Les techniques d'exploitation de données (Data Mining)

Présenté par :

Emer Mestiri, M.sc Finance, Data Scientist

Conseiller Gestion de risque de crédit, Mouvement Desjardins

Sommaire

I. Logiciel SAS Miner

I.1 Présentation

- ✓ Repose sur le socle commun de l'ensemble des solutions décisionnelles de SAS;
- ✓ Conçu pour fonctionner avec les technologies SAS d'intégration et d'analyse de données et de reporting;
- ✓ Industrialise le processus de data mining pour but de définir les modèles de prédiction et de segmentations;

I.2 Forces

- ✓ Modèles prédéfinis et auto-documenté ;
- ✓ Utilisation simple et flexible ;
- ✓ Interface intuitive et déploiement facile des processus ;
- ✓ Outils de validation et d'évaluation des modèles ;
- ✓ Une Plus-Value décisionnelle plus importante ;

I.2 Inconvénients

- ✓ Des serveurs de grande capacité;
- ✓ Une Licence plus dispendieuse;
- ✓ Peu ou pas de programmation ;

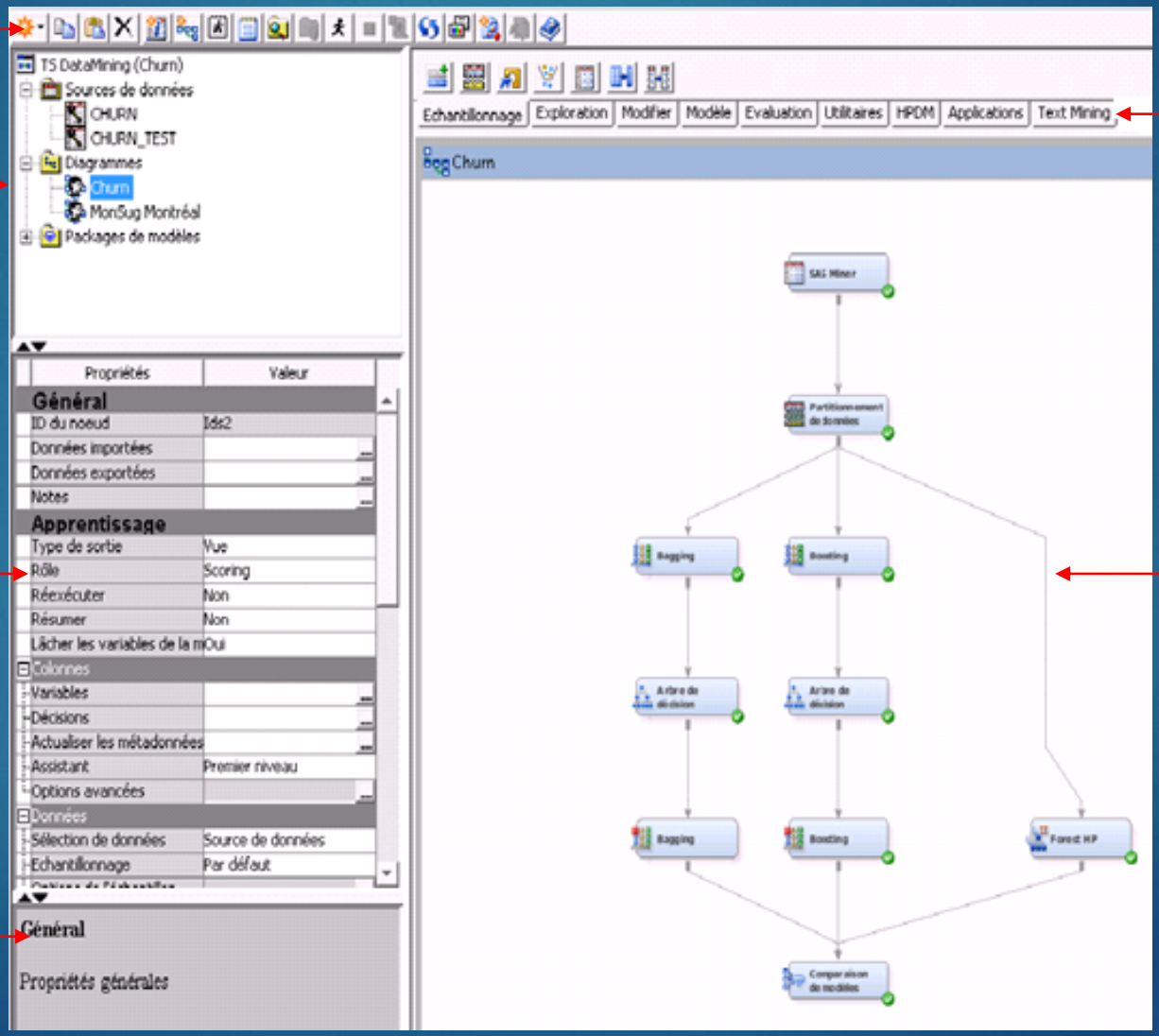
I.3 Interface de travail

La barre des raccourcis

La fenêtre des projets

Propriétés des projets

Propriétés générales et aide pour le projet



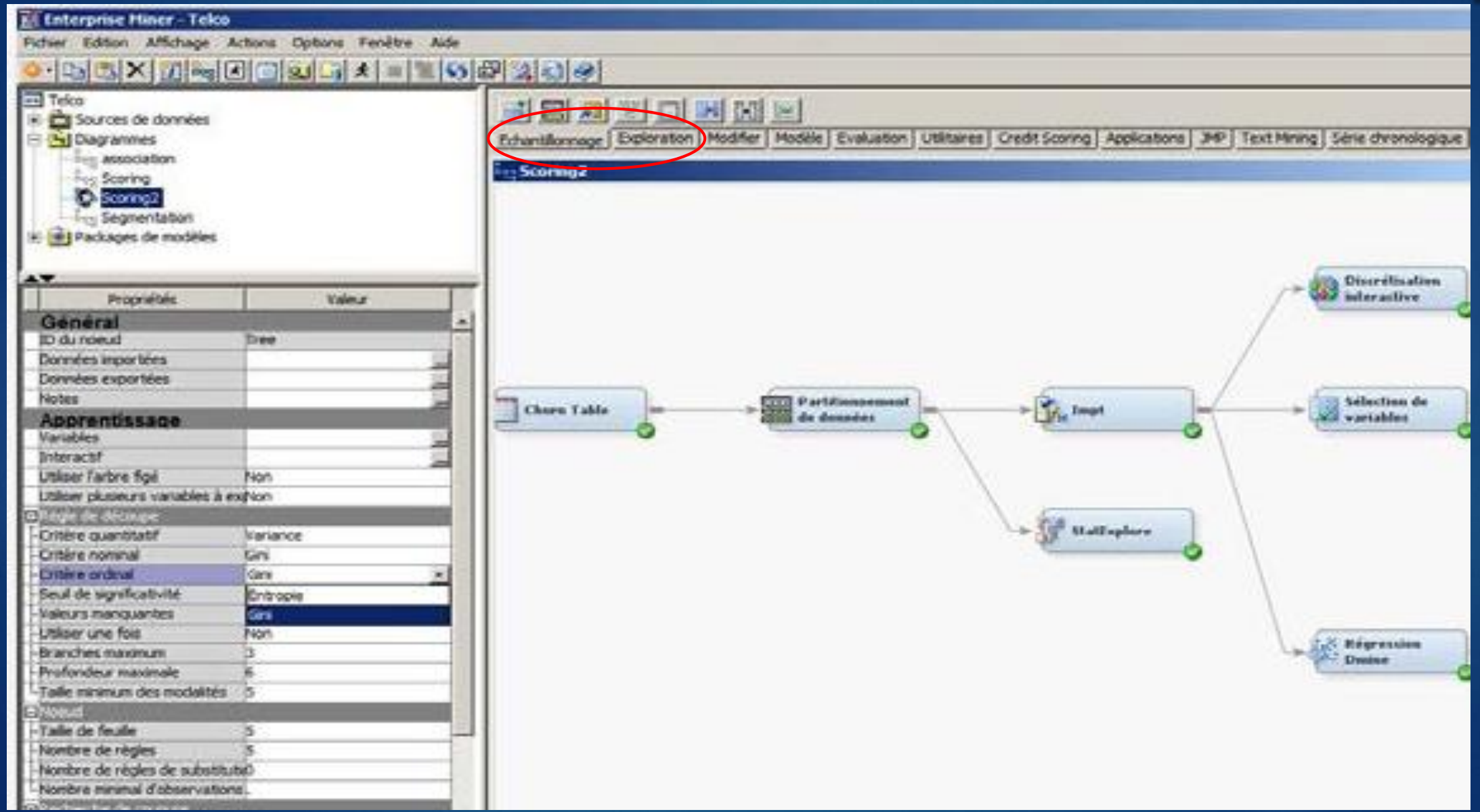
S.E.M.M.A (Sample, Explore, Modify, Model, Assess)

