



FOUILLE DE DONNEES

OPTION : Systeme Intelligent et Multi-média

Projet Final Fouille de données

Auteurs :

Saidi DAOUDA KADRI
Abdoul-Djalil O.HAMZA

Encadrant :

Dr. Nguyen Thi MINH HUYEN

Promotion XXII
3 décembre 2018

Table des matières

1	INTRODUCTION	2
2	DESCRIPTION DES ATTRIBUTS	2
3	ANALYSE EXPLORATOIRE	2
3.1	Attributs Continues	2
3.1.1	attribut age	3
3.1.2	hours-per-week	4
3.2	Attributs Discrets	5
4	ANALYSE DES LIENS ENTRE CHAQUE PAIRES D'ATTRIBUTS	6
4.1	Paires attributs continues	6
4.2	Paires attributs discrets	8
5	APPLICATION DES MÉTHODES FACTORIELLES A NOTRE JEU DE DONNÉES	9
5.1	Analyse en Composantes principale (ACP)	9
5.2	Valeurs propres	9
5.3	Corrélation entre les valeurs et les axes principaux	10
5.4	Plans factoriel	11
5.5	Cercle des corrélations	12
6	CONTEXTE ET ENONCE DU PROBLEME	12
6.1	Contexte	12
6.2	Enoncé du probleme	13
7	PRESENTATION DE LA METHODE	14
7.1	Representation d'un arbre CART	15
8	APPLICATION DE LA METHODE A NOS DONNEES	16
8.1	Apprentissage avec la methode C-RT	16
8.2	Evalution sur l'echantillon test	18
8.3	Courbe d'erreur en fonction de la complexité de l'arbre	19
8.4	Performance de l'arbre 0-SE RULE	21
9	PRESENTATION D'AUTRES METHODES POUR LA COMPARAISON	23
9.1	Pretraitement des données	23
9.2	Linear Disriminat Analysisie (LDA)	23
9.3	Support Vector Machine (SVM)	25
9.4	Comparaison des trois méthodes étudiées	26
10	Conclusion	27

1 INTRODUCTION

Nous avons choisi de travailler sur l'ensemble de données « Adult » dans le but de prédire si le revenu annuel des fonctionnaires d'un état donné est inférieur ou égale à 50k ou supérieur à 50k (≤ 50 ou > 50 k). Nous choisissons le cas des Etats-Unis d'Amérique.

2 DESCRIPTION DES ATTRIBUTS

Notre jeu de données « Adult » est composé de **32561 individus**, des valeurs manquantes et de 15 attributs dont 6 attributs continus et 9 attributs discrets. Ainsi, Nous avons les **attributs continus** qui sont :

1. age : âge de la personne
2. fnlwgt : poids final d'échantillonnage
3. education-num : le plus haut niveau d'éducation atteint sous forme numérique
4. capital-gain : gain en capital pour un individu
5. capital-loss : perte en capital pour un individu
6. hours-per-week : le nombre d'heure de travailler par semaine fonctionnalité ,forme des prédicateurs.

Et aussi des **attributs discrets** :

1. workclass : représenter le statut d'emploi d'un individu < Prive ,Self-emp-not-inc,Self-empinc- Federal-gouv,local-gouv,Etat-gouv,sans paye, jamais-travaille
2. education : Niveau d'éducation le plus élevé atteint par un individu
3. marital-status : Marié-conjoint-conjoint, Divorcé, Jamais-marié, Séparé, Veuf, Marié-conjointabsent, Marié-conjoint
4. occupation : Type générale d'occupation d'une personne
5. relationship : situation matrimonial d'une personne (épouse, propre enfant, mari, non en famille, autre parent, célibataire).
6. race : race d'un individu (Blanc, Asiatique-Pac-Insulaire, Amer-Indien-Esquimau, Autre, Noir).
7. sexe : le sexe de la personne Femme, Homme.
8. ative-contry : Permet de connaitre le pays d'origine d'une personne États-Unis, Cambodge, Angleterre, Porto
9. label : Salaire d'un particulier est plus de 50k par année.

3 ANALYSE EXPLORATOIRE

3.1 Attributs Continues

Nous allons utiliser la statistique descriptive détaillée pour la description des variables continues. Parmi les six (6) attributs continus, nous avons jugé utile de faire l'analyse sur deux cas :

l'attribut age et l'attribut hours-per-week c'est-à-dire le nombre d'heure de travail par semaine.

3.1.1 attribut age

Results					
Attribute	Stats		Histogram		
age	Statistics		Values	Count	Percent
	Average	38.8743	x_<_24.3000	534	16.37%
	Median	38.0000	24.3000_=<_x_<_31.6000	582	17.84%
	Std dev. [Coef of variation]	13.5410 [0.3483]	31.6000_=<_x_<_38.9000	584	17.90%
	MAD [MAD/STDDEV]	11.1063 [0.8202]	38.9000_=<_x_<_46.2000	652	19.99%
	Min * Max [Full range]	17.00 * 90.00 [73.00]	46.2000_=<_x_<_53.5000	419	12.84%
	1st * 3rd quartile [Range]	28.00 * 48.00 [20.00]	53.5000_=<_x_<_60.8000	257	7.88%
	Skewness (std-dev)	0.5255 (0.0429)	60.8000_=<_x_<_68.1000	160	4.90%
	Kurtosis (std-dev)	-0.1560 (0.0857)	68.1000_=<_x_<_75.4000	52	1.59%
			75.4000_=<_x_<_82.7000	16	0.49%
			x>=_82.7000	6	0.18%

FIGURE 1 – Statistique-histogramme-1

Les résultats de la figure 1 ci-dessus nous donnent les informations sur les sections suivantes :

Statistics :

- La moyenne (Average) est de 38.8743 ;
- La mediane (Median) est de 38.0000 ;
- L'ecart-type (std dev [Coef.of variation] est de 13.5410 ;
- Skewness(std-dev) 0.5255 (0.0429) ;
- Kurtosis (std-dev) est -0.1560 (0.0857).

Histogram :

D'après l'histogramme, nous pouvons constater que les individus dont l'âge est inférieure à 24.3000 sont au nombre de 534 au sein de la population avec un pourcentage de 16.37alors que ceux qui ont l'âge compris entre 38.9000 et 46.2000 sont considérablement plus nombreux (652) avec un pourcentage assez élevé de 19.99%.

