

Exploration et modélisation de relations au sein de gros volumes de données : le logiciel entreprise minertm

Eric Menguy

Ariane Sioufi

Sas Academic

The Power to Know[™]

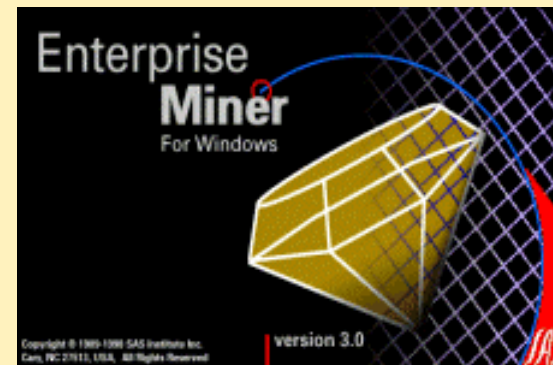
Plan

I. Présentation générale de Enterprise Miner

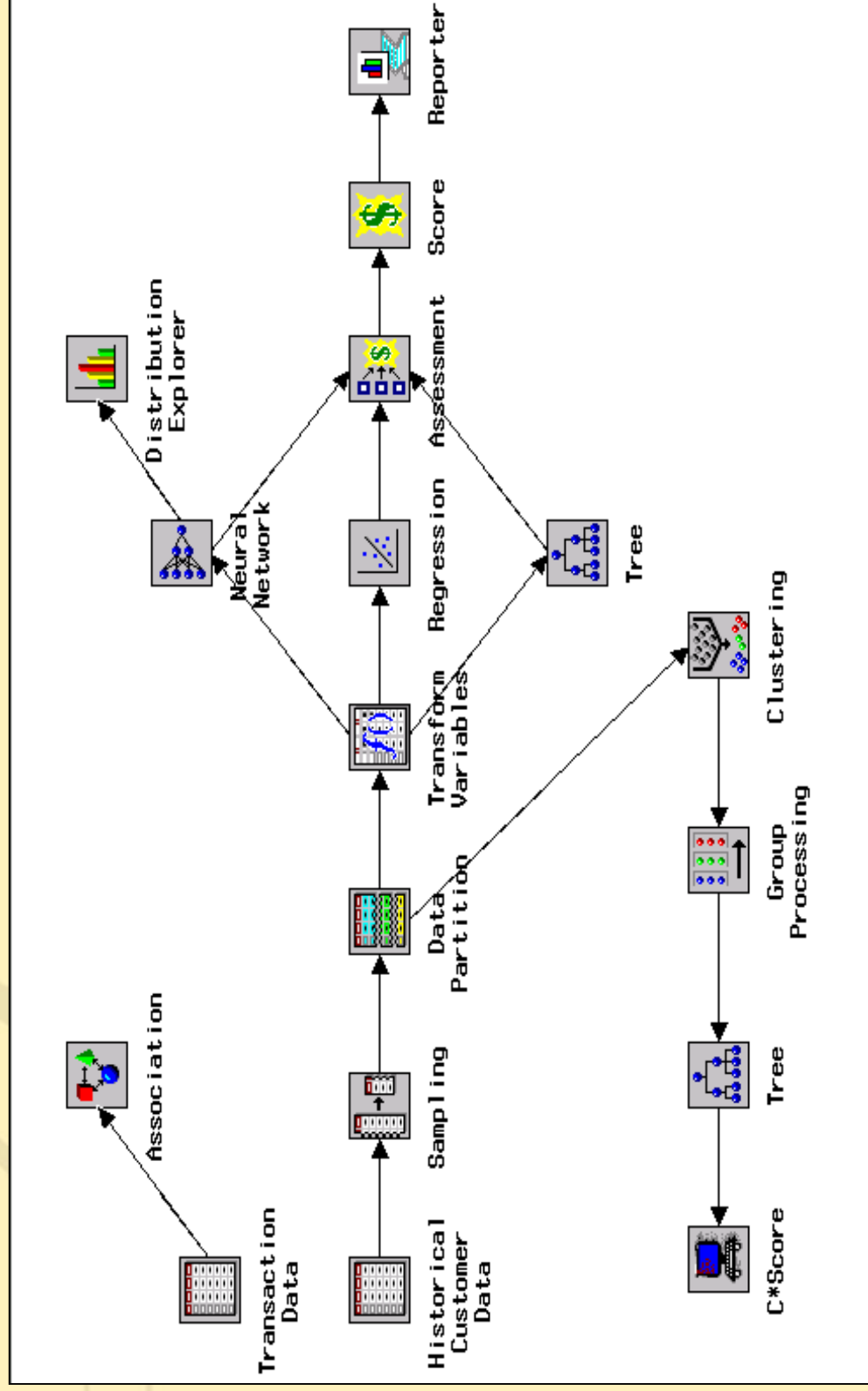
II. Démonstration

Présentation générale de Enterprise Miner

Présentation générale de Enterprise Miner



Enterprise Miner™ est une solution logicielle intégrée de SAS pour le Data Mining, qui, au travers d'une interface totalement graphique, donne à l'utilisateur l'accès à de nombreux outils analytiques.



