COURS DE FOUILLE DE DONNÉES

M1, P-20- IFI 2016

Enseignante: **NGUYỄN Thị Minh Huyền**

Rapport Final – G2

Données Utilisées: **Census Income Data Set**

Étudiant: **Ginel Dorleon**
Gervais Fotsing Sikadie S.

Table Des Matières

COURS DE FOUILLE DE DONNÉES M1, P-20- IFI 2016

Rapport Final – G2
Données Utilisées: Census Income Data Set

I.INTRODUCTION GÉNÉRALE

De nos jours toutes les entreprises collectent et stockent de grandes quantités de données. Ces mégabases de données, qui ne cessent d'augmenter jour après jour, sont peu exploitées, alors qu'elles cachent de connaissances décisives face au marché et à la concurrence. Pour combler ce besoin, une nouvelle industrie est née : la Fouille de Données ou la Science des Données.

La Fouille de Données a ainsi pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances. [1]

Pour appliquer ainsi les connaissances acquises durant l'introduction du cours de fouille de données, nous allons ainsi mener une étude sur un jeu de données. Le présent rapport est divisé en deux grandes parties:

L'Analyse Statistique des Variables dans laquelle on va faire un briefing pour prendre connaissance de notre jeu de données et enfin l'Apprentissage Supervisée qui constitue notre principal objectif

Pour faire ce travail, nous allons utilisé le logiciel spécialisé, tanagra

II.CONTEXTE

Jeu de Données: **Census Income Data Set** [2]

L'étude porte sur les rémunérations des salaries aux USA en 2001. Les données que nous utilisons ont été extraites de la base de données d'un bureau de recensement aux USA. 32561 instances avec 15 attributs . L'objectif est d'étudier un ensemble de paramètre et d'établir une tache de prévision afin de déterminer le profil des personnes qui gagnent plus ou moins de 50K par année.

III.DESCRIPTION DES VARIABLES

Age: Variable continue représentant l'age des personnes.
Workclass: Variable discrète prenant les valeurs suivantes : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

Fnlwgt: Final weight, une variable continue calculée à partir des données d'origine et du sexe de la personne.

Education: Variable discrète déterminant le niveau d'éducation de la personne, elle prend les valeurs suivantes : Licence, Certains collège, 11e, HS-grad, Prof-école, Assoc-ACDM, Assoc-voc, 9e, septième-huitième, 12e, Masters, 1ere-4ème, 10e, Doctorat, 5ème-6ème, Maternelle.
éducation-num: continu.
statut matrimonial: Marié-civ-conjoint, Divorcé, jamais marié, Séparé, Veuf, Marié-conjoint absent, Marié-AF-conjoint.

Profession: Tech-support, Craft-réparation, Autre service, ventes, Exec-gestion, Prof spécialité, Handlers-nettoyants, machine-op-inspct, Adm-clérical, agriculture-pêche, Transport-mobile, Priv-ménage serv, protection-serv, armées-Forces.

Relationship: Variable discrète représentant les types de relations des individus. Elle prend les valeurs suivantes: Femme, propre enfant, Marié , Non-en-famille, Autre-parent, Unmarried.

Race: Variable discrète représentant la race des individus. Elle prend les valeurs Noir Blanc, Asie-Pac-Islander, Amer-Indian-Eskimo, Autre.

Sexe: Variable discrète représentant le sexe de l'individu, 2 valeurs : Femme, Homme.
Capital-gain: Variable continue représentant le capital gagné par l'individu.
Capital-loss:Variable continue représentant le capital perdu par l'individu.
Hours-per-week : Variable continue représentant le nombre d'heur que l'individu travail
Pays: Variable discrète représentant le pays d'origine de l'individu.

Salary: Notre variable cible <=50K, >50 K

STATISTIQUE SUR LES VARIABLES

IV. STATISTIQUES SUR LES VARIABLES

Dans la plupart des études sur les données , le nombre de sujets est souvent trop important pour que l'on puisse présenter les données réelles de chaque individu. C'est pourquoi, il est nécessaire de trouver un moyen qui donne le maximum d'informations possible sous le format le plus utile. Une manière courante de présentation de données est la représentation graphique ou le tableau. Les tableaux sont commodes pour présenter l'information relative aux données individuelles et les graphiques pour donner un profil général des observations. Toutefois, il est également utile de donner un résumé chiffré. Pour les variables qualitative ou en catégorie (niveau d'étude, sexe, absence-présence d'une maladie, niveau d'éducation, etc.), la mesure la plus instructive est la proportion d'individus entrant dans chaque catégorie. Les variables quantitatives (poids, taille, âge, salaire, etc.) nécessitent quant à elle deux types de mesure pour avoir une idée complète de la distribution des observations : la mesure de la *position centrale* des observations et la mesure de leur *dispersion*, c'est-à-dire la mesure de la répartition des observations autour de cette position centrale. Ainsi, nous allons faire procéder à l'étude qualitative et quantitative de deux des variables de notre jeu de données.

1- **Étude d'une variable qualitative- Education**

Généralement, en statistique, une variable qualitative, on dit aussi catégorielle, est une variable pour laquelle la valeur mesurée sur chaque individu ne représente pas une quantité. Les différentes valeurs que peut prendre cette variable sont appelées les *catégories*, *modalités* ou *niveaux*.

Ainsi, parmi l'ensemble des variables qualitatives de notre jeu de données, nous avons choisi d'étudier la variable *Education*. En effectuant la statistique primaire de cette variable, on peut constater que la plus grande valeur valeur Hs-grad signifiant le degré en High School. On constate que le mode est cette valeur, Hs-grad. Le nombre d'individu ayant ce niveau est de 10501 et le pourcentage d'individus ayant ce niveau est de 32.25 % . Voir capture ci-dessous

Univariate discrete stat 1

Parameters

Attributes : 1
Examples : 32561

Results

Attribute	Gini	Distribution			
		Values	Count	Percent	Histogram
EDUCATION	0,8096	Bachelors	5355	16,45 %	
		HS-grad	10501	32,25 %	
		11th	1175	3,61 %	
		Masters	1723	5,29 %	
		9th	514	1,58 %	
		Some-college	7291	22,39 %	
		Assoc-acdm	1067	3,28 %	
		Assoc-voc	1382	4,24 %	
		7th-8th	646	1,98 %	
		Doctorate	413	1,27 %	
		Prof-school	576	1,77 %	
		5th-6th	333	1,02 %	
		10th	933	2,87 %	
		1st-4th	168	0,52 %	

2 - Étude d'une variable quantitative – L'Age

Une variable est *quantitative* si elle reflète une notion de grandeur, c'est-à-dire si les valeurs qu'elle peut prendre sont des nombres. Une grandeur quantitative est souvent exprimée avec une unité de mesure qui sert de référence

Parmi l'ensemble des variables quantitatives de notre jeu de données, nous avons choisi d'étudier la variable quantitative Age. En effectuant l'analyse statistique primaire de cette variable, on constate que l'age minimum est de 17 ans, le max est de 90 ans et l'age moyenne est 38,5816. Voir capture ci-dessous.

Univariate continuous stat 1

Parameters

Attributes : 1
Examples : 32561

Results

Attribute	Min	Max	Average	Std-dev	Std-dev/avg
AGE	17	90	38,5816	13,6404	0,3535

Computation time : 0 ms.
Created at 21/12/2016 15:14:43

More Univariate cont stat 1

Parameters

Attributes : 1
Examples : 32561

Results

Attribute	Stats		Histogram			
	Statistics		Values	Count	Percent	Histogram
AGE	Average	38,5816	x_<_24,3000	5570	17,11%	
	Median	37,0000	24,3000_=<_x_<_31,6000	5890	18,09%	
	Std dev. [Coef of variation]	13,6404 [0,3535]	31,6000_=<_x_<_38,9000	6048	18,57%	
	MAD [MAD/STDDEV]	11,1892 [0,8203]	38,9000_=<_x_<_46,2000	6163	18,93%	
	Min * Max [Full range]	17,00 * 90,00 [73,00]	46,2000_=<_x_<_53,5000	3967	12,18%	
	1st * 3rd quartile [Range]	28,00 * 48,00 [20,00]	53,5000_=<_x_<_60,8000	2591	7,96%	
	Skewness (std-dev)	0,5587 (0,0136)	60,8000_=<_x_<_68,1000	1595	4,90%	
	Kurtosis (std-dev)	-0,1661 (0,0271)	68,1000_=<_x_<_75,4000	496	1,52%	
			75,4000_=<_x_<_82,7000	174	0,53%	
			x>=_82,7000	67	0,21%	

On peut ensuite observer les informations suivantes :

- La médiane est de 37
- La classe modale est [38,9000 ; 46,2000[, le mode est 42.5500
- L'écart-type est de 13,6404

On peut ensuite constater la dispersion de la variable Age.
L'étendue (différence entre la valeur max et la valeur min) est de 73,00
la dispersion interquartile (différence entre le troisième et le premier quartile) est de 20,00;

La représentation graphique ici est sous forme d'histogramme.