3.1.2 hours-per-week

hours-per-week	Statistics		Values	Count	Percent	Histogram
	Average	40.6343	x_<_10.8000	82	2.51%	
	Median	40.0000	10.8000_=<_x_<_20.6000	185	5.67%	
	Std dev. [Coef of variation]	12.0653 [0.2969]	20.6000_=<_x_<_30.4000	229	7.02%	
	MAD [MAD/STDDEV]	7.4877 [0.6206]	30.4000_=<_x_<_40.2000	1786	54.75%	
	Min * Max [Full range]	1.00 * 99.00 [98.00]	40.2000_=<_x_<_50.0000	319	9.78%	
	1st * 3rd quartile [Range]	40.00 * 45.00 [5.00]	50.0000_=<_x_<_59.8000	397	12.17%	
	Skewness (std-dev)	0.0299 (0.0429)	59.8000_=<_x_<_69.6000	188	5.76%	
	Kurtosis (std-dev)	2.5676 (0.0857)	69.6000_=<_x_<_79.4000	48	1.47%	
			79.4000_=<_x_<_89.2000	20	0.61%	
			x>=_89.2000	8	0.25%	

FIGURE 2 – Tableau

Les resultats de la figure 2 sont les suivants :

Statistics

- La moyenne (Average) est de 40.6343 ;
- La mediane (Median) est de 40.0000 ;
- L’ecart-type (std dev [Coef.of variation] est de 12.0653 ;
- Skewness(std-dev) 0.0299 (0.0429) ;
- Kurtosis (std-dev) est 2.5676 (0.0857).

Histogram

La section histogramme, nous montre une forte concentration des individus qui ont un nombre d’heure de travail par semaine compris entre 30.4000 et 40.2000 avec un pourcentage de plus de la moitié de l’ensemble des individus (54.75%).

3.2 Attributs Discrets

Results				
Attribute	Gini	Distribution		
education	0.8093	Values	Count	Percent
		Bachelors	557	17.08 %
		HS-grad	1048	32.13 %
		11th	137	4.20 %
		Masters	179	5.49 %
		9th	49	1.50 %
		Some-college	720	22.07 %
		Assoc-acdm	100	3.07 %
		Assoc-voc	137	4.20 %
		7th-8th	63	1.93 %
		Doctorate	35	1.07 %
		Prof-school	60	1.84 %
		5th-6th	31	0.95 %
		10th	93	2.85 %
		1st-4th	17	0.52 %
		Preschool	4	0.12 %
		12th	32	0.98 %

FIGURE 3 – *Attribut education - univariate discrete stat-1*

Le taux le plus élevé dans le tableau correspond à la variable HS-grad de l'attribut education (Niveau d'éducation le plus élevé atteint par un individu) avec un pourcentage de 32.13% ce qui indique que presque tous les individus arrivent à atteindre ce niveau, par ailleurs une infime partie ont atteint le niveau doctorale.

Results					
Attribute	Gini	Distribution			
workclass	0.5140	Values	Count	Percent	Histogram
		State-gov	135	4.14 %	
		Self-emp-not-inc	257	7.88 %	
		Private	2230	68.36 %	
		Federal-gov	89	2.73 %	
		Local-gov	218	6.68 %	
		?	208	6.38 %	
		Self-emp-inc	124	3.80 %	
		Without-pay	1	0.03 %	

FIGURE 4 – *Attribut sex - univariate discrete stat-2*

Les résultats obtenus dans le tableau ci-dessus nous laisse croire que l'attribut workclass (représente les individus ayant un emploi stable) montre que la majorité des individus travaillent dans le secteur privé.

4 ANALYSE DES LIENS ENTRE CHAQUE PAIRES D'ATTRIBUTS

4.1 Paires attributs continues

Linear correlation 1

Parameters

Cross-tab parameters

Sort results

non

Input list

Target (Y) and input (X)

Results

Y	X	r	r ²	t	Pr(> t)
age	fnlwgt	-0.0950	0.0090	-5.4514	0.0000
age	education-num	0.0162	0.0003	0.9230	0.3561
age	capital-gain	0.0720	0.0052	4.1231	0.0000
age	capital-loss	0.0564	0.0032	3.2246	0.0013
age	hours-per-week	0.0547	0.0030	3.1301	0.0018

education-num	age	0.0162	0.0003	0.9230	0.3561
education-num	fnlwgt	-0.0769	0.0059	-4.4025	0.0000
education-num	capital-gain	0.1129	0.0127	6.4852	0.0000
education-num	capital-loss	0.0791	0.0063	4.5312	0.0000
education-num	hours-per-week	0.1697	0.0288	9.8293	0.0000
capital-gain	age	0.0720	0.0052	4.1231	0.0000
capital-gain	fnlwgt	-0.0026	0.0000	-0.1496	0.8811
capital-gain	education-num	0.1129	0.0127	6.4852	0.0000
capital-gain	capital-loss	-0.0338	0.0011	-1.9304	0.0536
capital-gain	hours-per-week	0.0702	0.0049	4.0184	0.0001
capital-loss	age	0.0564	0.0032	3.2246	0.0013
capital-loss	fnlwgt	-0.0142	0.0002	-0.8108	0.4175
capital-loss	education-num	0.0791	0.0063	4.5312	0.0000
capital-loss	capital-gain	-0.0338	0.0011	-1.9304	0.0536
capital-loss	hours-per-week	0.0789	0.0062	4.5192	0.0000
hours-per-week	age	0.0547	0.0030	3.1301	0.0018
hours-per-week	fnlwgt	-0.0201	0.0004	-1.1458	0.2520
hours-per-week	education-num	0.1697	0.0288	9.8293	0.0000
hours-per-week	capital-gain	0.0702	0.0049	4.0184	0.0001
hours-per-week	capital-loss	0.0789	0.0062	4.5192	0.0000

FIGURE 5 – *Evaluation de la corrélation linéaire entre chaque paire de variables continues*

Sur la figure 5, on détermine pour chaque variable, la variable qui est la plus corrélée avec elle. Le résultat obtenu est présenté sur la figure suivante :

VARIABLES	VARIABLE LA PLUS CORRELEE
education-num	capital-gain
education-num	hours-per-week
capital-gain	hours-per-week
hours-per-week	capital-loss
education-num	capital-loss
age	capital-gain

FIGURE 6 – Comparaison entre les variables la plus corrélée

4.2 Paires attributs discrets

Pour analyser les paires des variables discrets, nous avons le tableau de contingence suivant :