Présentation générale de Enterprise Miner

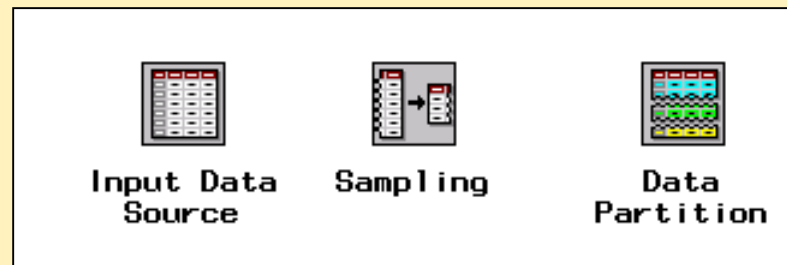
- Méthodologie Semma -

Décomposition de l'activité en Cinq étapes :

- *Sample*
- *Explore*
- *Modify*
- *Model*
- *Assess*

Présentation générale de Enterprise Miner

- Le groupe Sampling -

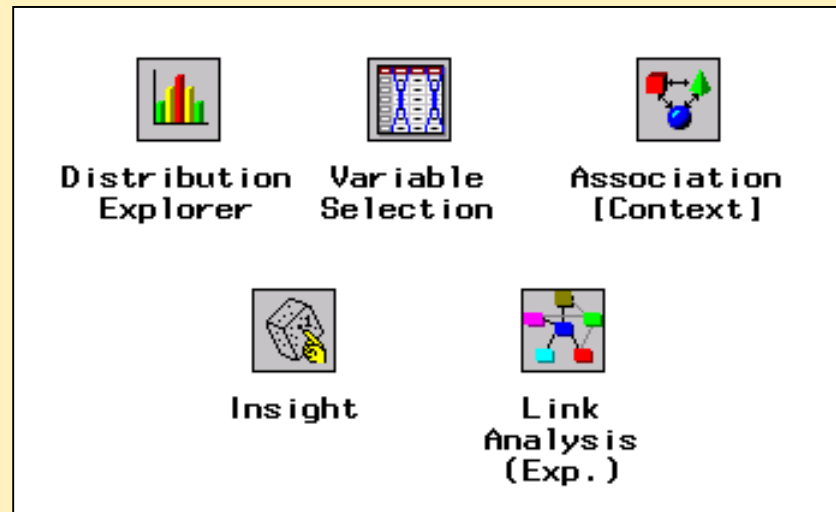


Cette étape permet de :

- l'identification des données (données initiales)
- l'échantillonnage (ou un sur-échantillonnage) des données initiales
- la définition de jeux de données d'apprentissage, de validation et de test en prévision de l'évaluation de la qualité de modèles de prédiction