La forme de la distribution de l'Age est, suite aux statistiques primaires, étalée à droite car la moyenne est > médiane > mode (on peut aussi le constater en observant que le coefficient d'asymétrie de Fisher (Skewness = 0,5587) est positif.

La forme de la distribution de l'Age est moins aplatie qu'une distribution normale car coefficient d'aplatissement observé ici (Kurtosis = -01661) est négatif.

3 - Corrélation entre chaque paire de variable quantitative

En probabilité et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques *numériques*, c'est étudier l'intensité de la liaison qui peut exister entre ces deux variables X et Y. La mesure de la corrélation linéaire entre les deux se fait alors par le calcul du coefficient de corrélation linéaire, noté r. Cette corrélation est dite:

- Positive, c'est-à-dire à toute augmentation au niveau de X correspond une augmentation au niveau de Y. Les deux variables varient dans le même sens et avec une intensité similaire. Exemple: EDUCATION-NUM et HOURS-PER-WEEK.

- Négative, c'est-à-dire à toute augmentation au niveau de X correspond une diminution au niveau de Y. Les deux variables varient dans deux sens opposés et avec une intensité similaire. Exemple: EDUCATION-NUM et FNLWGT.

Le coefficient:

Le coefficient de corrélation est un indice statistique qui exprime l'intensité et le sens (positif ou négatif) de la relation linéaire entre deux variables quantitatives. C'est une mesure de la liaison linéaire, c'est à dire de la capacité de prédire une variable X par une autre Y à l'aide d'un modèle linéaire. Il permet de mesurer l'intensité de la liaison entre deux caractères quantitatifs. C'est donc un paramètre important dans l'analyse des régressions linéaires (simples ou multiples). En revanche, ce coefficient est nul (r = 0) lorsqu'il n'y a pas de relation linéaire entre les variables (ce qui n'exclut pas l'existence d'une relation autre que linéaire). Par ailleurs, le coefficient est de signe positif si la relation est positive (directe, croissante) et de signe négatif si la relation est négative (inverse, décroissante). Ce coefficient varie entre -1 et +1 ; l'intensité de la relation linéaire sera donc d'autant plus forte que la valeur du coefficient est proche de +1 ou de -1, et d'autant plus faible qu'elle est proche de 0. •

Une valeur proche de +1 montre une forte liaison entre les deux caractères, la relation linéaire est ici croissante (c'est-à-dire que les variables varient dans le même sens).•

Une valeur proche de -1 montre également une forte liaison mais la relation linéaire entre les deux caractères est décroissante (les variables varient dans le sens contraire).

• Une valeur proche de 0 montre une absence de relation linéaire entre les deux caractères.

Cependant, il faut noter que l'existence d'une corrélation élevée entre deux variables X et Y ne conduit pas à l'existence d'une relation de cause à effet. On utilise la connaissance de la variable X pour prédire les valeurs de Y ; cela n'implique pas qu'un changement de X cause un changement de Y.

Dans notre cas, le coefficient de corrélation de nos variables se situe dans la majorité des cas entre 0 et 0.2 donc c'est très faible. On en conclut qu'il n'existe pratiquement aucune comparaison entre les paires de variables présentés.

Dans le tableau suivant, nous présentons la corrélation entre chaque paire de nos variables quantitatives.

Dans le tableau ci-dessus, r représente le coefficient de corrélation pour chaque paire de variable, r^2 est l'écart-type

Linear correlation 2

Parameters

Cross-tab parameters	
Sort results	non
Input list	Target (Y) and input (X)

Results

Y	X	r	r²	t	Pr(>|t|)
AGE	FNLWGT	-0,0766	0,0059	-13,8709	0,0000
AGE	EDUCATION-NUM	0,0365	0,0013	6,5954	0,0000
AGE	CAPITAL-GAIN	0,0777	0,0060	14,0581	0,0000
AGE	CAPITAL-LOSS	0,0578	0,0033	10,4423	0,0000
AGE	HOURS-PER-WEEK	0,0688	0,0047	12,4358	0,0000
FNLWGT	AGE	-0,0766	0,0059	-13,8709	0,0000
FNLWGT	EDUCATION-NUM	-0,0432	0,0019	-7,8014	0,0000
FNLWGT	CAPITAL-GAIN	0,0004	0,0000	0,0779	0,9379
FNLWGT	CAPITAL-LOSS	-0,0103	0,0001	-1,8499	0,0643
FNLWGT	HOURS-PER-WEEK	-0,0188	0,0004	-3,3872	0,0007
EDUCATION-NUM	AGE	0,0365	0,0013	6,5954	0,0000
EDUCATION-NUM	FNLWGT	-0,0432	0,0019	-7,8014	0,0000
EDUCATION-NUM	CAPITAL-GAIN	0,1226	0,0150	22,2958	0,0000
EDUCATION-NUM	CAPITAL-LOSS	0,0799	0,0064	14,4677	0,0000
EDUCATION-NUM	HOURS-PER-WEEK	0,1481	0,0219	27,0256	0,0000

4 - **Contingence des variables qualitatives**

Pour évaluer le lien existant entre deux de nos variables qualitatives, on va présenter un tableau de contingence. Le tableau de contingence est une méthode de représentation de données issues d'un comptage permettant d'estimer la dépendance entre deux caractères. Elle consiste à croiser deux caractères d'une population, par exemple dans notre cas, *workclass et education* , en dénombrant l'effectif correspondant à la conjonction «caractère 1» et «caractère 2». Les effectifs partiels sont rassemblés dans un tableau à double entrée, par ligne pour le premier caractère, et par colonne en fonction du second caractère: c'est le «tableau de contingence».

Cet outil simple répond à un problème crucial en statistique: la détection d'éventuelles dépendances entre les qualités relevées sur les individus d'une population.

Ainsi, nous présentons le tableau suivant avec la variable *Workclass* d'une part et le niveau d'éducation (variable *education*) de l'individu.

Contingency Chi-Square 2

Parameters

Cross-tab parameters	
Sort results	non
Input list	Target (Row) and input (Column)
Additional information	0
Contribution thresold	2,0

Results

Row (Y)	Column (X)	Statistical indicator	
WORKCLASS	EDUCATION	Stat	Value
		d.f.	120
		Tschuprow's t	0,084918
		Cramer's v	0,099369
		Phi²	0,078993
		Chi² (p-value)	2572,10 (0,0000)
		Lambda	0,000000
		Tau (p-value)	0,0177 (0,0000)
		U(R/C) (p-value)	0,0286 (0,0000)

	Bachelors	HS-grad	11th	Masters	9th	Some-college	Assoc-acdm	Ass
State-gov	270	268	14	169	6	325	41	
Self-emp-not-inc	399	866	60	124	34	486	71	
Private	3551	7780	923	894	387	5094	729	
Federal-gov	212	263	9	67	3	254	55	
Local-gov	477	503	36	342	23	387	88	
?	173	532	118	48	51	514	47	
Self-emp-inc	273	279	14	79	10	226	35	
Without-pay	0	9	0	0	0	3	1	
Never-worked	0	1	1	0	0	2	0	
Sum	5355	10501	1175	1723	514	7291	1067	

On peut constater que, le secteur privé est le secteur qui offre le plus grand nombre d'emploi.