Results							
Row (Y)	Column (X)	Statistical indicator		Cross-tab			
education	class	Stat	Value		<=50K	>50K	Sum
		d.f.	15	Bachelors	316	241	557
		Tschuprow's t	0.180807	HS-grad	858	189	1049
		Cramer's v	0.355826	11th	129	8	137
		Phi²	0.126612	Masters	86	93	179
		Chi² (p-value)	413.14 (0.0000)	9th	48	1	49
		Lambda	0.022584	Some-college	596	124	720
		Tau (p-value)	0.0142 (0.0000)	Assoc-acdm	65	35	100
		U(R/C) (p-value)	0.0309 (0.0000)	Assoc-voc	101	36	137
				7th-8th	58	5	63
				Doctorate	9	26	35
				Prof-school	19	41	60
				5th-6th	29	2	31
				10th	87	6	93
				1st-4th	16	1	17
				Preschool	4	0	4
				12th	29	3	32
				Sum	2450	811	3263

FIGURE 7 – Tableau de contingence des variables discrets

5 APPLICATION DES MÉTHODES FACTORIELLES A NOTRE JEU DE DONNÉES

5.1 Analyse en Composantes principale (ACP)

L'Analyse en Composante principale est une méthode de la famille de l'analyse des données et d'une façon plus générale de la statistique multi-variée, qui consiste à transformer des variables liées entre elles (dites "corrélées" en statistique) en nouvelles variables indépendantes les unes des autres (donc "non corrélées"). Ces nouvelles variables sont appelées "composantes principales", ou axes. Notre objectif est de partir d'un ensemble de données contenant 32561 observations dont 6 variables continues. Nous cherchons à résumer l'information disponible à l'aide des variables synthétiques que nous appelons ici composantes principales. Nous avons réalisé l'ACP grâce à la fonction «Principal Component Analysis » située sous l'onglet « Factorial Analysis ». Les résultats obtenus sont présentés dans la suite.

5.2 Valeurs propres

Le tableau contenant les valeurs propres de la matrice de corrélation nous est présenté en « figure 1 ». Il faut noter que nous avons spécifié le nombre d'axes à produire (3) ; Ensuite nous avons activé l'option qui permet de calculer les COS2 et les contributions pour chaque individu.

Eigen values

Matrix trace	6.000000				
Average	1.000000				







Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	1.310633	0.269666	21.84 %		21.84 %
2	1.040966	0.022367	17.35 %		39.19 %
3	1.018599	0.076807	16.98 %		56.17 %
4	0.941792	0.055348	15.70 %		71.87 %
5	0.886443	0.084877	14.77 %		86.64 %
6	0.801566	-	13.36 %		100.00 %
Tot.	6.000000	-	-	-	-

FIGURE 8 – valeur propre

D'après le tableau, l'inertie (l'inertie indique la dispersion autour du barycentre, c'est une variance multidimensionnelle « calculée sur p dimensions ») expliquée par le premier axe principal est $\lambda_1 = 1.31$ ainsi la part d'inertie expliquée par la première composante principale est $\lambda_1/p = 0,2183$, avec p = nombre de facteurs. C'est-à-dire la première composante

principale explique à elle seule 21,83 % de la variance totale. D'après les informations de la « figure 9 » ci-dessous on constate que seules les trois (3) premières valeurs propres sont importantes

Significance of Principal Components

Global critical values	
Kaiser-Guttman	1
Karlis-Saporta-Spinaki	1.07831

Eigenvalue table - Test for significance

Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	1.320896	2.450000
2	1.045666	1.450000
3	1.034964	0.950000
4	0.951422	0.616667
5	0.865841	0.366667
6	0.781212	0.166667

FIGURE 9 – Importance des composantes principales

5.3 Corrélation entre les valeurs et les axes principaux

La « figure 10 » indique la corrélation des variables avec les axes factoriels. Nous pouvons constater que le premier axe est fortement corrélé positivement avec les variables : « education-num », « hours-per-week », ce qui signifie qu'ils sont des aspects importants dans la détermination des revenus annuels. En outre, il existe une corrélation de liaison étroite entre les variables « âge » et « capital-loss ».

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2		Axis_3	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
education-num	0.64535	42 % (42 %)	-0.28836	8 % (50 %)	-0.05548	0 % (50 %)
hours-per-week	0.60206	36 % (36 %)	-0.29198	9 % (45 %)	-0.18493	3 % (48 %)
fnlwgt	-0.32478	11 % (11 %)	-0.64300	41 % (52 %)	-0.14998	2 % (54 %)
age	0.38446	15 % (15 %)	0.58980	35 % (50 %)	0.22964	5 % (55 %)
capital-loss	0.34576	12 % (12 %)	0.15026	2 % (14 %)	-0.72282	52 % (66 %)
capital-gain	0.41122	17 % (17 %)	-0.30557	9 % (26 %)	0.63244	40 % (66 %)
Var. Expl.	1.32090	22 % (22 %)	1.04567	17 % (39 %)	1.03496	17 % (57 %)

FIGURE 10 – Corrélation entre les variables et les axes principaux

5.4 Plans factoriel

Plan factoriel âge vs PCA_AXIS_1

L'ACP repose en grande partie sur les représentations graphiques qu'elle propose. Ces dernières nous permettent d'apprécier visuellement les proximités entre les observations. En ce qui nous concerne, nous projetons les observations dans le premier plan factoriel. Nous voulons mettre en oeuvre les variables « age » et « education-num » afin de pouvoir suivre le comportement des revenus annuels des individus. La fonction d'âge décrit l'âge des individus.