Présentation générale de Enterprise Miner

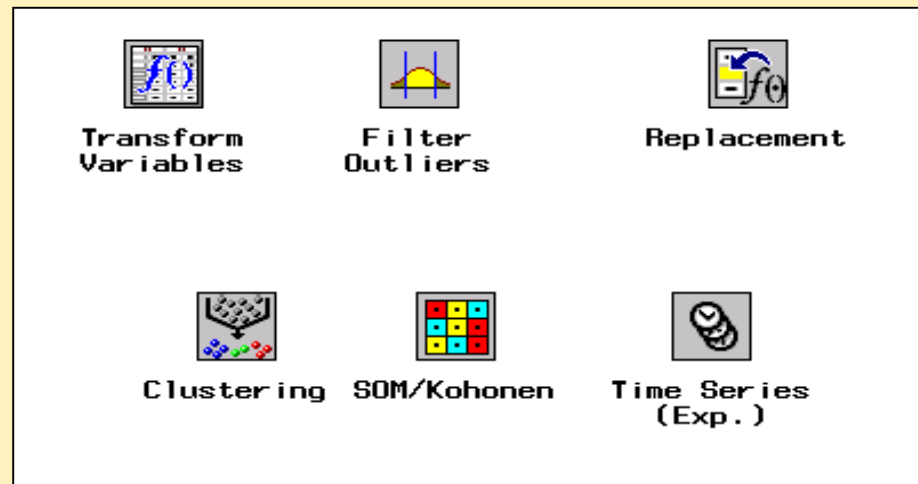
- Le groupe Explore -



Proposition d'outils d'exploration graphique, le calcul de statistiques descriptives, la mesure d'associations ou de liens entre items.

Présentation générale de Enterprise Miner

- Le groupe Modify -

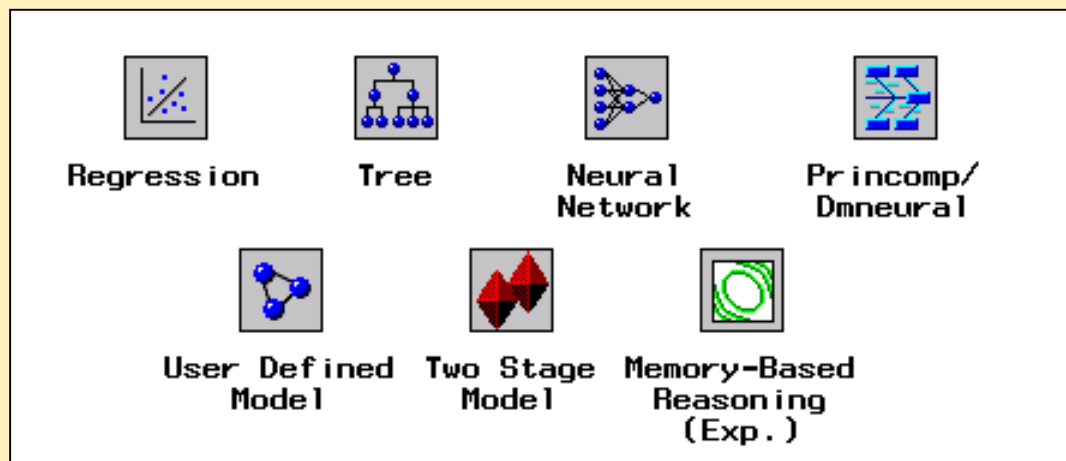


Cette étape permet de :

- modifier les données : élimination des données aberrantes ou atypiques, transformation des variables, remplacement automatique des données manquantes
- classer les données : classification mixte, réseaux de Kohonen.

Présentation générale de Enterprise Miner

- Le groupe Model -



Présentation générale de Enterprise Miner

- Le groupe Model -

C'est la phase prédictive où différents modèles peuvent être mis en compétition.
Les modèles suivants sont proposés :

- « *Neural Networks* »
- « *Decision Trees* »
- « *Regression Analysis* » : régression linéaire, régression logistique
- « *Memory Based Reasoning* » : algorithme des k-voisins les plus proches
- **Combinaison de modèles**
- « *Two-stage models* » : combinaison d'une cible discrète et d'une cible continue
- **Tout modèle disponible dans le système SAS (nœud « User Defined Model »)**

Présentation générale de Enterprise Miner

- Le groupe Model –

- Architectures de réseaux de neurones disponibles avec Enterprise Miner™

- **Modèles linéaires généralisés (GLIM)**
- **Perceptron multicouche (MLP)**
- **Fonction à base radiale ordinaire (ORBF)**
(possibilité de fixer la hauteur ou la largeur)
- **Fonction à base radiale normalisée (NRBF)**
(possibilité de fixer la hauteur, la « largeur » ou le volume)

En plus de ces architectures proposées, l'utilisateur dispose d'une interface graphique lui permettant la construction manuelle du réseau.