Diagramme ou espace de travail

I.4 Nœuds et tâches

- ✓ Échantillonnage :
  1. Identifier et définir l'ensemble des variables;
  2. Créer de façon aléatoire des échantillons;
  3. Partitionner les données sous forme de données d'apprentissage, de validation et de test;
- ✓ Exploration :
  1. Définir les relations entre des variables;
  2. Identifier les valeurs extrêmes et les valeurs aberrantes;
  3. Bâtir des règles d'association, de segmentation « Clustering » et/ou de l'analyse factorielle;
  4. Sélectionner les variables significatives à l'aide de critères comme R2 ,Khi-2, AIC, BIC etc.;

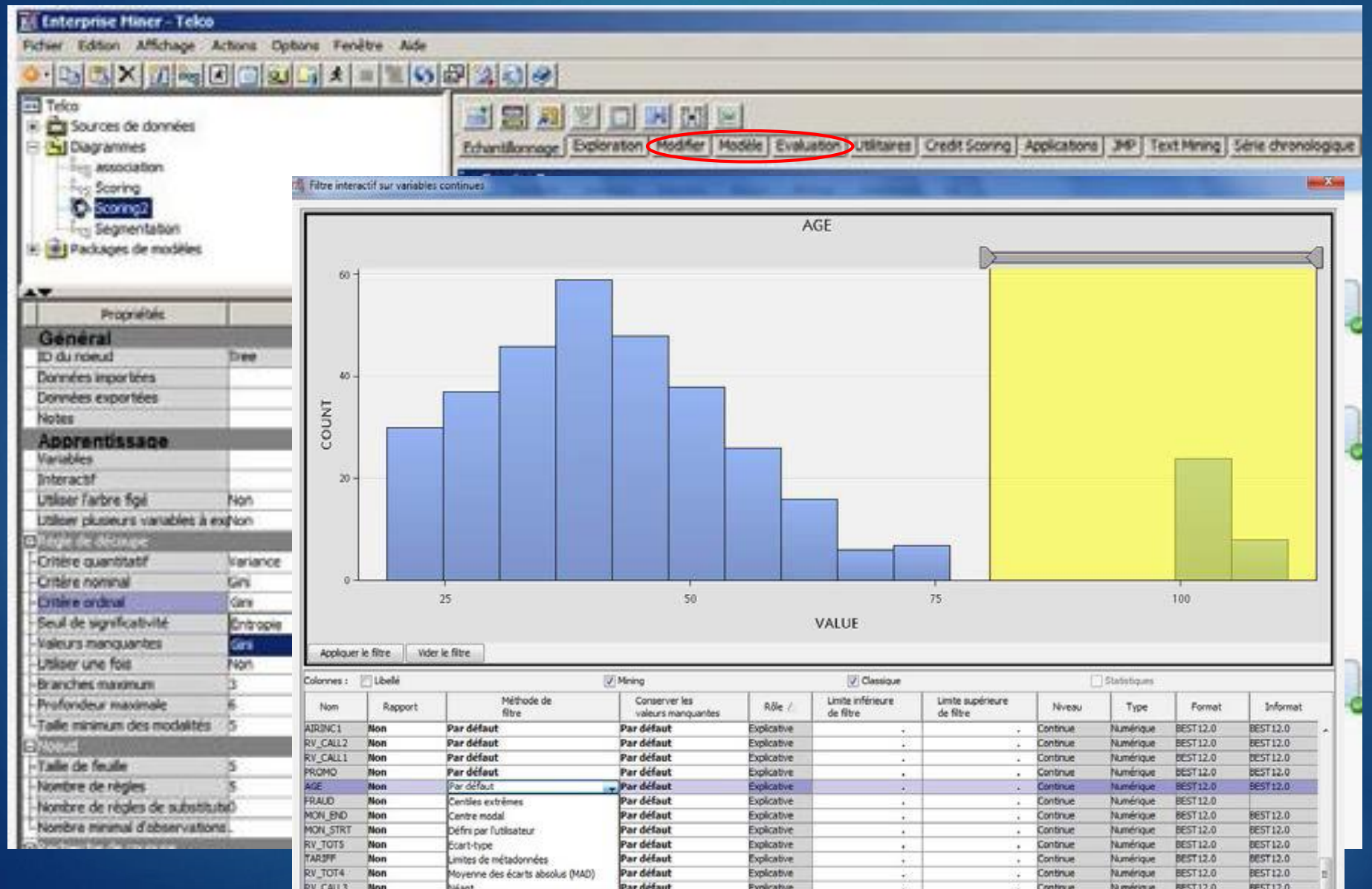
I.4 Nœuds et tâches



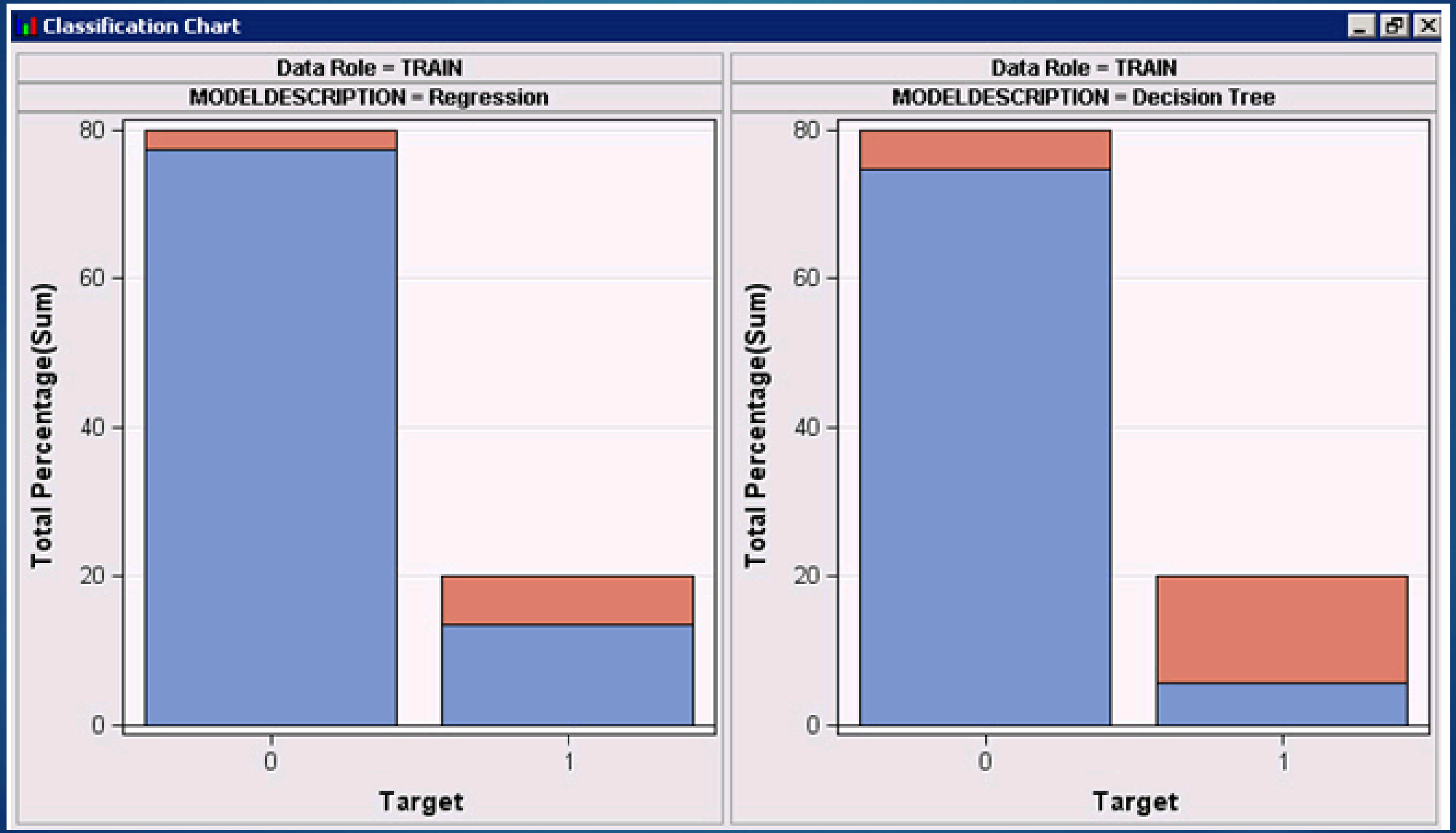
I.4 Nœuds et tâches (suite)

