VERSION EXPERIMENTALE

Cette version de Sipina v 3.0 n'est pas, et ne sera jamais, définitive, elle sert d'outil de recherche, elle a plus une vocation d'outil d'expérimentation que de logiciel dédié au Data Mining en entreprise. C'est la raison pour laquelle elle est distribuée gratuitement pour la recherche et l'enseignement.

Ce petit descriptif associé à un tutoriel est la seule documentation existante sur ce logiciel. Il est loin d'être exhaustif et ne montre pas toutes les facettes de cette version, il permet en revanche de se donner une idée sur les manipulations de base que l'on peut effectuer.

Bonne lecture.

INSTALLATION ET DEMARRAGE DE SIPINA 3

Configuration requise

- Système d'exploitation Windows 95, Windows 98, Windows 2000 et Windows NT
- Ram 16 Mo conseillé.
- Taille disque nécéssaire 30 Mo

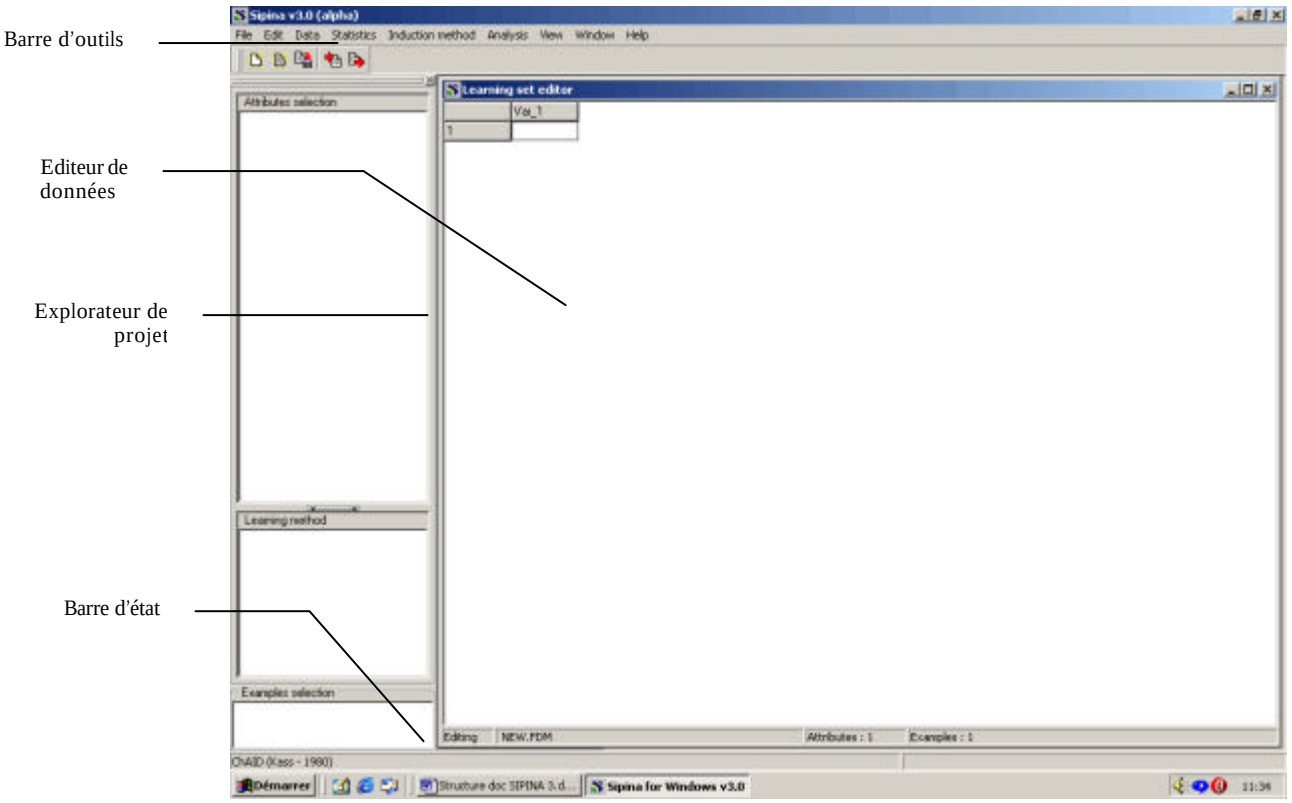
Installation de SIPINA 3 ®

- Démarrez votre ordinateur
- Si besoin est , connectez-vous en tant qu'administrateur (En cas de problème, veuillez vous renseigner auprès de votre administrateur réseau)
- Cliquez sur Démarrer
- Sélectionnez Exécuter…
- Tapez x:\setup.exe pour lancer l'installation où x représente la lettre assignée à votre lecteur de CD-ROM.

Démarrage de l'application

- Cliquez sur le bouton démarrer
- Sélectionnez l'option programmes puis Sipina for Windows
- Cliquez alors sur SIPINA