Les individus avec le niveau de Hs-grad représentent le plus grande valeur des employés. Il n'y a pas de Bacheliers, ni de niveau Masters, ni de niveau de 9th sans emplois ou qui ne sont jamais été employés .

5 -Test du Khi 2

Le *Test du Khi 2* est un test statistique permettant de tester l'adéquation d'une série de données à une famille de lois de probabilités ou de tester l'indépendance entre deux variables aléatoires.

Results

Row (Y)	Column (X)	Statistical indicator

Stat	Value
d.f.	120
Tschuprow's t	0,084918
Cramer's v	0,099369
Phi²	0,078993
Chi² (p-value)	2572,10 (0,0000)
Lambda	0,000000
Tau (p-value)	0,0177 (0,0000)
U(R/C) (p-value)	0,0286 (0,0000)

WORKCLASS / EDUCATION

	Bachelors	HS-grad	11th	Masters	9th	Some-college	Assoc-acdm
State-gov	270 (+ 1 %)	268 (- 2 %)	14 (- 1 %)	169 (+ 6 %)	6 (- 0 %)	325 (+ 0 %)	41 (- 0 %)
Self-emp-not-inc	399 (- 0 %)	866 (+ 0 %)	60 (- 0 %)	124 (- 0 %)	34 (- 0 %)	486 (- 0 %)	71 (- 0 %)
Private	3551 (- 0 %)	7780 (+ 1 %)	923 (+ 1 %)	894 (- 3 %)	387 (+ 0 %)	5094 (+ 0 %)	729 (- 0 %)
Federal-gov	212 (+ 1 %)	263 (- 0 %)	9 (- 1 %)	67 (+ 0 %)	3 (- 0 %)	254 (+ 0 %)	55 (+ 1 %)
Local-gov	477 (+ 2 %)	503 (- 2 %)	36 (- 1 %)	342 (+ 19 %)	23 (- 0 %)	387 (- 1 %)	88 (+ 0 %)
?	173 (- 2 %)	532 (- 0 %)	118 (+ 2 %)	48 (- 1 %)	51 (+ 1 %)	514 (+ 1 %)	47 (- 0 %)
Self-emp-inc	273 (+ 2 %)	279 (- 1 %)	14 (- 1 %)	79 (+ 0 %)	10 (- 0 %)	226 (- 0 %)	35 (- 0 %)
Without-pay	0 (- 0 %)	9 (+ 0 %)	0 (- 0 %)	0 (- 0 %)	0 (- 0 %)	3 (- 0 %)	1 (+ 0 %)
Never-worked	0 (- 0 %)	1 (- 0 %)	1 (+ 0 %)	0 (- 0 %)	0 (- 0 %)	2 (+ 0 %)	0 (- 0 %)
Sum	5355	10501	1175	1723	514	7291	1067

Il caractérise le degré de liaison entre les variables *Workclass* et *Education* . Il quantifie l'écart entre le tableau construit sur les données et le tableau que l'on aurait obtenu si l'hypothèse d'indépendance entre les variables était vraie. Pour nos deux variables, le *Test du Khi 2* est égal à 2572,10.

Observons la p-value de la statistique du KHI-2 de la figure ci-dessus, elle permet de déterminer s'il y a lieu de rejeter ou non l'hypothèse d'indépendance (H0), si elle est inférieure au niveau de signification (par défaut 5% sur Tanagra). Dans notre cas elle est de 0 % donc il ne faut pas rejeter H0. Le CHI-2 étant un critère additif, il serait intéressant pour nous de savoir quelles sont les cases qui y ont le plus contribué.

Les contributions sont indiquées en pourcentage du total. Le signe apparu dans les cases permet de déterminer s'il s'agit d'une attraction ou une répulsion entre les caractéristiques des variables étudiées. Lorsque la contribution d'une case est 2 fois plus élevée que la contribution moyenne, elle est surlignée en rouge, d'où la présence de la couleur rouge dans notre tableau. On peut déduire que la liaison entre la variable secteur de travail(workclass) et Education repose avant tout sur une forte attraction évaluée à 19 % entre la valeur Federal-gov et le niveau d'éducation Masters.

APPRENTISSAGE SUPERVISÉ

V. APPRENTISSAGE SUPERVISÉ

1 - Introduction

L'apprentissage supervisé (supervised learning en anglais) est une technique d'exploitation des données du passé ou bien de données connues pour prévoir ou expliquer le futur. Le résultat est un modèle prédictif, évalué au moment de sa production et en cours d'utilisation. Ce modèle doit avoir les qualités comme la précision et la robustesse en généralisation

2 - Définition mathématique

Une base de données d'apprentissage est un ensemble de couples entrée-sortie tel que

$$(x_n, y_n)_{1 \leq n \leq N} \text{ avec } x_n \in X \text{ et } y_n \in Y,$$

que l'on considère être tirées selon une loi sur $X \times Y$ inconnue, par exemple x_n suit une loi uniforme et $y_n = f(x_n) + w_n$ où w_n est un bruit centré.

La méthode d'apprentissage supervisé utilise cette base d'apprentissage pour déterminer une représentation compacte de f notée g et appelée *fonction de prédiction*, qui à une nouvelle entrée x associe une sortie $g(x)$.

Le but d'un algorithme d'apprentissage supervisé est donc de généraliser pour des entrées inconnues ce qu'il a pu «apprendre» grâce aux données déjà traitées par des experts, ceci de façon «raisonnable».

On distingue deux types de problèmes solvables avec une méthode d'apprentissage automatique supervisée:

- $Y \subset \mathbb{R}$: lorsque la sortie que l'on cherche à estimer est une valeur dans un ensemble continu de réels, on parle d'un problème de régression.
- $Y = \{1, \ldots, I\}$: lorsque l'ensemble des valeurs de sortie est fini, on parle d'un problème de classification, qui revient à attribuer une *étiquette* à chaque entrée.

3 - Méthodes d'apprentissage supervisé

Parmi les méthodes d'apprentissage supervisé on peut citer

Boosting - Machine à vecteurs de support - Mélanges de lois -
Réseau de neurones - Méthode des k plus proches voisins - Arbre de décision
Classification naïve bayésienne - Inférence grammaticale - Espace de versions

4 - Buts de l'apprentissage supervisé

Le but de l'apprentissage supervisé peut être une prédiction alors dans le cas ci on doit

– Trouver un modèle ou une application qui va lier correctement les variables explicatives et les cibles (régression linéaire multiple, régression logistique, etc.)

ou une classification, dans ce cas

– Ajuster un modèle ou un arbre de décision qui associe correctement les variables d'entrée aux classes en sortie (analyse discriminante, arbres d'induction, réseaux neuronaux, SVM, etc.)

Notre Travail

5 - Énoncé du problème

Nous avons un jeu de données sur un recensement qui permet, à partir d'un ensemble de paramètres de déterminer les individus qui gagnent un salaire annuel de moins ou de plus de 50K. Nous allons appliquer deux méthodes d'apprentissage afin d' associer correctement les variables d'entrée aux classes en sortie. L'objectif est d'étudier un ensemble de paramètre et d'établir une tache de prévision de classe afin de déterminer le profil des personnes qui gagnent plus ou moins de 50K par année à partir de données d'entrées telles que l'age, le niveau d'éducation, le secteur de travail, le sexe, le nombre d'heures de travail par semaine, le statut, et la situation familiale, etc.. . Dans ce jeu de données, les attributs des exemples sont à la fois quantitatif et qualitatif. Il est question dans ce travail de trouver un modèle capable de lier correctement nos variables explicatives (Age, éducation, secteur de travail, ..) avec notre variable cible (salary-per-year – le gain annuel) : c'est un apprentissage supervisé. Pour le faire la première étape et la plus importante est le choix d'une méthode d'analyse appropriée.

6 - Choix de la méthode

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. Dans notre cas la variable cible (Salary-per-Year) est qualitative (<=50K, >50K) et tandis que les variables explicatives sont à la fois quantitatives et qualitatives, nous appliquons donc la méthode d'apprentissage de l'**arbre de décision** par l'application de l'algorithme ID3.