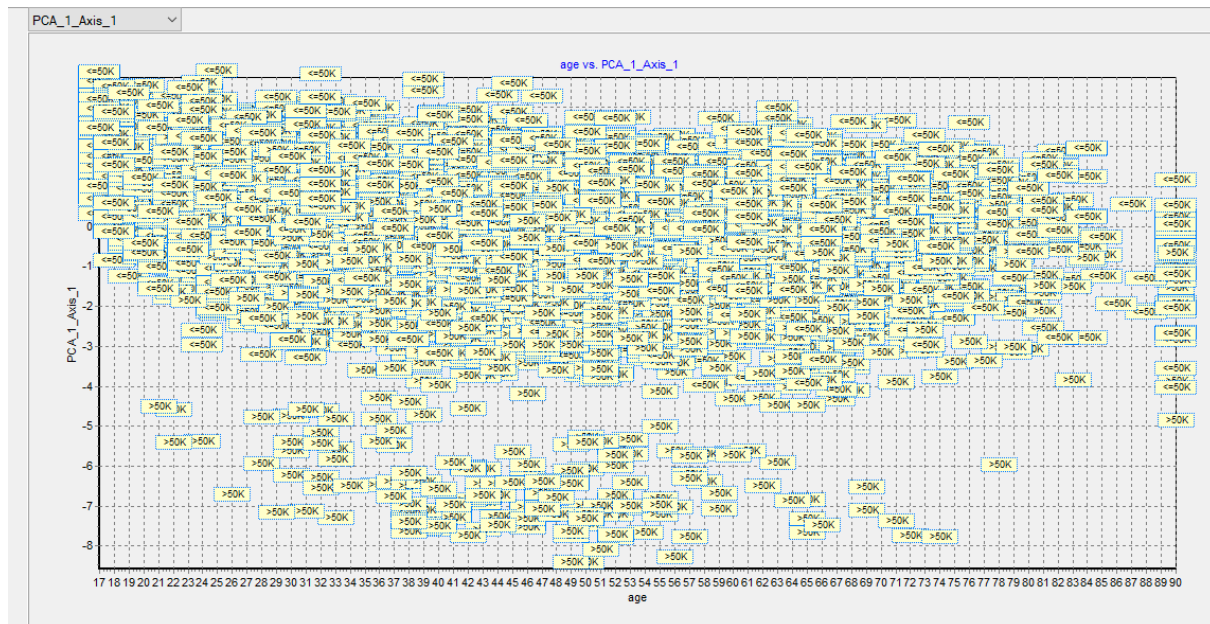


FIGURE 11 – âge vs PCA_AXIS_1

« La figure 11 » montre la repartitions d'âge avec les entrées de notre ensemble de donnée. Les

âges varient entre 17 à 90 ans, les tranches d'âges comprises entre 17 à 25 ans et celles comprises entre 75 à 90 ans ont un revenu inférieur ou égale à 50k ($\leq 50k$). Par contre, le revenu de la majorité des individus appartenant à la tranche d'âge de 26 à 74 ans est significatif ($> 50k$).

5.5 Cercle des corrélations

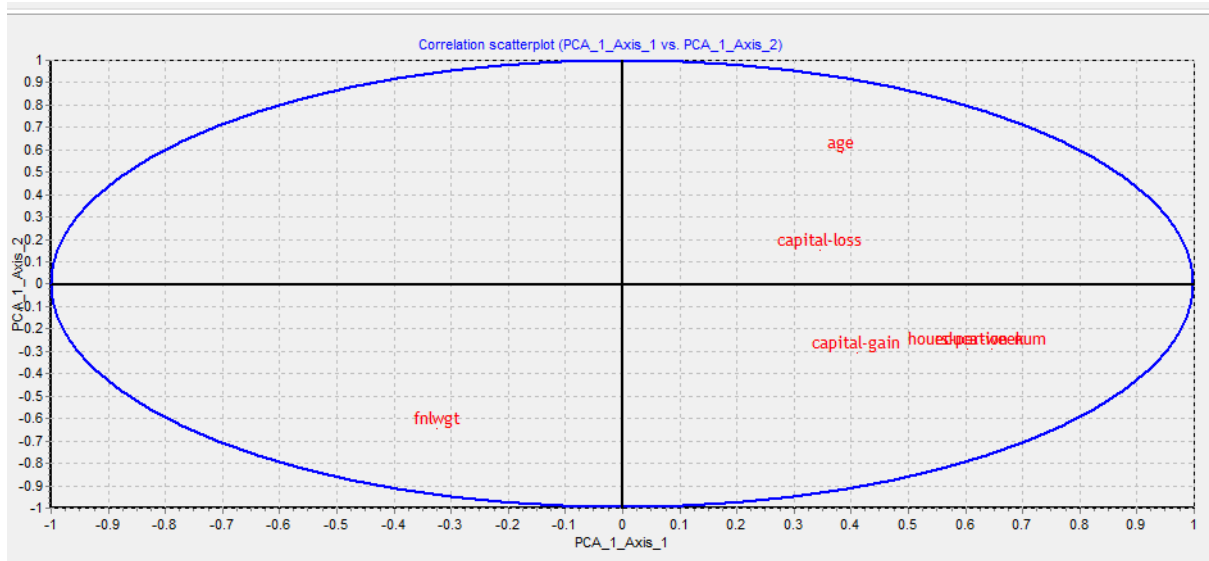


FIGURE 12 – cercle des corrélations

Nous constatons que les variables « education-num » et « hours-per-week » sont très proches du cercle l'une de l'autre et du premier axe principal (valeurs positives) ce qui indique leur forte corrélation positives avec le premier axe principal, en même temps ces deux variables sont proches de la variable « capital-gain » ; on remarque ainsi leur opposition avec la variable « fmlwgt ». La variable « age » situant sur le deuxième axe et celle de « capital-loss » sur le troisième axe sont faiblement proche, on peut aussi voir que la variable « age » est plus proche du cercle que celle de capital-loss, il n'y a donc pas d'informations les caractérisant. En définitive, on peut en déduire que l'analyse faite sur le cercle des corrélations confirme nos observations soulignées précédemment.

6 CONTEXTE ET ENONCE DU PROBLEME

6.1 Contexte

Le choix d'une méthode d'analyse pour effectuer un apprentissage supervisé est déterminé par plusieurs critères, notamment la nature des données et le type de résultats attendus.

L'apprentissage par arbre de décision est une méthode classique en apprentissage automatique. Son but est de créer un modèle qui prédit la valeur d'une variable-cible depuis la valeur de plusieurs variables d'entrée.

Une des variables d'entrée est sélectionnée à chaque nœud intérieur (ou interne, nœud qui n'est pas terminal) de l'arbre selon une méthode qui dépend de l'algorithme. Chaque arête

vers un nœud-fils correspond à un ensemble de valeurs d'une variable d'entrée, de manière à ce que l'ensemble des arêtes vers les nœuds-fils couvrent toutes les valeurs possibles de la variable d'entrée.

L'arbre est en général construit en séparant l'ensemble des données en sous-ensembles en fonction de la valeur d'une caractéristique d'entrée. Ce processus est répété sur chaque sous-ensemble obtenu de manière récursive, il s'agit donc d'un partitionnement récursif.

La récursion est achevée à un nœud soit lorsque tous les sous-ensembles ont la même valeur de la caractéristique-cible, ou lorsque la séparation n'améliore plus la prédiction.

En fouille de données, les arbres de décision peuvent aider à la description, la catégorisation ou la généralisation d'un jeu de données fixé.