Présentation générale de Enterprise Miner

- Le groupe Model -

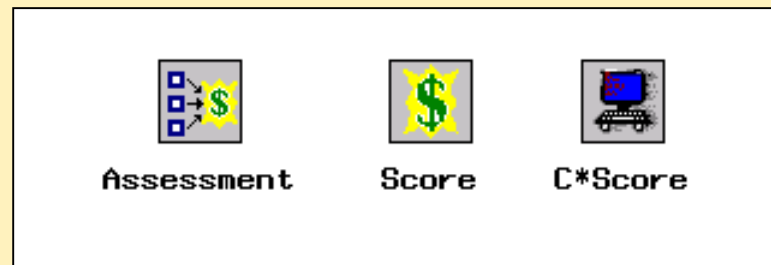
- Particularités du nœud Tree -

- critères de division :
 - Variable cible discrète : Chi2 de Pearson, indice d'Entropie, indice de Gini
 - Variable cible continue : Test de Fisher, réduction de la variance
- possibilité d'implémenter les méthodes CHAID, CART, C4.5
- possibilité de définir des variables suppléantes
- élagage automatique et sélection du sous-arbre optimal
- calcul de la statistique d'importance pour chaque variable explicative

Le nœud *Tree* propose également une interface permettant la construction manuelle du modèle.

Présentation générale de Enterprise Miner

- Le groupe Assess -



Phase de comparaison des résultats produits par les différentes méthodes à l'aide d'interfaces graphiques.

La comparaison peut se faire sur des critères statistiques, mais également financiers.

Enterprise Miner™ permet également d'exporter les règles de « Scoring » dans un format « simple » afin de permettre l'application des règles dans toutes les bases de données (récupération du code de « Scoring » en langage SAS ou en langage C)

Démonstration

Démonstration

- Le conteste -

Les données proviennent d'une société nucléaire fictive qui fabrique des pastilles de Plutonium.

Ces pastilles en fin de processus sont empilées dans des tubes (les tubes sont ensuite placés dans les réacteurs).

Les tubes subissent un contrôle final ; on analyse les pastilles une à une. Si une pastille a une concentration en Plutonium trop importante (par exemple), le tube complet est rejeté (ce qui engendre beaucoup de perte).

Démonstration

- Le conteste -

Objectif : cibler les pastilles de mauvaise qualité en insistant sur :

- L'identification des paramètres de production produisant des pastilles de mauvaises qualités.

→ Analyse du système de production

- L'identification des pastilles de mauvaise qualité avant le passage dans les tubes

→ Amélioration du rendement

Démonstration

- Description du processus -

En début de processus, les matières premières se trouvent sous forme de poudre (poudre d'Uranium, Plutonium, d' UO_2 ...).

Cette poudre est broyée, mélangée puis compressée.

Après la compression, la poudre est transformée en plusieurs pastilles. Ces pastilles sont ensuite chauffées puis éventuellement rectifiées (ajustement du diamètre par exemple).

Démonstration

- Description des paramètres -

Au cours du processus, un certain nombres de paramètres sont mesurés. Les paramètres inclus dans le fichier de données correspondent aux phases suivantes :

Poudre -> broyage -> compresse -> chauffage.

Démonstration

- Description des paramètres -

Matières premières :	Poids des lots Coulabilité Qualité Matière première Pourcentage d'AZB (matière additive) Teneur en Pu % de matière recyclée
Broyage :	Ouverture des clapets Vitesse
Mélange :	Grains de Pu Nature du tamisage Taille des tamis
Pressage :	Cadence de compression Numéro des presses Hauteur de remplissage
Frittage :	Numéro des fours Température Diamètre Densité

Démonstration

- Description des paramètres -

Column Name	Type	Length	Format	Informat	Label
 CLAPETS	Text	12			OUVERTURE DES CLAPETS DU BROEUR
 GRAIN	Text	10			PRESENCE OU NON DE GRAINS DE Pu
 POIDS	Number	8	4.		POIDS DES LOTS
 COULAB	Number	8	4.2		COULABILITE DE LA POUDRE
 MP	Number	8			QUALITE DE LA MATIERE PREMIERE
 CADENCE	Number	8			CADENCE DE COMPRESSION
 TAMISAGE	Text	8			NATURE DU TAMISAGE
 PRESSE	Text	3			NUMERO DE LA PRESSE
 BROEUR	Text	3			NUMERO DU BROEUR
 FOUR	Text	3			NUMERO DU FOUR
 TAMIS	Text	4			TAILLE DU TAMIS
 AZB	Text	4			POURCENTAGE D'AZB
 TENEUR	Text	8			TENEUR EN Pu DES PASTILLES
 RECYCL	Text	4			POURCENTAGE DE MATIERE RECYCLEE
 TEMP	Text	4			TEMPERATURE DU FOUR
 REMPLI	Text	4			HAUTEUR DE REMPLISSAGE DES PRESSES
 QUALITE	Text	8			QUALITE DES PASTILLES
 DENSITE	Number	8			DENSITE DES PASTILLES
 DIAMETRE	Number	8			DIAMETRE DES PASTILLES
 VITESSE	Text	4			VITESSE DE BROYAGE