- Modification :
  1. Transformer les variables existantes ou créer des variables supplémentaires;
  2. Remplacer les valeurs manquantes et filtrer les données ;
- Modèle :
  1. Décider des modèles statistiques ou des modèles de prévision (moindres carrés, régression linéaire/logistique, arbre de décision, réseaux de neurones ou la méthode d'ensemble);
  2. Ajuster les modèles par l'utilisateur dans la fenêtre des propriétés des projets;
- Évaluation :
  1. Comparer la performance des différents modèles en affichant les estimations des erreurs quadratiques, le taux de mauvaises classifications ou des graphiques/tableaux (exemple: la courbe ROC).
  2. Choisir le groupe d'échantillon des données (Validation ou test).

I.4 Nœuds et tâches (Modifier)



I.4 Nœuds et tâches (Évaluation)



II. Techniques d'exploitation de données

II.1 Méthodes des ensembles

- Définition:

La méthode d'ensembles est un ensemble de modèles dont les prédictions sont combinées d'une certaine manière (typiquement une moyenne pondérée ou non des valeurs prédites) afin de prédire de nouvelles données.

II.1 Méthodes des ensembles

- ✓ Objectifs des méthodes d'ensemble :

1. Précision :

Une moyenne de plusieurs mesures répétées est plus précise qu'une mesure individuelle.

Une moyenne de valeurs prédites par différents modèles est souvent plus précise que la valeur prédite à l'aide d'un seul modèle.

1. Stabilité :

Une moyenne est beaucoup plus stable qu'une seule mesure.

II.1 Méthodes des ensembles

- ✓ Variables explicatives

1. Variables continues:

   La moyenne des valeurs prédites par chaque modèle.

2. Variables nominales, ordinales ou intervalles :

   La moyenne de la probabilité a posteriori de chaque modèle pour chaque classe de la variable cible (La classe la plus populaire).

II.1 Méthodes des ensembles

- ✓ Présentation du sujet :

a) Une compagnie de télécommunication ;

b) Une base de données qui contient des informations sur le comportement de ses clients (nbre de minutes d'appels, nbre de plaintes, les forfaits etc.) ;

c) Prédire le taux de rétention de ses clients ;

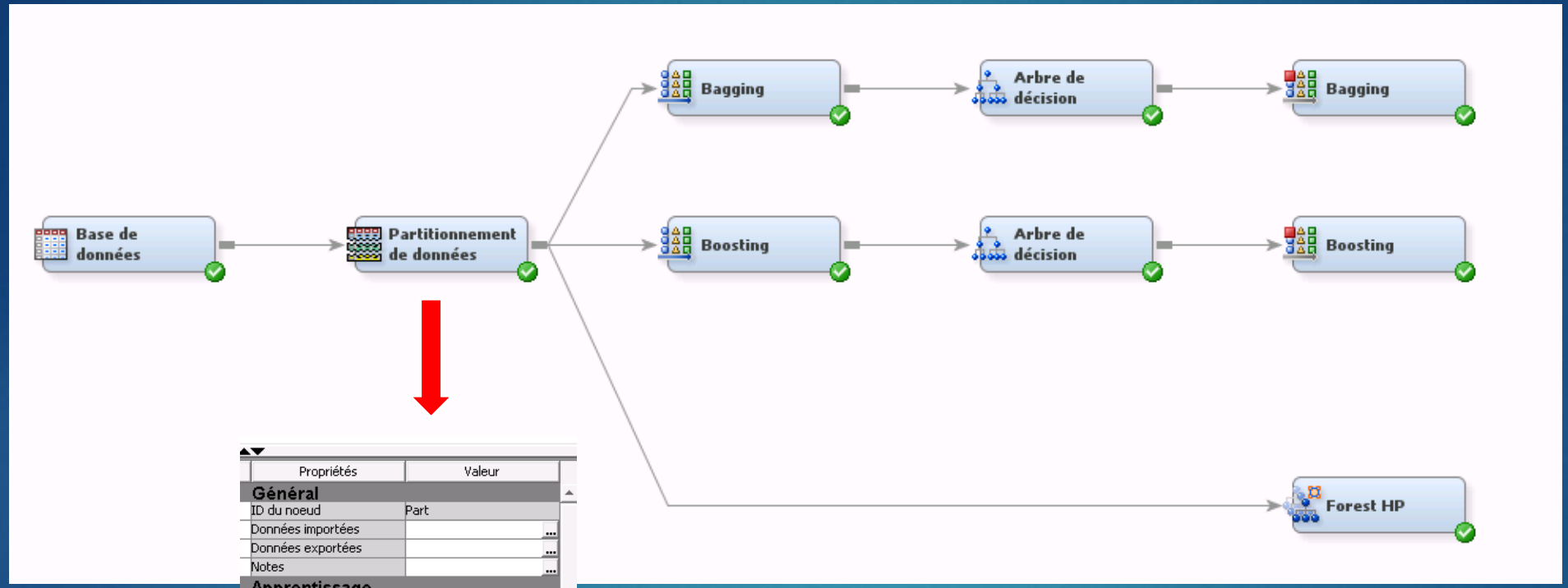
- ✓ Les modèles utilisés :