L'ensemble d'apprentissage est généralement fourni sous la forme d'enregistrements du type :

$$(\mathbf{x}, \mathbf{Y}) = (x_1, x_2, x_3, \dots, x_k, \mathbf{Y})$$

La variable \mathbf{Y} désigne la variable-cible que l'on cherche à prédire, classer ou généraliser. Le vecteur \mathbf{x} est constitué des variables d'entrée :

$$x_1, x_2, x_3 \text{ etc.}$$

qui sont utilisées dans ce but.

Il existe **deux principaux types** d'arbre de décision en fouille de données :

- Les arbres de **classification** (Classification Tree) permettent de prédire à quelle classe la variable-cible appartient, dans ce cas la prédiction est une étiquette de classe,
- Les arbres de **régression** (Regression Tree) permettent de prédire une quantité réelle (par exemple, le prix d'une maison ou la durée de séjour d'un patient dans un hôpital), dans ce cas la prédiction est une valeur numérique.

Les arbres utilisés dans le cas de la régression et dans le cas de la classification présentent des similarités mais aussi des différences, en particulier en ce qui concerne la procédure utilisée pour déterminer les séparations des branches

6.2 Enoncé du problème

De nombreuses tâches d'exploration de données requièrent la classification des données en différentes classes. Par exemple, les demandes de prêt peuvent être classées dans des classes «approuver» ou «désapprouver». Cependant, [1] Certains classificateurs sont naturellement plus faciles à interpréter que d'autres ; par exemple, les arbres de décision (Quinlan 1993) sont faciles à visualiser, tandis que les réseaux neuronaux sont beaucoup plus difficiles. Notre choix s'est alors porté sur les algorithmes d'arbre de décision notamment sur la méthode **CART**. Cette méthode sera axée sur notre jeu de données «**Adult**» déjà décrit dans les TPs précédents. L'objectif est de prédire la variable «**CLASS**», un individu a-t-il un revenu annuel supérieur ou inférieur à 50.000 \$, à partir de ses caractéristiques signalétiques (niveau d'éducation, âge, etc.). Pour cela,

nous avons subdivisé la base en 2 parties :

- Une partie qui va constituer d'échantillon d'apprentissage de 1000 observations réservé pour la construction du modèle ;
- Et une autre partie correspondant à l'échantillon test, comprenant 2262 individus pour en évaluer les performances. Nous avons introduit une variable indicatrice INDEX permettant de spécifier l'appartenance d'un individu à l'un ou l'autre des sous échantillons.

7 PRESENTATION DE LA METHODE

[2]L'algorithme **CART** dont l'acronyme signifie « Classification And Regression Trees », s'attelle à construire un arbre de décision en classifiant un ensemble d'enregistrements. Cet arbre fournit un modèle pour classer de nouveaux échantillons. Il a été publié par Leo Breiman en 1984. L'algorithme construit un arbre de décision d'une manière analogue à l'algorithme ID3. Contrairement à ce dernier, l'arbre de décision généré par CART est binaire (un nœud ne peut avoir que deux fils) et le critère de segmentation est l'indice de diversité de Gini.

Nous dénombrons trois (3) étapes successives de l'algorithme CART : Construction de l'arbre et phase d'expansion, Elagage, Sélection finale.

- Construction de l'arbre et phase d'expansion :

- On suppose prédéfini un ensemble de tests binaires.
 - ① Variables qualitatives à n modalités.
 - autant de tests binaires que de partitions en deux classes.
 - $2^{n-1} - 1$ tests possibles.
 - ② Variables quantitatives
 - Une infinité de découpage selon des seuils.
 - Le meilleur seuil est choisi par un expert ou de manière automatique.
- On dispose d'un échantillon S découpé en un ensemble d'apprentissage A et un ensemble de test T .

- Entrée : ensemble d'apprentissage A
- On utilise la fonction Gini.
- Décider si un nœud est terminal :
Un nœud à la position p est terminal si $Gini(p) \leq s_0$ ou $n(p) \leq n_0$ où s_0 et n_0 sont des paramètres à fixer.
- Sélectionner un test à associer à un nœud :
On choisit le test qui maximise $\Delta(p, t)$, avec p une position, t un test et P_g, P_d la proportion d'éléments qui vont sur la position p_1 , respectivement p_2

$$\Delta(p, t) = Gini(p) - (P_g \times Gini(p_1) + P_d \times Gini(p_2))$$

- Affecter une classe à une feuille : on choisit la classe majoritaire
- Sortie : un arbre de décision.

- Elagage :

- **Entrée** : l'arbre de décision obtenu dans la phase d'expansion.
- **Construction** d'une suite d'arbres $t_0 t_1 \dots t_k$.
- On **calcule** pour chaque t_j l'erreur apparente sur l'ensemble T
- La suite est donnée par :
 - ① t_0 est l'arbre obtenu dans la phase d'expansion.
 - ② t_k est une feuille.
 - ③ A l'étape t_i : pour toute position p de t_i , on calcule $g(p)$ et on choisit la position p qui minimise $g(p)$. L'arbre t_{i+1} est un élagué de t_i en position p .
- **Sortie** : l'arbre de la suite dont l'erreur apparente est minimale.

Ainsi, la fonction $g(p)$ se définit comme suit :

Calcul de $g(p)$: soit u_p le sous-arbre de t_i à la position p et

$$g(p) = \frac{\Delta_{app}(p)}{|u_p| - 1}$$

, où $\Delta_{app}(p) = \frac{MC(p) - MC(u_p)}{N(p)}$, nombre d'erreurs supplémentaires que commet l'arbre sur l'échantillon lorsqu'on élague à la position p . $|u_p| - 1$ mesure le nombre de feuilles supprimées.

- $|u_p|$ taille de l'arbre u_p
- $N(p)$ est le nombre d'exemples de A associés à p .
- $MC(p)$ est le nombre d'exemples de A mal classés à p si on élague t_i en position p .
- $MC(u_p)$ est le nombre d'exemples de A associés à p de t_i mal classés par u_p

On choisit la position p pour laquelle $g(p)$ est minimale.

— Selection finale :

- t_{i+1} est obtenu à partir de t_i , auquel on coupe la branche qui permet un g minimal.
- Soit t_0, \dots, t_k la suite obtenue, t_k est réduit à une feuille.
- Sélection de l'arbre t_i dont le nombre d'erreurs calculées sur l'ensemble de validation est minimal.