Démonstration

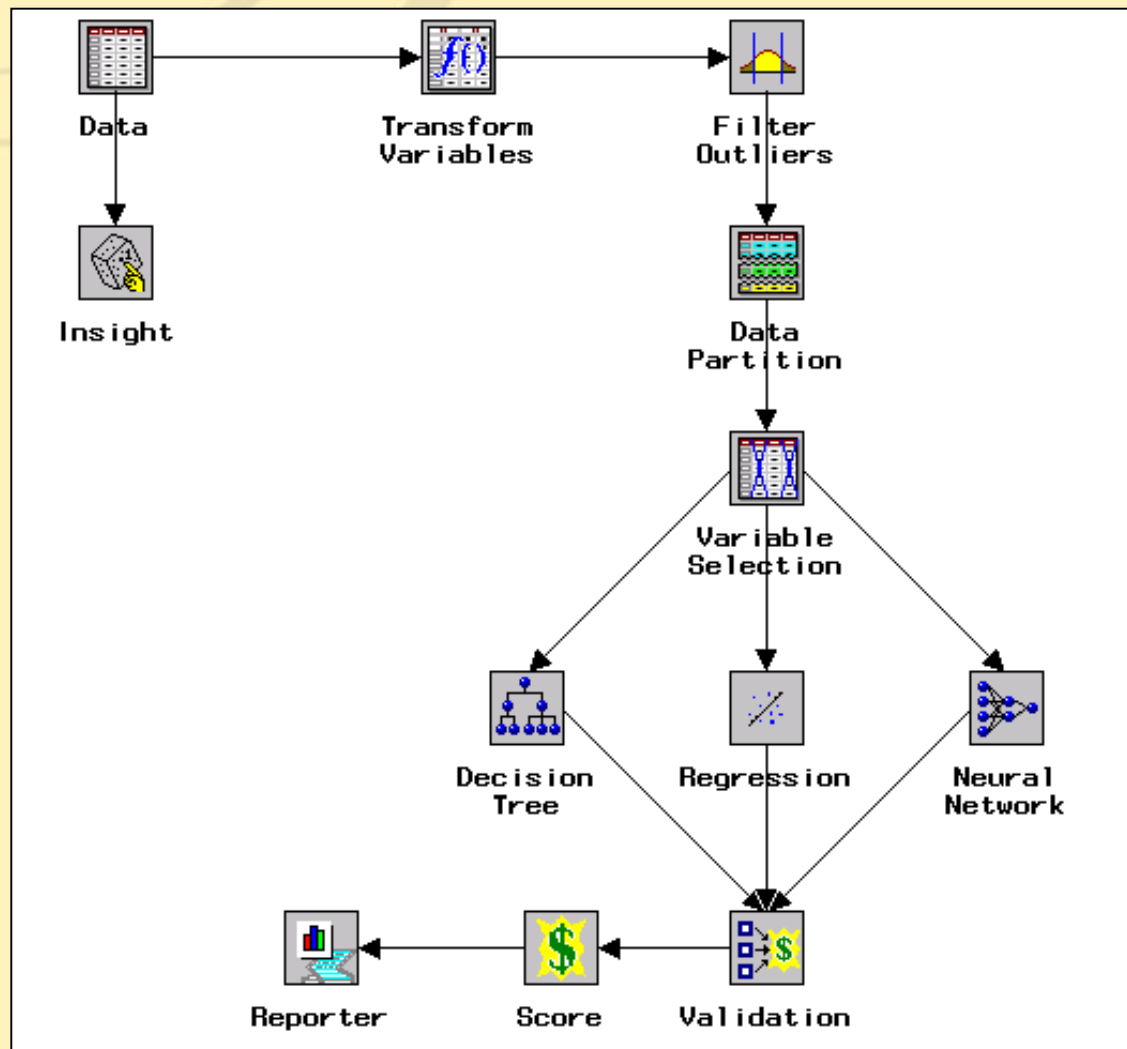
- Matrice de Coûts -

Prédire qu'une pastille de mauvaise qualité sera bonne,
nous fait perdre 15 000 euros
(cette pastille sera mise dans le tube et le contaminera).