Au lancement de l'application, l'interface de SIPINA se présente sous la forme de différentes fenêtres ayant chacune une utilité particulière. Comme cela est souvent le cas, vous pouvez si vous désirez, cacher ou montrer ces différentes fenêtres. Pour ce faire il vous suffit de cliquer sur l'option de menu view . En cliquant sur le menu fenêtre vous pouvez réorganiser celles-ci comme bon vous semble, en cascade ou horizontalement.


Barre d'outils
Editeur de données
Explorateur de projet
Barre d'état
Sipina v3.0 (alpha)
File Edit Data Statistics Induction method Analysis View Window Help
Attributes selection
Learning method
Examples selection
Learning set editor
Var_1
1
Editing
NEW.PDM
Attributes : 1
Examples : 1
ChAID (Kass - 1980)
Démarrer
Structure doc SIPINA 3.d...
Sipina for Windows v3.0

Barre d’outils (Toolbars)

Cette barre d’outils reprend la plupart des options du menu File. Pour votre confort visuel, celle-ci peut être cachée en cliquant sur l’option View. De plus, vous pouvez la dissocier et la positionner où vous le désirez sur l’écran. Pour ce faire double-cliquez sur les doubles traits situés à gauche de la barre et déplacer cette dernière.

Les options

 Création d’un nouveau fichier de données

 Ouverture d’un fichier de données

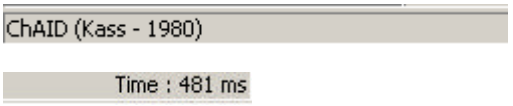
 Sauvegarde du fichier de données

 Importation d’un fichier de données (Format Texte)

 Exportation d’un fichier de données

Barre de statut (Status Bar)

Cette barre vous indique, à gauche, la méthode en cours d'utilisation et à droite le temps nécessaire à l'exécution de cette dernière.

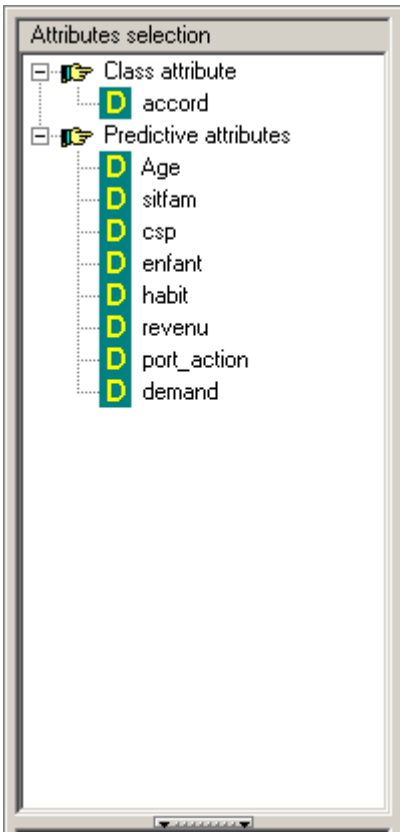


Information sur la méthode en cours d'utilisation.