7.1 Représentation d'un arbre CART

Voici l'aspect d'un arbre CART

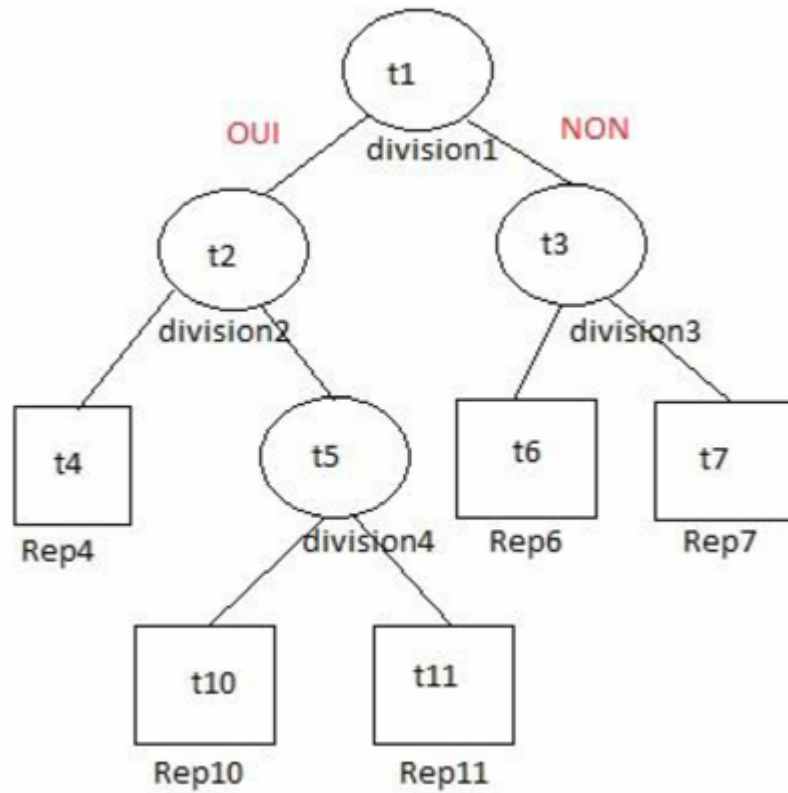


FIGURE 13 – Exemple d'un arbre CART

8 APPLICATION DE LA METHODE A NOS DONNEES

8.1 Apprentissage avec la methode C-RT

Dans Tanagra, le composant C-RT correspond à la méthode CART telle qu'elle est décrite dans l'ouvrage de référence (Breiman et al., 1984). Dans la phase d'expansion, l'indice de Gini est utilisé pour choisir les variables de segmentation.

Supervised Learning 1 (CS-CRT)

Parameters

Classification tree (C-RT) parameters	
Size before split	10
Pruning set size (%)	33
x-SE rule	1,00
Random generator	1
Show all tree seq (even if > 15)	0

Misclassification Cost Matrix

	Observed vs. Predicted	
Cost(i,j)	1	2
1	0	1
2	1	0

Results

Classifier performances

Error rate			0,1742			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,9297	0,1439	<=50K	714	54	768
>50K	0,4805	0,3273	>50K	120	111	231
			Sum	834	165	999

La matrice de confusion nous montre les vraies valeurs et les valeurs prédites de SALARY sur les 1000 observations ayant participé à l'apprentissage (growing + pruning).

Partition des données

Data partition

Growing set	669
Pruning set	330

TANAGRA nous indique que parmi les 1000 observations dédiées à l'apprentissage, il a réservé 669 observations pour l'expansion de l'arbre (growing set) et 330 pour le post élagage (pruning set). La partition a été effectuée de manière aléatoire

Trees sequence (# 8)

N°	# Leaves	Cost (growing set)	Cost (pruning set)	SE (pruning set)	x
8	1	0,2317	0,2303	0,0232	2,603341
7	3	0,1719	0,1788	0,0211	0,144630
6	5	0,1584	0,1848	0,0214	0,433890
5	6	0,1540	0,1758	0,0210	0,000000
4	9	0,1420	0,1848	0,0214	-
3	13	0,1360	0,2121	0,0225	-
2	21	0,1271	0,1970	0,0219	-
1	25	0,1241	0,2000	0,0220	-

T . . .

On constate à ce niveau que l'arbre le plus grand lors de la phase d'expansion comporte 25 feuilles, le taux d'erreur sur le growing set est de 0.1241, sur le pruning set 0.2000.

Tree description

Number of nodes	5
Number of leaves	3

Decision tree

- marital-status in [Never-married, Divorced, Married-spouse-absent, Separated, Married-AF-spouse, Widowed] then class = <=50K (92,0 % of 99 examples)
- marital-status in [Married-civ-spouse]
 - education in [Bachelors, Masters, Assoc-acdm, Assoc-voc, Doctorate, Prof-school, Preschool, 12th] then class = >50K (69,23 % of 99 examples)
 - education in [HS-grad, 11th, 9th, Some-college, 7th-8th, 5th-6th, 10th, 1st-4th] then class = <=50K (71,66 % of 99 examples)

Les attributs les plus déterminants sont MATRIAL-STATUS et EDUCATION.

8.2 Evaluation sur l'échantillon test

Evaluation sur l'échantillon test Les ensembles growing et pruning participent, chacun à leur manière, à l'élaboration du modèle de prédiction. A ce titre, ils fournissent une estimation optimiste des performances puisque l'arbre est optimisé pour ces données. Pour obtenir une évaluation réellement non biaisée, il faut utiliser un ensemble test qui n'a jamais participé, de près ou de loin, à l'apprentissage. C'est à ce stade que nous mettons à contribution les individus « Index = test »

Test 1

Parameters

Evaluation set : unselected examples

Results

pred_SpvInstance_1

Error rate			0,1958			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,9138	0,1624	<=50K	1537	145	1682
>50K	0,4862	0,3396	>50K	298	282	580
			Sum	1835	427	2262

Nous activons le menu VIEW. Nous obtenons un taux d'erreur de 0.1958 calculé sur les 2262 individus que nous avons mis de côté initialement. Ce n'est qu'une estimation bien sûr. Mais étant calculé sur un effectif aussi élevé, nous pouvons penser qu'il est relativement fiable.

8.3 Courbe d'erreur en fonction de la complexité de l'arbre

Pour obtenir le détail de la courbe d'erreur, nous activons le menu contextuel SUPERVISED PARAMETERS du composant SUPERVISED LEARNING 1 (C-RT) dans le diagramme.

