z_{1i}+1.46.Z_2+2.14.Z_3+1.80.Z_4+0.82.Z_5-1.14.Z_6-2.96.Z_7-4.66.Z_8)}}{1 + e^{(3.25-3.40.Z_{1i}+1.46.Z_2+2.14.Z_3+1.80.Z_4+0.82.Z_5-1.14.Z_6-2.96.Z_7-4.66.Z_8)}}$$

Tableau n°4 Notation des variables utilisées dans le modèle Logit dans la prévision

Variables	tetha chapeau	Notation dans l'application
(Intercept)	3.2481	tetha0
Affectation.Init.CELong	-3.4048	Z1
SERIEBACSc.Eco	1.4613	Z2
SERIEBACSc.Ex	2.1485	Z3
SERIEBACSc.N.V	1.8093	Z4
SERIEBACTechnologie	0.8245	Z5
RECH.EMPLOIOui	-1.1466	Z6
MOYb[9,12)	-2.9664	Z7
MOYb[12,13)	-4.6616	Z8

Exemples de situations hypothétiques:

- Un étudiant qui a été affecté en cycle long (Affectation.Init.CELong=1), issu de la série de Bac sciences économiques (SERIEBACSc.Eco=1), n'a pas cherché d'emploi pendant les études (RECH.EMPLOI=0), sa moyenne au Bac est entre 9 et 12 (MOYb[9,12)=1), et le reste des modalités est 0. Sa probabilité d'être diplômé pendant trois ans sera de 0.159.
- Un étudiant qui n'a pas été affecté en cycle long (Affectation.Init.CELong=0), issu de la série de Bac sciences exactes (SERIEBACSc.Ex=1), n'a pas cherché d'emploi pendant les études (RECH.EMPLOI=0), sa moyenne au Bac est inférieure a 9 c-à-d (MOYb[9,12)=0) et (MOYb[12,13)=0), et le reste des modalités est 0. Sa probabilité d'être diplômé pendant trois ans sera de 0.996.

IV. CONCLUSION:

Dans le but d'améliorer notre système d'orientation des bacheliers à chaque entrée universitaire, nous pouvons créer des bases de données récentes et actuelles qui poursuivent les bacheliers au cours de leurs études universitaires. Ce qui nous permettra de voir les facteurs influençant positivement ou négativement l'achèvement des études à temps. Selon notre étude les variables qui influencent négativement l'achèvement des études à temps sont "affectation initiale en cycle long", les moyennes au Bac entre 9 et 13, afin de diminuer l'effet négatif de ces deux variables nous pouvons prendre en considération le vœu de l'étudiant. Tandis que pour la "recherche d'emploi pendant les études" les autorités concernées peuvent améliorer les conditions de vie de l'étudiant, ainsi que son pouvoir d'achat. Par contre la série de Bac influence positivement l'obtention du diplôme ainsi que d'autres facteurs qui n'ont

pas été introduit dans notre modèle ce que révèle le coefficient de la constante. Donc les critères à prendre en considération pour une meilleure orientation des bacheliers sont: la moyenne au Bac, la série de Bac, et le niveau de vie de l'étudiant, son vœu (son choix avant inscription universitaire) pour les études (ce qu'il veut étudier) et enfin nous devons chercher d'autres facteurs qui peuvent influencer la réussite des études, et empêchent les étudiants de persévérer dans leurs études d'où un mauvais rendement universitaire.

V. REFERENCES BIBLIOGRAPHIQUES :

- 1- Agresti Alan Catégorical data analysis. Second édition. University of Florida, Gaines ville, Florida.
- 2- Alain Jaquot. Les modèles économétriques logit-probit-tobit. CNAF Bureau des prévisions
- 3- Alban Thomas-Econométrie des variables qualitatives-Collection Eco-Sup-Edition : Dunod, 2000.
- 4- Chelsea Moore. An Introduction to Logistic and Probit Regression Models
- 5- Hugo Harari-Kermadec. Econometrie 2 : donnees qualitatives, probit et logit. URCA 2008-2009
- 6- Lingjian Yang a , Songsong Liu a , b , Sophia Tsoka c , Lazaros G. Papageorgiou a . A regression tree approach using mathematical programming, a Centre for Process Systems Engineerig, Department of Chemical Engineering, University College London, Torrington Place, London WC1E 7JE, UK b School of Management, Swansea University, Bay Campus, Fabian Way, Swansea SA1 8EN, UK c Department of Informatics, King's College London, Strand, London WC2R 2LS, UK
- 7- Ludwig-Maximilians , James Malley , Gerhard Tutz. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests Carolin Strobl -Universita't Munich National Institutes of Health Gerhard Tutz Ludwig-Maximilians-Universita't Munich