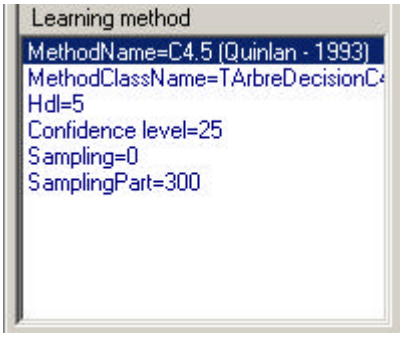
Temps d 'exécution

Explorateur de projet (Project explorer)

Cette fenêtre est découpée en trois modules



Dans cette exemple, nous avons sélectionné l'information *accord* comme variable à prédire (attribut) et les champs Age, situation familiale, csp etc.. comme variables prédictives (Predictive attributes).

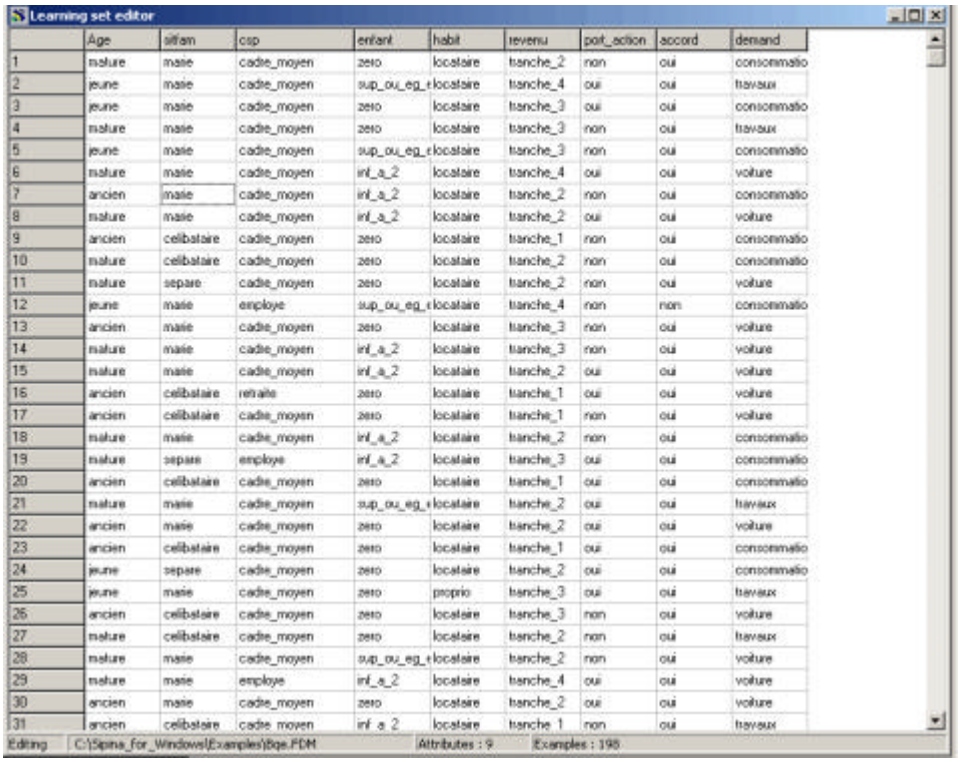


La méthode C4.5 (Quilan – 1993) est ici utilisée avec les paramètres suivants.

Pour ce projet nous avons décidé de travailler sur la totalité des enregistrements du fichier de données.

Editeur de données (Learning set editor)

Les données sont organisées sous la forme d'un tableau.