Test 1

Parameters

Evaluation set : unselected examples

Results

pred_SpvInstance_1

Error rate	0,1958		
Values prediction			
Value	Recall	1-Precision	
<=50K	0,9138	0,1624	<=50K
>50K	0,4862	0,3396	>50K
			Sum

Computation time : 0 ms.
Created at 30/06/2018 21:58:42

Cost Sensitive C-RT

Parameters

Cost matrix

Min size of node to split :

10

Pruning set size (%) :

33

x-SE Rule :

1

Random number generator

☐ Random
 ☒ Standard

☒ Show all tree sequence (even if > 15)

OK

Cancel

Help

Nous actionnons le menu VIEW, le tableau retraçant les erreurs est détaillé maintenant. Nous remarquons plusieurs choses : le nombre de feuilles des arbres qui ont été testés n'est pas régulier, le mécanisme de coût complexité permet de réduire considérablement les solutions à évaluer ; l'erreur sur l'échantillon growing diminue constamment à mesure que le nombre de feuilles augmente ; l'erreur sur le pruning set diminue rapidement d'abord, semble stagner sur un palier, puis se dégrade lorsque le nombre de feuilles devient exagéré.

Trees sequence (# 12)

N°	# Leaves	Cost (growing set)	Cost (pruning set)	SE (pruning set)	x
12	1	0,2317	0,2303	0,0232	3,273620
11	3	0,1719	0,1788	0,0211	0,744004
10	4	0,1614	0,1636	0,0204	0,000000
9	6	0,1480	0,1697	0,0207	-
8	8	0,1360	0,1727	0,0208	-
7	9	0,1315	0,1636	0,0204	-
6	11	0,1241	0,1879	0,0215	-
5	13	0,1181	0,1939	0,0218	-
4	16	0,1106	0,1939	0,0218	-
3	20	0,1046	0,2061	0,0223	-
2	28	0,0957	0,1909	0,0216	-
1	32	0,0927	0,1939	0,0218	-

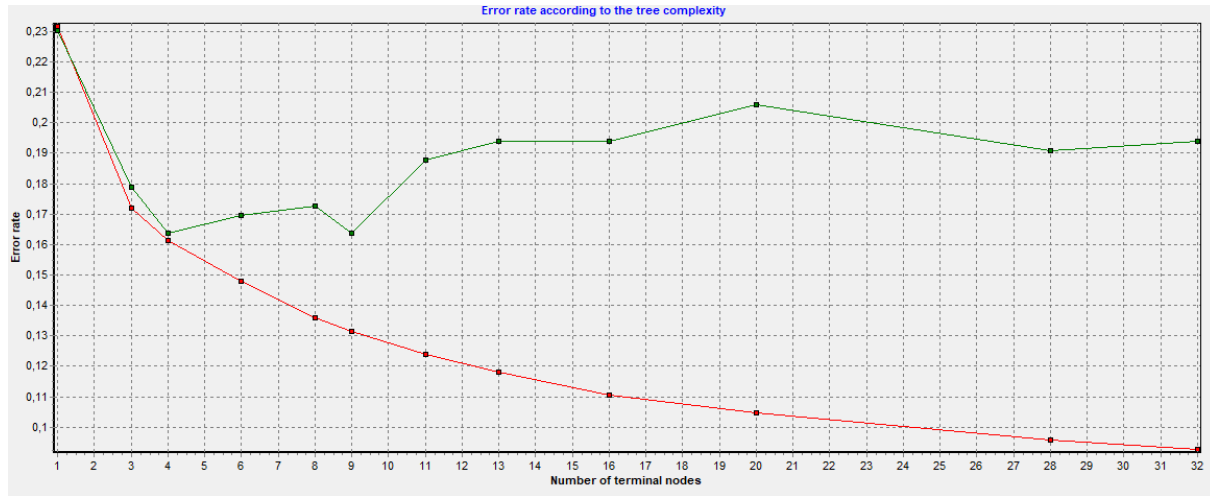


FIGURE 14 – Evolution du taux d'erreur en fonction de la complexité de l'arbre

8.4 Performance de l'arbre 0-SE RULE

Nous voulons évaluer les performances de l'arbre optimal. Nous paramétrons de nouveau le composant SUPERVISED LEARNING 1 (C-RT) en imposant la valeur 0-SE RULE.

Trees sequence (# 8)

N°	# Leaves	Cost (growing set)	Cost (pruning set)	SE (pruning set)	x
8	1	0,2317	0,2303	0,0232	2,603341
7	3	0,1719	0,1788	0,0211	0,144630
6	5	0,1584	0,1848	0,0214	0,433890
5	6	0,1540	0,1758	0,0210	0,000000
4	9	0,1420	0,1848	0,0214	-
3	13	0,1360	0,2121	0,0225	-
2	21	0,1271	0,1970	0,0219	-
1	25	0,1241	0,2000	0,0220	-

Classifier performances

Error rate			0,1612			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,9570	0,1483	<=50K	735	33	768
>50K	0,4459	0,2426	>50K	128	103	231
			Sum	863	136	999

Voyons maintenant ce qu'il en est sur l'échantillon test, nous activons pour cela le menu VIEW du composant TEST 1 au bout du diagramme de traitement.

Error rate			0,1989			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,9358	0,1785	<=50K	1574	108	1682
>50K	0,4103	0,3121	>50K	342	238	580
			Sum	1916	346	2262

Le taux d'erreur en test est 0.1989 c'est qui n'est pas significativement différent de l'erreur de l'arbre à 6 feuilles produit par la 1-SE RULE.

9 PRESENTATION D'AUTRES METHODES POUR LA COMPARAISON

9.1 Pretraitement des données

Nous allons utilisé deux méthodes (LDA et SVM) pour la comparaison des résultats obtenus précédemment. Comme annoncé ci-haut, notre jeu de données est composé de deux types de variables : continues et discrètes. De ce fait, pour pouvoir utilisé LDA et SVM, nous devons d'abord convertir les attributs discrets en attributs continus.