7 - Présentation de la méthode d'arbre de décision

Un arbre de décision est un graphe orienté acyclique dont les nœuds correspondent aux variables choisies sur la base de critères de qualité, quant aux arcs, ils représentent les modalités d'une variable prédictive. Les nœuds terminaux sont appelés feuilles et évoquent les classes. La construction de l'arbre consiste à partitionner les données selon la variable explicative la plus discriminante. Ce processus est répété localement sur chaque nœud de l'arbre jusqu'à l'obtention de feuilles pures (correspondant à des feuilles constituées d'individus d'une même classe), ou sur ordre d'un arrêt volontaire de la progression de l'arbre. Les différents nœuds de l'arbre sont caractérisés par la distribution des effectifs de la population cible.
Les performances de prédiction dépendent directement de la taille de l'arbre appris. Une première difficulté concerne le choix des variables persistantes. Ce choix est fait sur la base d'un critère de séparation. Parmi les critères les plus fréquemment utilisés figurent:

- L'entropie de Shannon, applicable à tout type de variable explicatives. Cette mesure est notamment utilisée par Quillan dans C4.5 et C5.0 pour mesurer l'incertitude :

Entropie(noeud$_t$) ¿ $\sum_{i=1}^{k} (-fi \log 2 fi)$

L'algorithme CART produit des arbres de décision binaires et applique l'indice de Gini appelé entropie quadratique pour sélectionner des variables explicatives de tout type

Gini(nœud$_t$)= $1-fi$ $fi(¿)$ ¿ ¿ $\sum_{i=1}^{k}$ ¿

L'algorithme ID3(Inductive Decision Tree)
L'algorithme CHAID (Chi-Square Automatic Interaction Detection),
L'algorithme QUEST (Quick, Unbiased, Efficient Statistical Trees),

pour ne citer que ceux-ci.

8 - Particularités de l'arbre de décision

L'arbre de décision met l'accent sur sur la convivialité et l'intelligibilité (ou la lisibilité) des résultats;

=> en classification supervisée: la sortie de résultats sous la forme de règles logiques de classification: "SI tel ensemble de conditions sur telles variables est satisfait ALORS le cas appartient à telle classe".

=> résultats plus facilement interprétables et donc exploitables => communication plus aisée avec les spécialistes du domaine traité.

Principes de la méthode de l'arbre de décision

Nous pouvons établir ces principes en 2 phases

– Phase1: Construction

Sur base d'un ensemble d'apprentissage, processus récursif de division (souvent binaire) de l'espace des données en sous régions de plus en plus pures en terme de classes (estimé sur base d'un critère).

– Phase 2: Élagage ("pruning"):

On supprimer les branches (parties terminales) peu représentatives pour garder de bonnes performances prédictives (généralisation) alors on a la nécessité d'un critère pour désigner les branches à élaguer. Après élagage, les nouvelles feuilles sont labellisées sur base de la distribution des exemples d'apprentissage (classe majoritaire). Dans cette phase on calcule le taux d'erreur en test ou par validation croisée des différents sous arbres et on retient le plus bas possible. le mécanisme employé durant cette étape est appelé le coût complexité minimal.

Le temps d'apprentissage est certes plus long mais les performances de l'arbre sont les meilleures. Ces méthodes utilisent un critère d'élagage basé sur l'estimation du taux d'erreur de classification

Le gain d'information

Les algorithmes d'arbres de décision procèdent généralement au calcul du gain en information. Cette mesure fournit l'information gagnée après la séparation du nœud parent en nœuds fils. L'attribut discriminant sélectionné est celui qui dispose du gain informationnel maximal et donc d'un besoin d'information minimal. Il est donné par la différence :

Gain = critère du parent – ∑critère des fils

Une fois l'arbre de segmentation construit, on le défait progressivement pour générer les règles de décision. Le modèle résultant correspond à l'ensemble des chemins menant de la racine à une feuille. Ainsi, pour prédire la classe associée à un nouvel objet, il suffit de lui faire parcourir l'arbre de la racine jusqu'à l'une de ses feuilles, en prenant soin de vérifier toutes les conditions sur les meilleures variables au sens des critères évoqués ci-dessus, pour finalement lui attribuer la classe associée à feuille terminale dans laquelle elle se trouve.

Un arbre de décision est déclaré correct et complet et complet si tous les individus étiquetés sont correctement classifiés. Cette configuration idéale n'est jamais atteinte sur des applications réelles. Afin de s'approcher de cette solution, des opérations de pré ou post-élagage de l'arbre sont souvent nécessaires. Elles consistent à supprimer les feuilles les moins significatives et à les remplacer par un nœud terminal qui représente la classe majoritaire des individus classés par cette partie de l'arbre.

9 - Apports et limites des arbres de décision

Les arbres de décision ont de nombreux avantage et se comptent parmi les méthodes de fouille de données de données les plus appréciées. Parmi les avantages on cite :

- Elles permettent de traiter d'importants volumes de données hétérogènes (catégorielles ou numériques), pour des temps de calculs faibles.
- La méthode est de nature exploratoire, elle nécessite peu d'hypothèse sur les données.
- Les chemins résumant des décisions transcrites sous forme de règles (Si... alors) sont compréhensibles et donc facilement interprétables par un utilisateur non initié.
- Les arbres de décision n'en souffrent pas d'outiliers. Ces valeurs extrêmes sont isolées dans des feuilles et sont facilement détectables. Cependant, les arbres extrêmes sont isolés dans des feuilles et sont facilement détectables.

Les arbres de décision souffrent de quelques inconvénients :

- Le manque de précision dans les prédictions
- Face à un nombre trop important de classes, les arbres de décisions ont tendance à devenir très complexes et complètement illisibles.
- Impossible de revenir sur les affectations des niveaux antérieurs
- La recherche d'un arbre de décision optimal est un problème difficile
- Le problème de généralisation confronté au sur-apprentissage est omniprésent avec les arbres de décision.

10 - Construction de la base d'analyse

Les modèle que l'on construit ne représentent qu'un reflet approché de la réalité. Pour apprécier la capacité d'un modèle à bien représenter les données, la démarche classiquement utilisée consiste à séparer les instances en deux groupes: les données d'apprentissage (ou d'entraînement) et les données tests.

L'algorithme commence par apprendre et construire son modèle, type arbres de décision à partir du jeu de données d'apprentissage. Une fois l'apprentissage terminé, on mesure la fiabilité du modèle obtenu sur l'échantillon test qui doit être différent de l'échantillon d'apprentissage. Généralement on utilise 70% des individus pour apprendre le modèle et le reste pour mesurer la performance.

11- L'algorithme ID3 (Inductive Decision Tree)

Pour mettre en application la méthode de l'arbre de décision, on applique l'algorithme ID3.

L'algorithme ID3 a été développé à l'origine par Ross Quinlan. Il a tout d'abord été publié dans le livre "Machine Learning" en 1986.

C'est un algorithme de classification supervisé, c'est-à-dire qu'il se base sur des exemples déjà classés dans un ensemble de classes pour déterminer un modèle de classification. Le modèle que produit ID3 est un arbre de décision. Cet arbre servira à classer de nouveaux échantillons.

12 - Principe Général de l'Algorithme ID3

Chaque exemple en entrée est constitué d'une liste d'attributs. Un de ces attributs est l'attribut «cible» et les autres sont les attributs «non cibles». On appelle aussi cette "cible" la "classe". En fait l'arbre de décision va permettre de prédire la valeur de l'attribut «cible» à partir des autres valeurs. Bien entendu, la qualité de la prédiction dépend des exemples: plus ils sont variés et nombreux, plus la classification de nouveaux cas sera fiable.

Un arbre de décision permet de remplacer un expert humain dont il modélise le cheminement intellectuel. À chaque nœud correspond une question sur un attribut non cible. Chaque valeur différente de cet attribut sera associée à un arc ayant pour origine ce nœud. Les feuilles de l'arbre, quant à elles, indiquent la valeur prévue pour l'attribut cible relativement aux enregistrements contenus par la branche (indiqués par les différents arcs) reliant la racine à cette feuille.