	Age	situfam	csp	enfant	habit	revenu	pot_action	accord	demand
1	mature	mase	cadre_moyen	zero	locataire	tranche_2	non	oui	consommatio
2	jeune	maie	cadre_moyen	sup_ou_eg_t	locataire	tranche_4	oui	oui	travaux
3	jeune	maie	cadre_moyen	zero	locataire	tranche_3	oui	oui	consommatio
4	mature	maie	cadre_moyen	zero	locataire	tranche_3	non	oui	travaux
5	jeune	maie	cadre_moyen	sup_ou_eg_t	locataire	tranche_3	non	oui	consommatio
6	mature	maie	cadre_moyen	inf_a_2	locataire	tranche_4	oui	oui	voiture
7	ancien	maie	cadre_moyen	inf_a_2	locataire	tranche_2	non	oui	consommatio
8	mature	maie	cadre_moyen	inf_a_2	locataire	tranche_2	oui	oui	voiture
9	ancien	celibataire	cadre_moyen	zero	locataire	tranche_1	non	oui	consommatio
10	mature	celibataire	cadre_moyen	zero	locataire	tranche_2	non	oui	consommatio
11	mature	separe	cadre_moyen	zero	locataire	tranche_2	non	oui	voiture
12	jeune	maie	employe	sup_ou_eg_t	locataire	tranche_4	non	non	consommatio
13	ancien	maie	cadre_moyen	zero	locataire	tranche_3	non	oui	voiture
14	mature	maie	cadre_moyen	inf_a_2	locataire	tranche_3	non	oui	voiture
15	mature	maie	cadre_moyen	inf_a_2	locataire	tranche_2	oui	oui	voiture
16	ancien	celibataire	retraite	zero	locataire	tranche_1	oui	oui	voiture
17	ancien	celibataire	cadre_moyen	zero	locataire	tranche_1	non	oui	voiture
18	mature	maie	cadre_moyen	inf_a_2	locataire	tranche_2	non	oui	consommatio
19	mature	separe	employe	inf_a_2	locataire	tranche_3	oui	oui	consommatio
20	ancien	celibataire	cadre_moyen	zero	locataire	tranche_1	oui	oui	consommatio
21	mature	maie	cadre_moyen	sup_ou_eg_t	locataire	tranche_2	oui	oui	travaux
22	ancien	maie	cadre_moyen	zero	locataire	tranche_2	oui	oui	voiture
23	ancien	celibataire	cadre_moyen	zero	locataire	tranche_1	oui	oui	consommatio
24	jeune	separe	cadre_moyen	zero	locataire	tranche_2	oui	oui	consommatio
25	jeune	maie	cadre_moyen	zero	proprio	tranche_3	oui	oui	travaux
26	ancien	celibataire	cadre_moyen	zero	locataire	tranche_3	non	oui	voiture
27	mature	celibataire	cadre_moyen	zero	locataire	tranche_2	non	oui	travaux
28	mature	maie	cadre_moyen	sup_ou_eg_t	locataire	tranche_2	non	oui	voiture
29	mature	maie	employe	inf_a_2	locataire	tranche_4	oui	oui	voiture
30	ancien	maie	cadre_moyen	zero	locataire	tranche_2	oui	oui	voiture
31	ancien	celibataire	cadre_moyen	inf_a_2	locataire	tranche_1	non	oui	travaux

Informations : Approchez le curseur de la barre des titres des colonnes pour en modifier la largeur.

Exemple d'application : différencier les bons des mauvais payeurs dans une demande de crédit

Le contexte

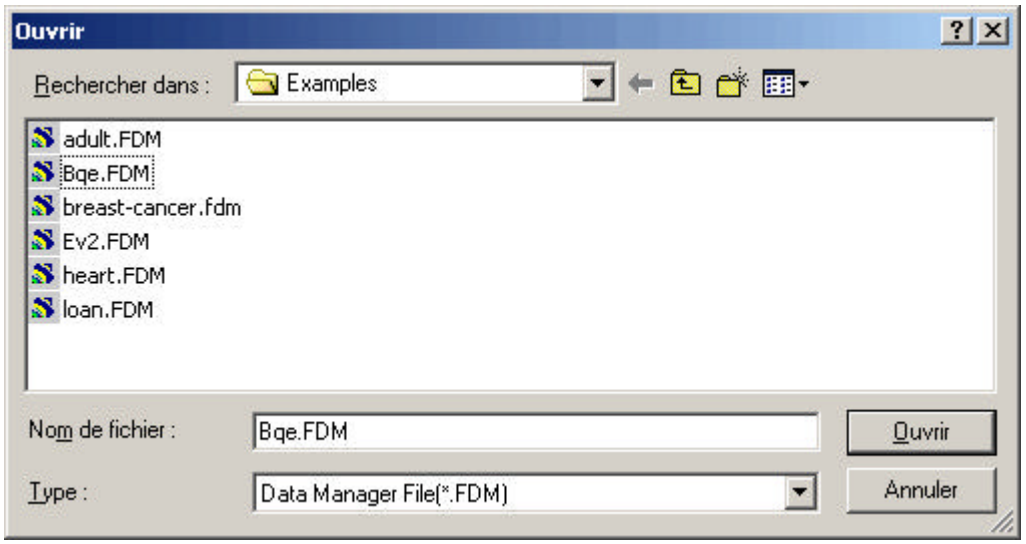
Votre entreprise	Etablissement de crédit
Votre activité	Vente de solutions de crédit aux particuliers.
Vos moyens	Vous disposez d'un fichier de données contenant les informations suivantes :
Votre objectif	Optimiser votre crédit checking par une meilleure connaissance de vos clients.