Prédire qu'une pastille de bonne qualité sera mauvaise,
nous fait perdre 300 euros
(On perdra cette pastille, elle passera au rebut et subira un
certain nombre de traitements ce qui entraîne des coûts).

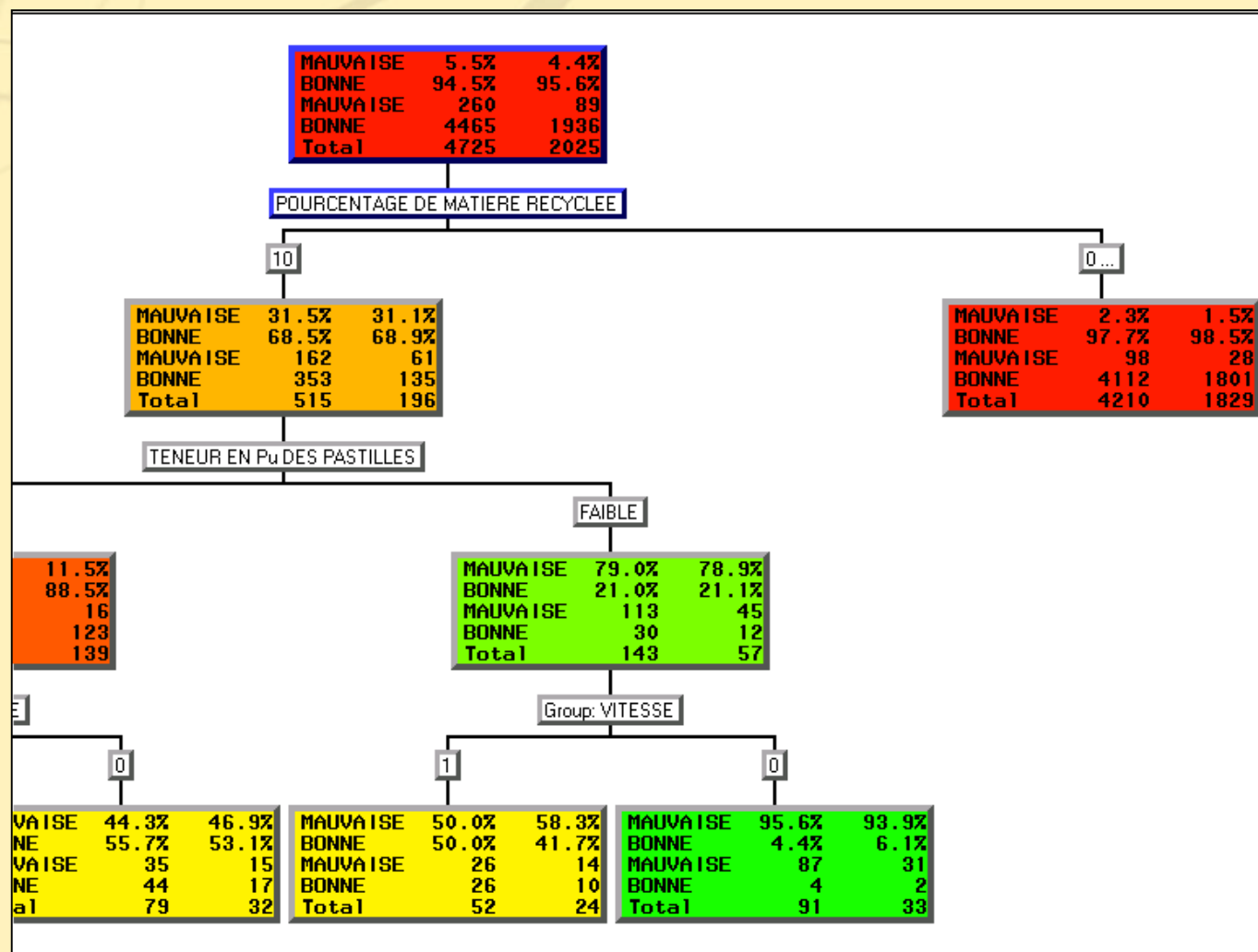
Démonstration

- Diagramme -



Démonstration

- Résultats -



Démonstration

- Résultats -