ID3 construit l'arbre de décision récursivement. À chaque étape de la récursion, il calcule parmi les attributs restant pour la branche en cours, celui qui maximisera le gain d'information. C'est-à-dire l'attribut qui permettra le plus facilement de classer les exemples à ce niveau de cette branche de l'arbre. On appelle ce calcul l'entropie de Shannon.

13 - Pseudo-code de l'algorithme

```
fonction ID3(exemples, attributCible, attributsNonCibles)
  si exemples est vide alors /* Nœud terminal */
    retourner un nœud Erreur
  sinon si attributsNonCibles est vide alors /* Nœud terminal */
    retourner un nœud ayant la valeur la plus représentée pour attributCible
  sinon si tous les exemples ont la même valeur pour attributCible alors /* Nœud terminal */
    retourner un nœud ayant cette valeur
  sinon /* Nœud intermédiaire */
    attributSélectionné = attribut maximisant le gain d'information parmi attributsNonCibles
    attributsNonCiblesRestants = suppressionListe(attributsNonCibles, attributSélectionné)
    nouveauNœud = nœud étiqueté avec attributSélectionné

    pour chaque valeur de attributSélectionné faire
      exemplesFiltrés = filtreExemplesAyantValeurPourAttribut(exemples, attributSélectionné, valeur)
      nouveauNœud->fils(valeur) = ID3(exemplesFiltrés, attributCible, attributsNonCiblesRestants)
    finpour

    retourner nouveauNœud
```

14 - Application de la méthode choisie aux données

Les paramètres du modèle

Les paramètres utilisés dans notre modèle sont donc :

- Les variables explicatives : Age, Sexe, Éducation, Statut Marital, Secteur de Travail, Occupation, Nombre d'heures de Travail par semaine, Capital-gain, Capital-loss, Race, Pays
- La variable à prédire : Salaire Annuel (Salary-per-week ,,,,errata de Salary-per-Year)
- L'entropie: 0,03

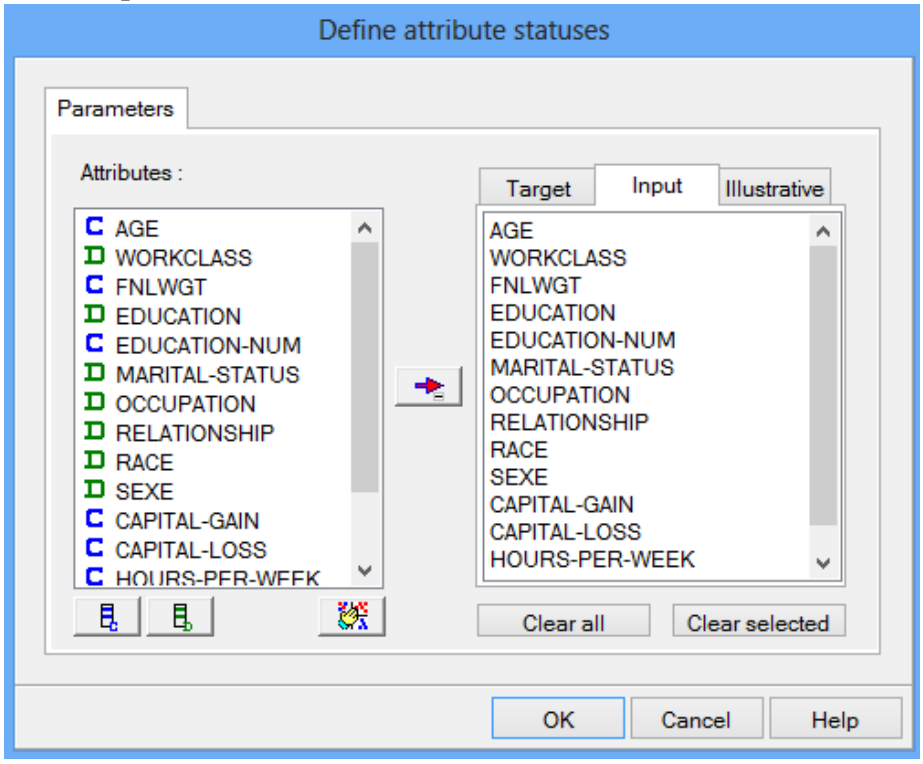
15- Ensemble d'Entraînement et Données de Test

L'ensemble d'entraînement considéré représente 70% de nos données et les données de Test représentent 30% ce qui permet de minimiser le taux d'erreur. La performance du modèle est calculée par la détermination du taux d'erreur sur le nombre d'instance du jeu de données Test. Il est possible de calculer le taux d'erreur pour chacune d'erreur dans chacune des classes mais l'inconvénient de cette approche c'est qu'il faut beaucoup d'individu pour l'apprentissage.

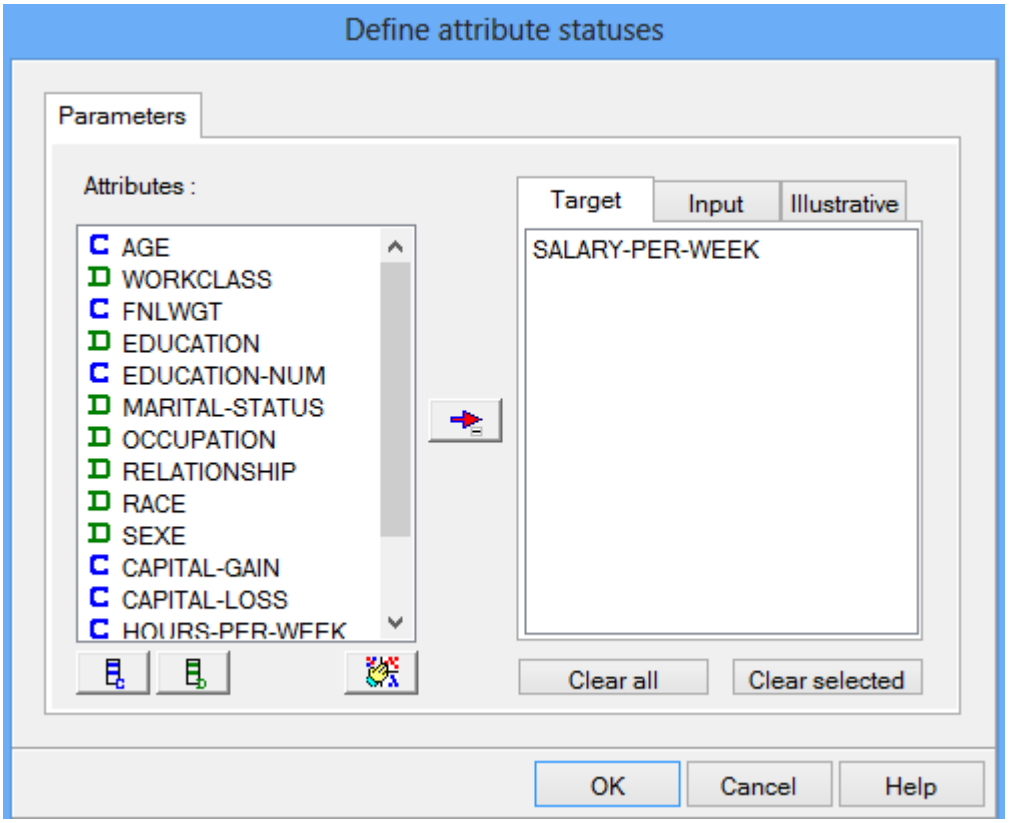
16 - Résultats des expérimentations

Nous allons voir dans cette partie comment mettre en œuvre l'apprentissage supervisé sous Tanagra. Nous allons mettre en œuvre la méthode d'apprentissage arbre de décision par le biais ID3 (apprentissage supervisé). Notre but ici est de connaître les variables qui jouent important dans la valeur du Salaire annuel d'un individu. A l'aide d'un arbre de décision fait sur un ensemble d'exemple et des tests réalisés sur des entraînements on pourra même mesurer le pourcentage d'erreur et ainsi connaître la fiabilité de l'arbre de décision.
En supposant que les données sont déjà chargées dans Tanagra, on va suivre les étapes suivantes:

- Définir le statut des variables
    1. Pour cela se placer sur le nœud «Dataset» et ajouter un opérateur **Define Status** en cliquant sur son icône dans la barre des raccourcis. La fenêtre de dialogue permettant de définir le statut des variables apparaît automatiquement.
    2. Assurez-vous que c'est l'onglet « Input » qui est actif. Sélectionnez les variables continues de la liste en cliquant sur le bouton correspondant (cf ci-dessous), et cliquez enfin sur le bouton flèche pour les passer dans la liste des Input.