Nous utilisons le composant **0_1_Binarize** pour binariser les attributs discret en attributs continus sauf la variable à prédire CLASS

Parameters	
Used values : K-1 (ignore last value)	
Results	
Attribute binarization	
Source att	New attributes
workclass	(workclass_State-gov_1,workclass_Self-emp-not-inc_1,workclass_Private_1,workclass_Federal-gov_1,workclass_Local-gov_1,workclass_?_1,workclass_Self-emp-inc_1)
education	(education_Bachelors_1,education_H5-grad_1,education_11th_1,education_Masters_1,education_9th_1,education_Some-college_1,education_Assoc-acdm_1,education_Assoc-voc_1,education_7th-8th_1,education_Doctorate_1,education_Prof-school_1,education_5th-6th_1,education_10th_1,education_1st-4th_1,education_Preschool_1)
marital-status	(marital-status_Never-married_1,marital-status_Married-civ-spouse_1,marital-status_Divorced_1,marital-status_Married-spouse-absent_1,marital-status_Separated_1,marital-status_Married-AF-spouse_1)
occupation	(occupation_Adm-clerical_1,occupation_Exec-managerial_1,occupation_Handlers-cleaners_1,occupation_Prof-specialty_1,occupation_Other-service_1,occupation_Sales_1,occupation_Craft-repair_1,occupation_Transport-moving_1,occupation_Farming-fishing_1,occupation_Machine-op-inspct_1,occupation_Tech-support_1,occupation_?_1,occupation_Protective-serv_1,occupation_Armed-Forces_1)
relationship	(relationship_Not-in-family_1,relationship_Husband_1,relationship_Wife_1,relationship_Own-child_1,relationship_Unmarried_1)
race	(race_White_1,race_Black_1,race_Asian-Pac-Islander_1,race_Amer-Indian-Eskimo_1)
sex	(sex_Male_1)
native-countr	(native-countr_United-States_1,native-countr_Cuba_1,native-countr_Jamaica_1,native-countr_India_1,native-countr_?_1,native-countr_Mexico_1,native-countr_South_1,native-countr_Puerto-Rico_1,native-countr_Honduras_1,native-countr_England_1,native-countr_Canada_1,native-countr_Germany_1,native-countr_Iran_1,native-countr_Philippines_1,native-countr_Italy_1,native-countr_Poland_1,native-countr_Columbia_1,native-countr_Cambodia_1,native-countr_Thailand_1,native-countr_Ecuador_1,native-countr_Laos_1,native-countr_Taiwan_1,native-countr_Haiti_1,native-countr_Portugal_1,native-countr_Dominican-Republic_1,native-countr_El-Salvador_1,native-countr_France_1,native-countr_Guatemala_1,native-countr_China_1,native-countr_Japan_1,native-countr_Yugoslavia_1,native-countr_Peru_1,native-countr_Outlying-US(Guam-USVI-etc)_1,native-countr_Scotland_1,native-countr_Trinidad&Tobago_1,native-countr_Greece_1,native-countr_Nicaragua_1,native-countr_Vietnam_1,native-countr_Hong_1)

FIGURE 15 – *Binarisation des attributs*

9.2 Linear Disriminat Analysis (LDA)

Après avoir appliquer LDA sur les données, nous avons obtenu les résultats suivants :

Parameters	
Matrix inversion	approximated

Classifier performances

Error rate			0,1451			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,9360	0,1214	<=50K	702	48	750
>50K	0,6104	0,2400	>50K	97	152	249
			Sum	799	200	999

FIGURE 16 – Apprentissage avec LDA

pred_SpvInstance_2						
Error rate			0,1808			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,8965	0,1326	<=50K	1524	176	1700
>50K	0,5854	0,3485	>50K	233	329	562
			Sum	1757	505	2262

FIGURE 17 – test avec LDA

9.3 Support Vector Machine (SVM)

Nous avons les résultats de l'apprentissage et du test

Classifier performances

Error rate			0,1522			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,9480	0,1371	<=50K	711	39	750
>50K	0,5462	0,2229	>50K	113	136	249
			Sum	824	175	999

FIGURE 18 – Apprentissage avec SVM

Evaluation set : unselected examples

pred_SpvInstance_3						
Error rate			0,1927			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,9118	0,1558	<=50K	1550	150	1700
>50K	0,4911	0,3521	>50K	286	276	562
			Sum	1836	426	2262

FIGURE 19 – Test avec SVM

9.4 Comparaison des trois méthodes étudiées

La comparaison portera sur les résultats des taux d'erreurs observés sur les différents modes de classifications étudiés ci-haut :

Méthode de classification	taux d'erreur	taux d'erreur test
CART	0.1612	0.1989
LDA	0.1451	0.1808
SVM	0.1522	0.1927

D'après les résultats obtenus au niveau des deux(2) méthodes de classifications (**LDA et SVM**), nous pouvons constater que les taux d'erreurs d'apprentissages chez LDA et SVM sont significativement inférieure aux taux d'erreur de la méthode CART du tree. Cela s'explique du faite que le CART met énormément de temps dans l'apprentissage que les LDA et SVM.

Les mêmes observations peuvent être constater au niveau de l'élément test des trois méthodes, aux quels cas le taux d'erreur de CART reste toujours supérieure aux deux autres.

10 Conclusion

Ce travail pratique effectué dans le cadre du module **Fouille de données** avait pour but de nous initier sur les différents algorithmes d'apprentissages de data mining. Notre travail est subdivisé en deux parties : une première partie qui traite sur la préparation des données et une seconde partie qui porte la prédiction et l'apprentissage. Nous avons choisi un jeu de données dénommé «Adult» qui constitue l'objet d'études pour ce projet. Le but de ce jeu de données est de prédire la monotonie du revenu annuel d'un individu n'excédant pas 50.000\$ à partir de ses attributs. Nous avons ensuite fait une description détaillée de ce jeu de données (description des attributs, nombre d'individus, valeurs manquantes, problème posé) ; puis nous avons fait l'analyse exploratoire de chaque attribut et l'analyse de lien entre chaque paire d'attributs.

Dans la seconde partie nous avons, notre choix s'est d'abord porté sur la méthode **CART** car nous pensons qu'elle est mieux adaptée à nos données. D'après les différents test, nous remarquons que les résultats sont satisfaisants.

Ensuite, nous avons étudié deux (2) autres méthodes avec les mêmes jeux de données pour observer les différences vis à vis de la méthode choisie. Le résultat à été concluant.

Enfin, parmi les innombrables variantes des techniques d'apprentissage des arbres de décision, CART est probablement celle qui détecte le mieux la bonne profondeur de l'arbre. Elle produit de ce fait, bien souvent, des modèles performants.

Références

- [1] R Kohavi. Scaling up the accuracy of naive-bayes classifiers : a decision-tree hybrid accuracy scale-up : the learning. *Data Min. Vis., no. Utgo 1988*, pages 1–6, 1996.
- [2] Wikipédia. Algorithme cart — wikipédia, l'encyclopédie libre, 2018. [En ligne ; Page disponible le 21-mars-2018].