Ce que vous apprendrez

- Ouvrir un fichier de données.
- Choisir la méthode d'analyse.
- Définir les variables à prédire et prédictives.
- Sélectionner des individus. Filtrage des données.
- Apprentissage
- Interprétation des résultats.

Ouvrir le fichier de données

Sélectionner l'option de menu *file* puis *open* et ouvrir le fichier Bqe.FDM.

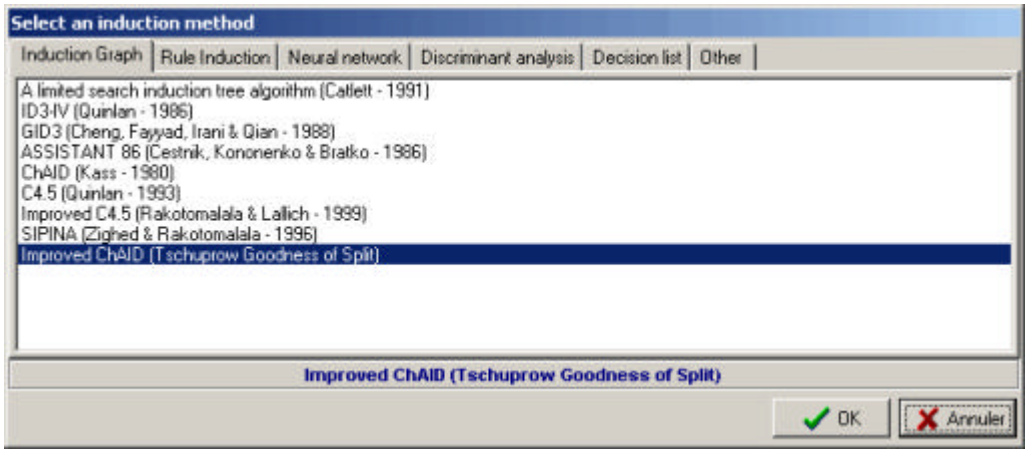


Choix de la méthode d'analyse

Les méthodes proposées par SIPINA 3 sont très nombreuses et ….. Dans notre exemple, nous allons travailler sur la méthode nommée Improved ChAID (Tschuprow Goodness of Split).

Sélection de méthode

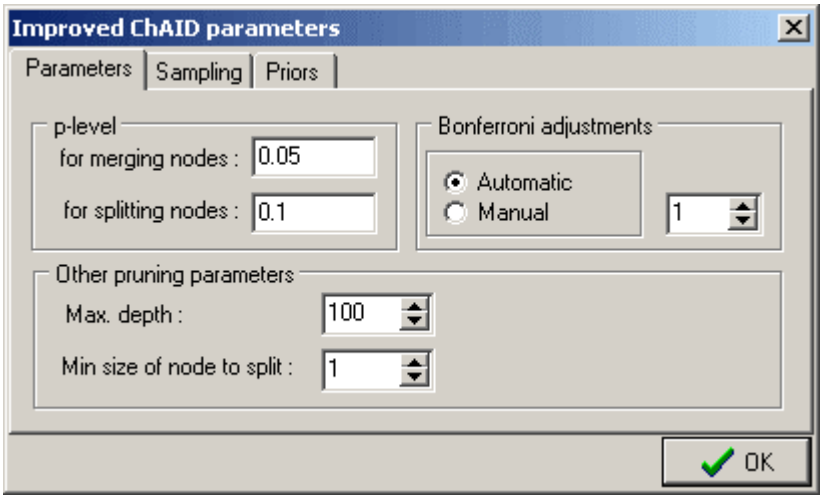
Cliquez sur l'option de menu *Induction method puis Standard algorithm*



Paramètres de la méthode

Comme pour toutes les méthodes, vous pouvez préciser les paramètres qui seront utilisés pour l'apprentissage des données. Dans notre exemple, nous allons seulement modifier le paramètre *For Splitting nodes* pour lui affecter la valeur de 0.1.