3. Toujours en restant dans la fenêtre de dialogue, activez l'onglet Target. Cliquez sur la variable « class » pour la sélectionner, puis sur le bouton flèche. Ici notre Target sera le Salaire Annuel.

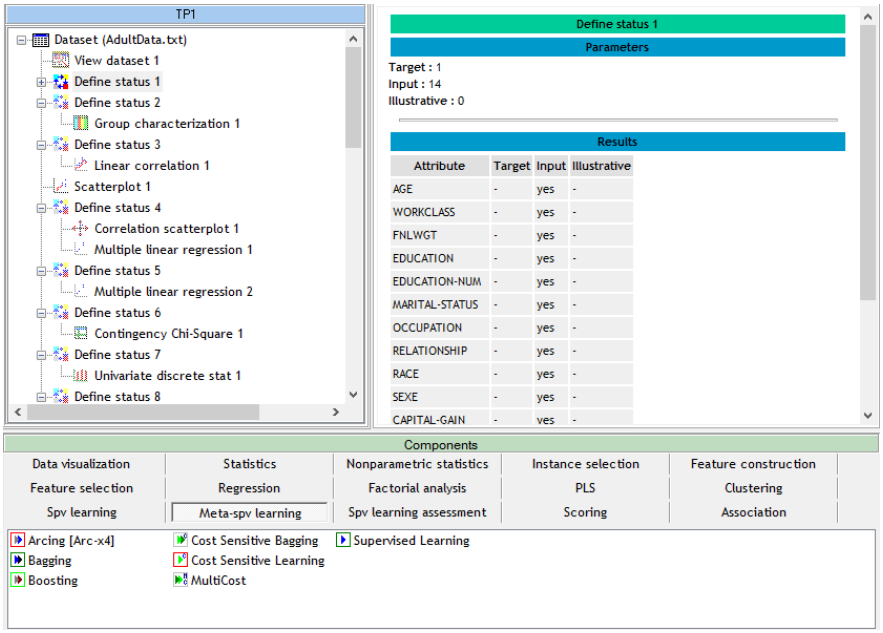


4. Après avoir défini la variable à prédire (« class » = Target) et les variables explicatives (les autres = Input). En appuyant sur OK, on valide et on ferme la fenêtre.

- Choisissons maintenant un méta-opérateur d'apprentissage supervisé

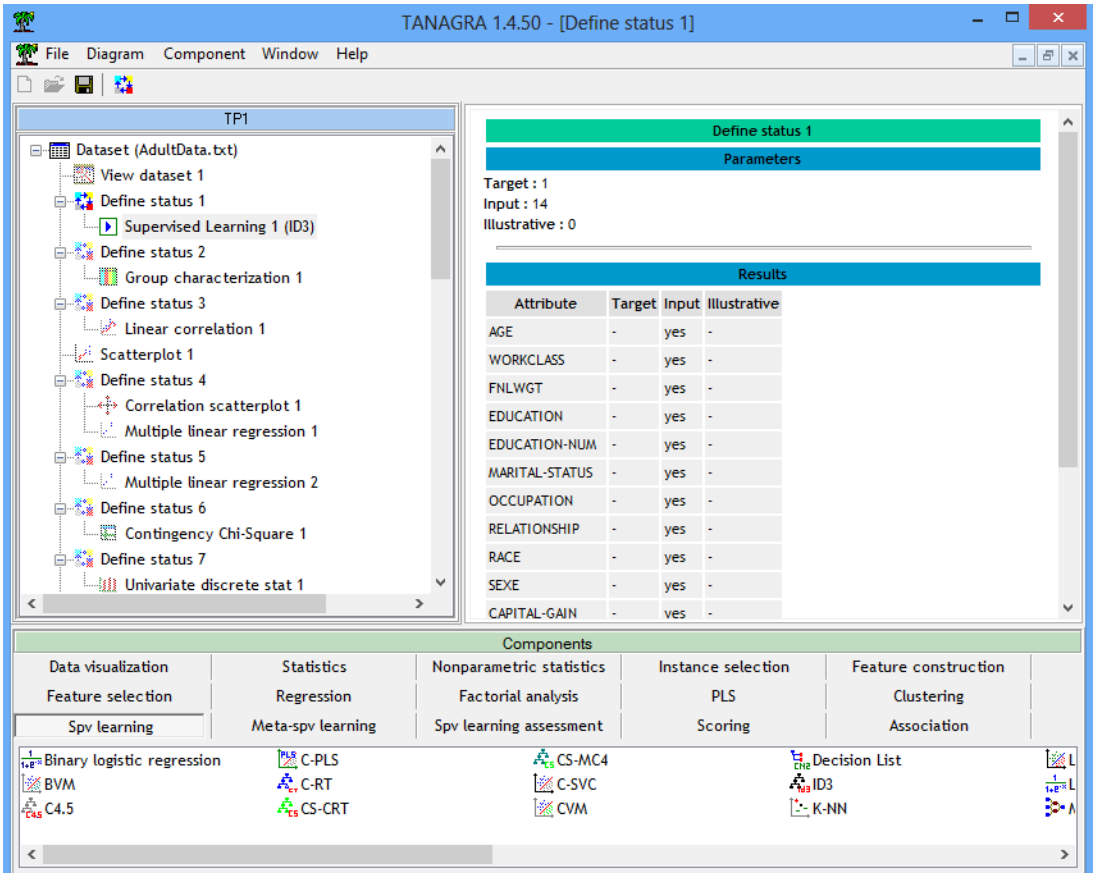
  Tanagra oblige à utiliser un méta-opérateur. Il en propose toutefois un pour les lancements uniques de méthode. Il s'agit de l'opérateur **Supervised Learning.**

1. Ajoutons un opérateur **Supervised learning** (onglet META-SPV LEARNING) au diagramme, sous le nœud «**Define status 1**».



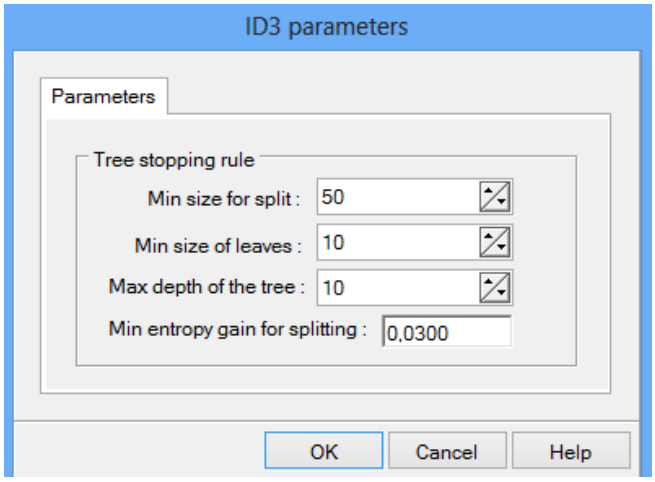
- Choisissons un opérateur d'apprentissage supervisé

1. Dans la palette des opérateurs, on clique sur l'onglet **SPV LEARNING**, et fait glisser un opérateur ID3 sur le nœud « **Supervised Learning** » que nous venons d'ajouter. L'opérateur est inclus dans le méta-opérateur, aussi voit-on son libellé dans celui du nœud du méta-opérateur, et non pas en-dessous de celui-ci.



- Définissons maintenant les paramètres de l'apprentissage (opérateur ID3)

1. On fait apparaître le menu contextuel du nœud « Supervised learning (ID3) » par clic droit sur ce dernier. En plus de la commande **Parameters**… habituelle, on trouve une commande « **Supervised parameters** »
2. Dans la fenêtre de dialogue qui s'affiche, compte tenu de la taille du fichier étudié (32561 individus), on modifie les paramètres de ID3 comme suit :



3. Validons en cliquant sur OK.

- Maintenant on va effectuer l'apprentissage

1. Dans le menu contextuel du nœud ID3, on choisit View. Les résultats s'affichent dans le cadre de droite.

ID3 parameters	
Size before split	50
Size after split	10
Max depth of leaves	10
Goodness of split thresold	0,0300

Results

Classifier performances

Error rate			0,1400			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		<=50K	>50K	Sum
<=50K	0,9469	0,1218	<=50K	23408	1312	24720
>50K	0,5860	0,2221	>50K	3246	4595	7841
			Sum	26654	5907	32561

Le taux d'erreur calculé sur l'apprentissage paraît moyen (11,400 %). On voit dans la matrice de confusion que l'erreur se répartit également entre avoir un salaire ***hebdomadaire <50 k*** et avoir un ***salaire >50K***. On constate dans l'arbre retranscrit l'importance de la variable **RELATIONSHIP** dans le diagnostic automatique. Comme le montre le début de l'arbre de décision suivant:

Data description

Target attribute	SALARY-PER-WEEK (2 values)
# descriptors	14

Tree description

Number of nodes	1286
Number of leaves	1104

Decision tree

- RELATIONSHIP in [Not-in-family]
  - CAPITAL-GAIN < 8296,0000
    - EDUCATION in [Bachelors]
      - AGE < 27,5000 then SALARY-PER-WEEK = <=50K (97,99 % of 498 examples)
      - AGE >= 27,5000
        - OCCUPATION in [Adm-clerical] then SALARY-PER-WEEK = <=50K (94,74 % of 114 examples)
        - OCCUPATION in [Exec-managerial]
          - HOURS-PER-WEEK < 44,5000
            - AGE < 36,5000 then SALARY-PER-WEEK = <=50K (100,00 % of 46 examples)
            - AGE >= 36,5000
              - HOURS-PER-WEEK < 39,0000 then SALARY-PER-WEEK = <=50K (100,00 % of 15 examples)
              - HOURS-PER-WEEK >= 39,0000
                - FNLWGT < 103865,5000 then SALARY-PER-WEEK = <=50K (100,00 % of 11 examples)

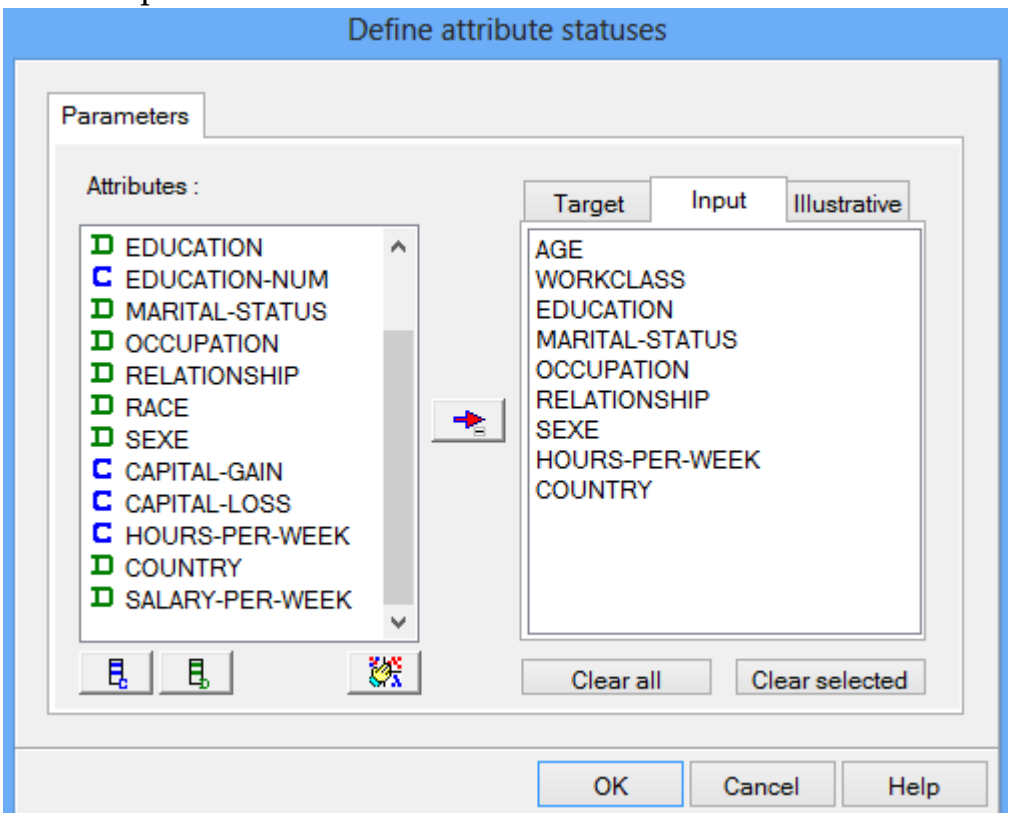
On constate aussi que l'arbre de décision est extrêmement long. On se propose donc d'éliminer quelques paramètres qui sont négligeables ou équivalents. Le critère qu'on utilise pour faire cela est d'éliminer les paramètres qui n'interviennent pas ou presque pas dans la prédiction du salaire hebdomadaire.

Ces paramètres sont:
FNLWGT (qui est un coefficient calculé à partir des autres paramètres de l'individu il est assez normal qu'il n'a pas un poids dans le salaire final parce que il est calculé à partir des autres paramètres intervenants dans les paramètres d'entrée de la prédiction).

EDUCATION_NUM qui est équivalent au paramètre EDUCATION donc il faut le supprimer des entrées pour éviter la redondance.

RACE qui n'a pas d'importance sur l'arbre de décision fait avec toutes les entrées.
CAPITAL-GAIN et *CAPITAL-LOSS* qui sont des variables équivalentes et qui n'ont pas d'importance dans l'arbre de décision final.

Ainsi les paramètres d'entrées retenus sont les suivants:



Alors, le taux d'erreur calculé sur l'apprentissage avec ces nouveaux paramètres paraît toujours moyen mais moins bon que le précédent (15,34 %). En diminuant les entrées on obtiendra un taux d'erreur qui n'est pas acceptable il est alors préférable de tester une autre méthode d'apprentissage et de comparer les résultats obtenus avec ce qui a été fait par l'arbre de décision. Nous allons choisir la méthode d'apprentissage SVM.

Apprentissage avec SVM

L'objectif de SVM c'est de trouver une séparation généralement non-linéaire en utilisant des noyaux et assurer que la séparation a de bonnes propriétés statistiques pour éviter le sur-apprentissage.

Deux étapes sont mises en évidence au cours de SVM

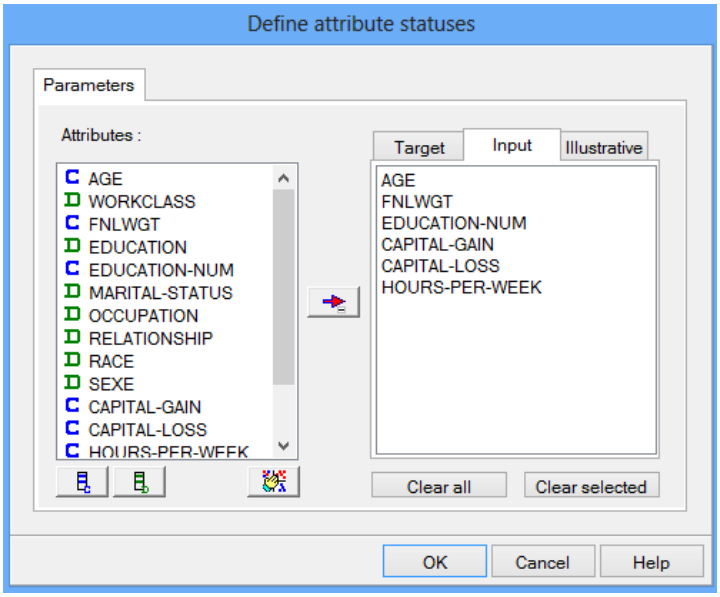
La transformation non linéaire Φ pour passer dans un espace de dimension plus grande que l'espace d'origine, mais doté d'un produit scalaire.