Cliquez sur la zone *for splitting nodes* et changer sa valeur en 0.1.



Cliquez sur OK

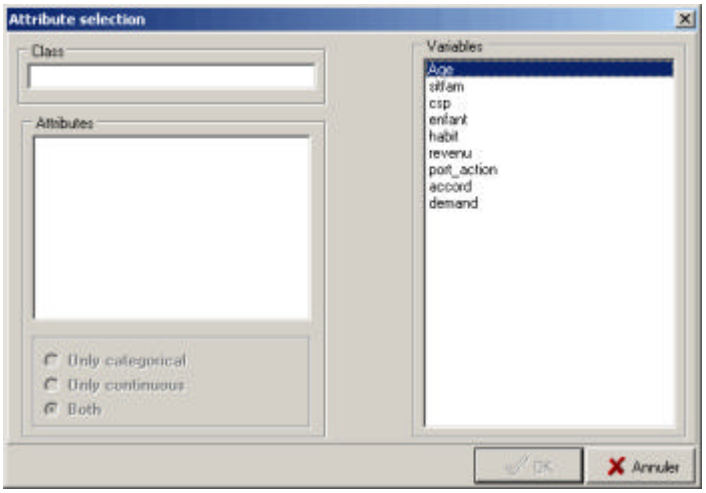
Informations

L'explorateur de projet vous affiche désormais les valeurs affectées à chacun des paramètres de la méthode. Si vous avez correctement effectué les opérations précédentes, la donnée Split doit avoir 0.1 comme valeur.

Variables à prédire et prédictives.

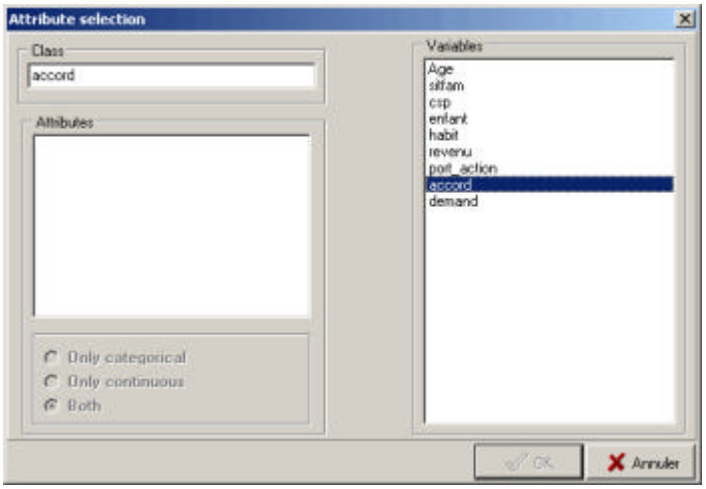
Jusqu'à là nous avons spécifié à SIPINA, la source des donnée puis la méthode qui sera appliquée à l'analyse. Il ne nous reste plus qu'a préciser ce que l'on souhaite savoir et quelle sont les informations dont SIPINA peut disposer pour nous répondre.

Définir la variable à prédire



Que voulez-vous savoir ?

De quelles informations SIPINA peut disposer ?