Dans cet espace, on cherche un séparateur linéaire f (x) = ax + b (par ex. : fonction discriminante de Fisher), qui est un hyperplan optimal séparant bien les groupes (précision du modèle) f (x) > 0 ⇒ classe A ; f (x) ≤ 0 ⇒ classe B

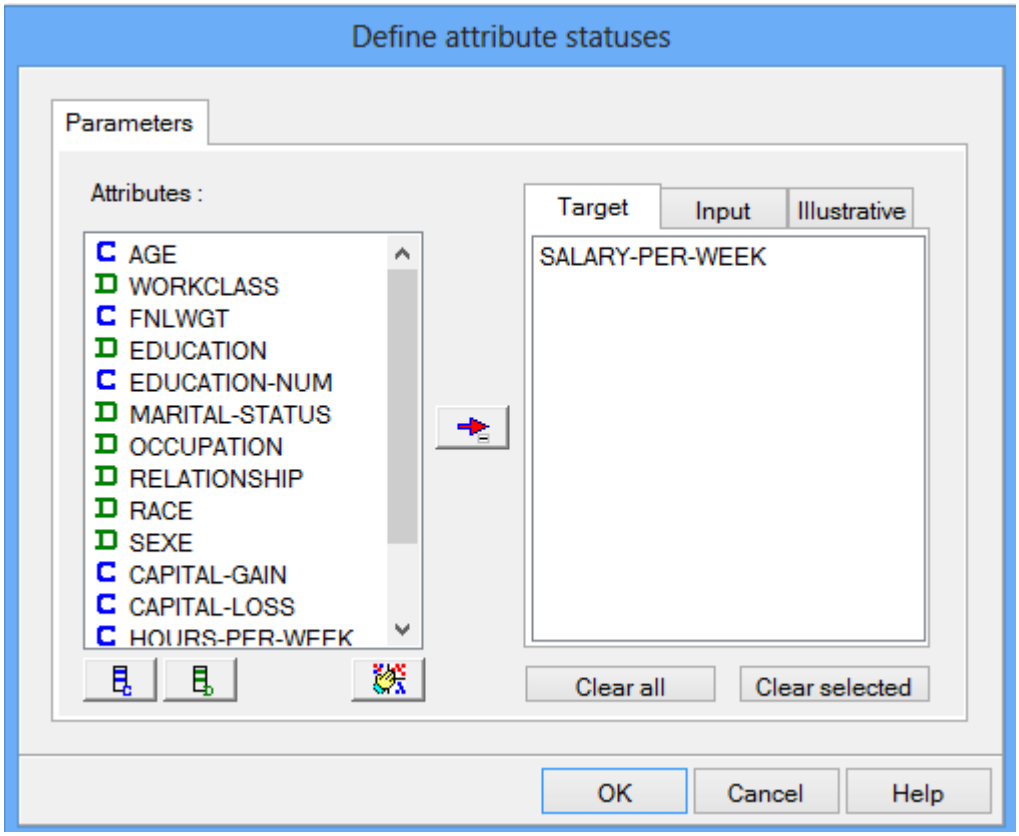
le plus loin possible de tous les cas et enfin on exprime f (Φ(x)) sans faire intervenir explicitement Φ.

Dans Tanagra, cette technique prend en entrée uniquement des paramètres continus. Donc il faut voir que l'input RELATIONSHIP qui jouait le rôle le plus important dans l'apprentissage avec l'arbre de décision n'apparaîtra pas dans les entrées avec SVM. Donc le taux d'erreur risquera d'être plus élevé.

- Définition des paramètres d'entrée (uniquement les attributs continus)



- Définition du paramètre de sortie (binaire)



- En procédant comme précédemment c'est-à-dire choisir **Supervised Learning** puis **SVM** on obtient le résultat suivant :

Classifier performances

Error rate	0,1943
Values prediction	Confusion matrix

Value	Recall	1-Precision
<=50K	0,9694	0,1886
>50K	0,2896	0,2500

	<=50K	>50K	Sum
<=50K	23963	757	24720
>50K	5570	2271	7841
Sum	29533	3028	32561

En lisant la section Confusion matrix, nous constatons que le taux d'erreur et de 19.43% ce qui est plus élevé que le taux d'erreur enregistré avec l'arbre de décision qui était de 11.4%. Ceci est sans aucun dû à l'absence du paramètre RELATIONSHIP dans l'input de l'apprentissage avec SVM dû au fait qu'il n'est pas numérique.

Le tableau ci-dessous montre le poids de chaque input dans la prédiction du salaire hebdomadaire. Nous constatons que l'attribut le plus important est EDUCATION-NUM avec un poids de 0.05 qui est d'ailleurs sensé car dans la vie réelle le diplôme joue dans la

plupart du temps un rôle important dans le salaire final. Il est aussi force de constater que le paramètre FNLWGT qui est le poids obtenu à partir de tous les autres paramètres n'a pas d'importance dans la prédiction car se paramètre résume tous les autres qui sont eux même présents déjà dans la liste d'inputs.

Classifier characteristics

Data description

Target attribute	SALARY-PER-WEEK (2 values)
# descriptors	6

Linear classifier

"Reference" class value : >50K

Attribute	Weight
AGE	0,006393
FNLWGT	0,000000
EDUCATION-NUM	0,054511
CAPITAL-GAIN	0,000211
CAPITAL-LOSS	0,000755
HOURS-PER-WEEK	0,006764
constant	-2,163494

Comparaisons des 2 méthodes

Dans Tanagra, SVM n'accepte que les variables continues comme paramètres d'entrée pour faire l'apprentissage. Or il se trouve que l'attribut RELATIONSHIP joue le rôle le plus important dans la détermination du salaire annuel (<= 50K ou >50K). C'est la raison pour laquelle le modèle obtenue avec ID3 a un taux d'erreur de 11.4 % et celui obtenue avec SVM a un taux d'erreur de 19.43 %. De plus on remarque que parmi les attributs continus, le plus important est EDUCATION-NUM qui représente le niveau d'éducation de l'individu. Donc on retient à partir des deux méthodes d'apprentissages appliquées sur notre jeux de données que le niveau d'éducation et le statut marital d'un individu jouent un rôle important dans sa rémunération.

VI. CONCLUSION

Ce travail pratique effectué dans le cadre du cours de Fouille de données avait pour but de définir un modèle capable de prédire les facteurs qui influencent le salaire annuel d'un individu.
En conclusion, nous avons opté pour le modèle de l'arbre de décision par ses caractéristiques intéressantes. D'abord, il hiérarchise les variables par ordre d'importance dans l'arbre, celles qui sont situées au plus près de la racine sont les plus performantes. Ensuite, il synthétise de manière intelligible et visuel le résultat de l'analyse.
Lors de cette étude nous avons cherché à extraire les variables les plus pertinentes qui peuvent orienter les décideurs dans la prise de décision. A cette fin nous avons décidé de comparer les méthodes d'arbre de décision par l'intermédiaire de ID3 et et le SVM. On constate que le taux d'erreur avec ID3 est plus acceptable. Cependant, à l'analyse des résultats obtenus, nous remarquons que ce jeu de données ne nous donne pas un modèle parfait.

Le lien entre la variable cible, Salaire Annuel et les différentes variables explicatives a montré que l'attribut RELATIONSHIP joue le rôle le plus important dans la détermination du salaire annuel (<= 50K ou >50K). Cependant, on peut par contre avec un jeu de données plus important vérifier la pertinence de ce modèle et ses règles.

VII. RÉFÉRENCES

[1] http://www.analyse-donnees.fr/services-analyse-enquetes-traitement/

[2] http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names

[3] Cours de Fouilles de Données, Master 1 , IFI 2016, Enseignante NGUYỄN Thị Minh Huyền

Tutoriels:

- http://tutoriels-data-mining.blogspot.com/2009/05/analyse-factorielle-des-correspondances.html
- http://tutoriels-data-mining.blogspot.com/search/label/App.%20Supervis%C3%A9%20-%20Scoring