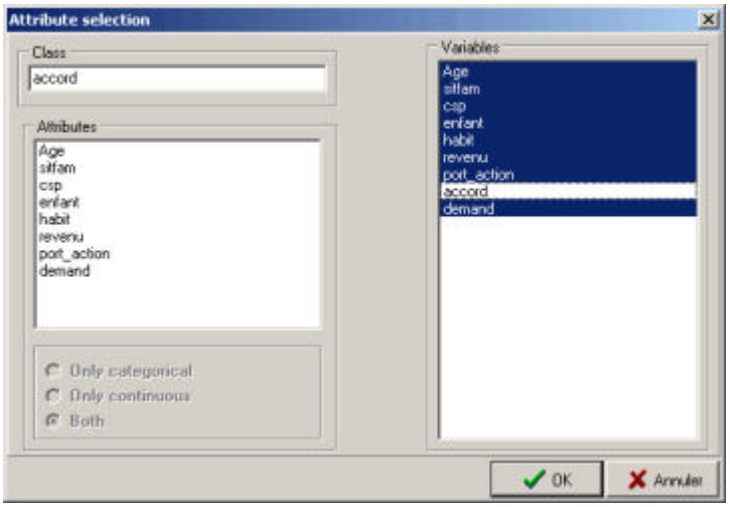


Dans notre exemple, l'utilisateur veut savoir si il peut donner son accord à l'attribution d'un crédit. Nous allons donc sélectionner l'information *accord* comme variable à prédire (attribut).

Pour ce faire cliquez sur la variable accord, maintenez le bouton de la souris appuyé et déplacez cet objet vers la zone Class.

Définir les variables prédictives

Choisissons maintenant le reste des informations, Age, situation familiale, csp etc.. comme variables prédictives (Predictive attributs)



Pour sélectionner plusieurs variables, cliquez sur la première (age), maintenez la touche shift enfoncée et cliquez sur la dernière (demand).

Déplacer cette sélection vers le zone attributes.

Remarquez que SIPINA a automatiquement retiré le champ accord. Ce qui vous l'avouerez est bien pratique.

Cliquez sur OK pour continuer.

Sélection d'individus

Nous avons ouvert notre fichier de données, précisé quelle méthode utiliser, paramétré notre variable à prédire ainsi que nos données prédictives. Si nous voulons nous pouvons nous arrêter là et lancer directement l'apprentissage.

Dans un but éducatif et surtout pour vous faire découvrir la richesse des fonctionnalités de SIPINA nous allons pousser plus en avant notre paramétrage.

Selon les cas, il peut effectivement être intéressant de travailler, soit sur la totalité du fichier de données, soit de se limiter à une partie de ce dernier.

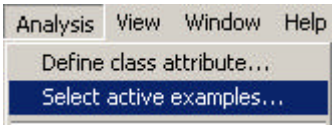
Choisir le mode de filtrage

SIPINA vous propose 4 modes de filtrage ou d'échantillonnage

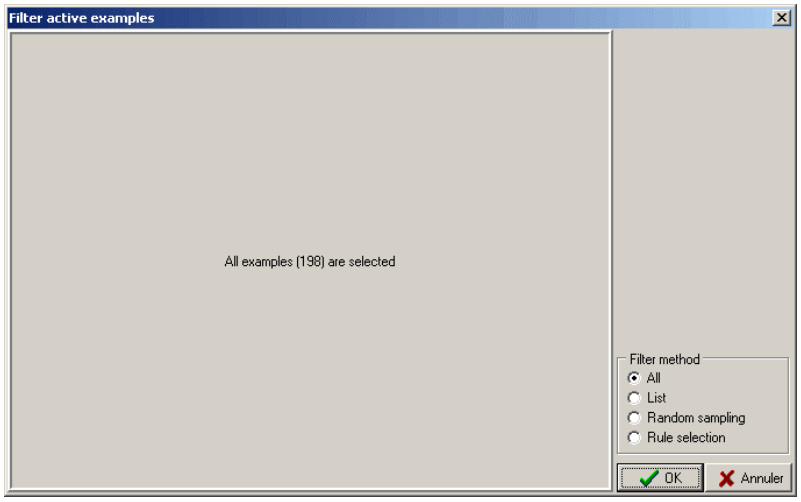
- Sans filtrage
  Comme son nom l'indique SIPINA travaillera sur la totalité des enregistrements de votre fichier de données.
- Filtrage manuel
  Dans ce cas, nous allons préciser manuellement le numéro du premier et du dernier enregistrement d'analyse. Par exemple nous allons demander à SIPINA de prendre en compte les enregistrements à partir du numéro 50 jusqu'au numéro 100.
- Echantillonnage aléatoire
  Cette fois nous allons laisser SIPINA faire son échantillonnage, en le limitant tout de même à un pourcentage. Par exemple vous pouvez lui demander de ne travailler que sur 50% des données.
- Filtrage suivant des règles
  Si vous n'avez pas trouvé votre bonheur avec les autres options, celle-ci vous satisfera certainement. En effet, SIPINA vous permet de filtrer vos données grâce à de puissantes règles de filtrage.

  Ce module sera abordé plus en détails dans notre prochain tutorial, mais sachez dés à présent que nous pourrons, par exemple, demander à SIPINA de ne travailler que sur les clients possesseur d'un portefeuille d'actions et qui, par exemple, ont demandé un crédit automobile etc…

Définir le filtrage



Sélectionnez *all* et cliquez sur OK.



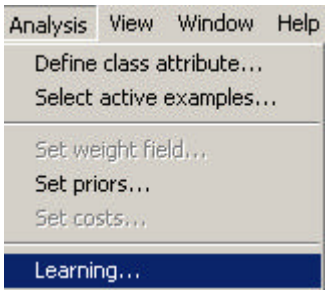
Par défaut, l'analyse s'effectue sur la totalité du fichier de données.

Apprentissage

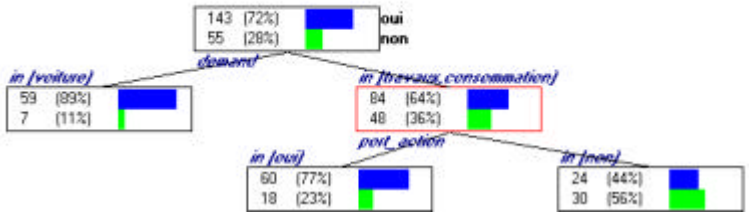
Nous en avons terminé avec la phase de paramétrage. Reste l'essentiel : l'apprentissage de ces données et les conclusions qui l'on peut en tirer.

Lancement de l'apprentissage

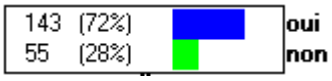
Pour démarrer l'apprentissage rien de plus simple. Cliquez simplement sur



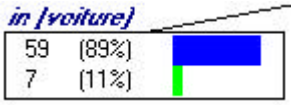
L'arbre de décision apparaît……



Interprétation des résultats

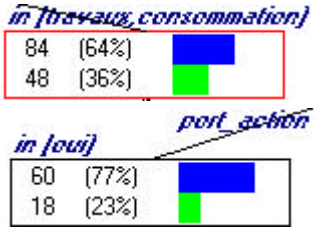


Parmi les 198 enregistrements de notre fichier de données, 72% sont des clients à qui le crédit a été accordé, 28% refusé.



De plus parmi les clients ayant fait la demande d'un crédit voiture, 89% ont été accepté contre seulement 11% refusé.

Nous pourrions donc conclure tout de suite qu'un bon client, et donc un client solvable, est un client qui a fait la demande d'un crédit voiture. Cependant nous pouvons aller plus loin dans notre analyse comme le montre les autres nœuds de notre arbre.



En effet, parmi les clients ayant demandé un crédit pour travaux ou consommation, 64% ont été validé, 36% refusé

De plus, parmi ces mêmes personnes, ce sont celles qui détenaient des actions (port_action) pour lesquels le plus de crédit a été accordé.

Conclusion

Si le client demande un crédit auto, il y a de grande probabilité que ce dernier soit accepté sans trop de difficulté. Pour d'autres demandes, ce sont les clients détenteurs de porte-feuille d'actions et demandeurs de crédits pour travaux ou consommation qui seront privilégiés.

Résumé

Action	Comment ?
Ouvrir le fichier de données	File / Open / Bqe.fdm
Choisir la méthode d'analyse	Induction Method / Standard Algorithm / ...
Définir la variable à prédire et les variables prédictives	Analysis / Define Class Attribut
Définir l'échantillon de données	Analysis / Select Active Examples
Apprendre	Analysis / Learning