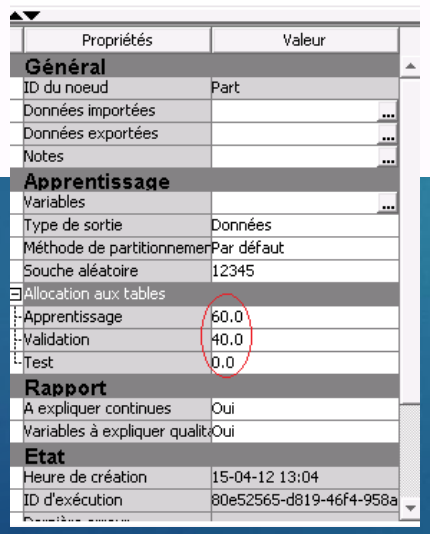
a) Le Bagging ;

b) Le Boosting ;

c) Les forêts aléatoires ;

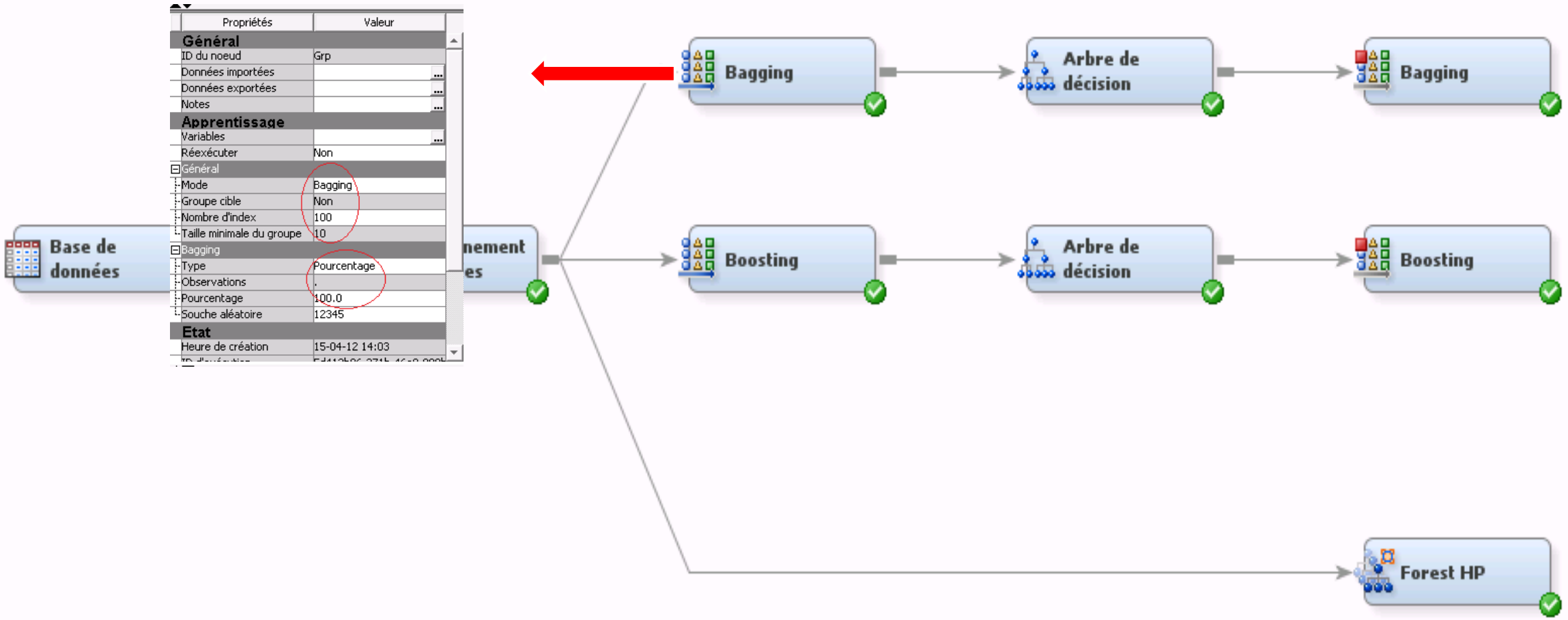
II.1 Méthodes des ensembles



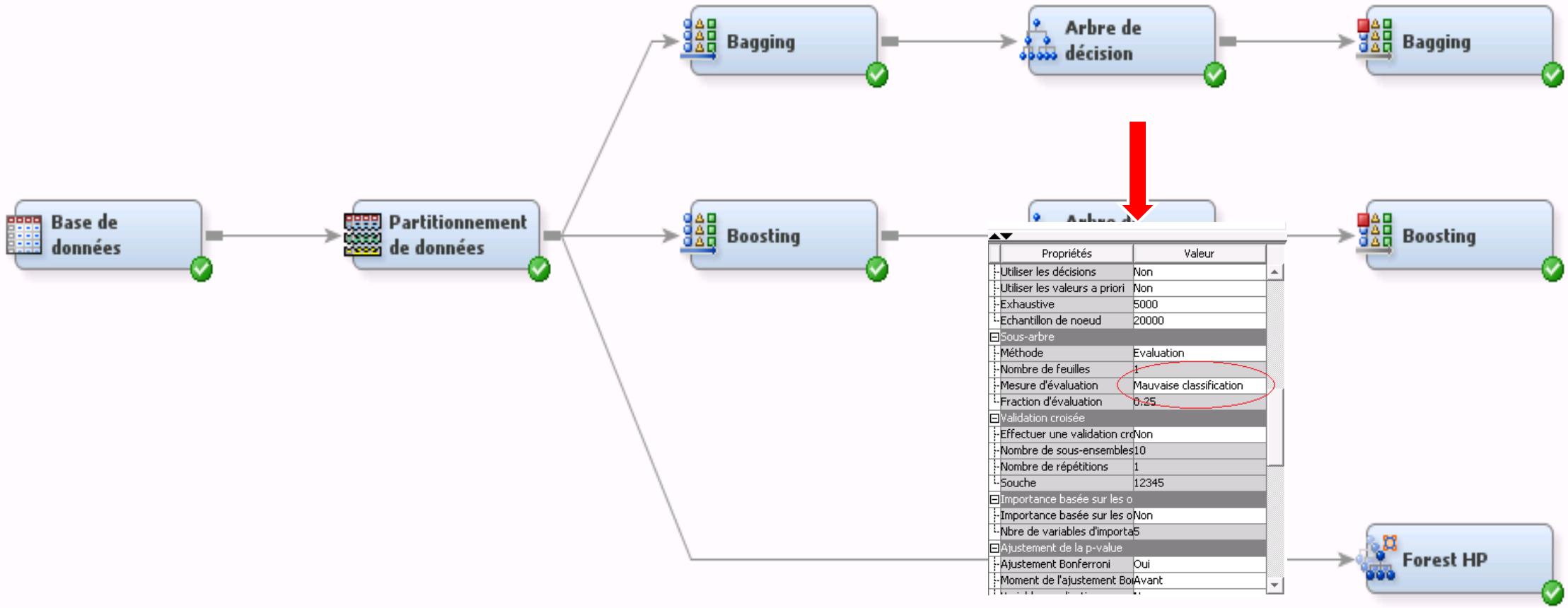


Propriétés	Valeur
Général	
ID du noeud	Part
Données importées	
Données exportées	
Notes	
Apprentissage	
Variables	
Type de sortie	Données
Méthode de partitionnemer	Par défaut
Souche aléatoire	12345
Allocation aux tables	
Apprentissage	60.0
Validation	40.0
Test	0.0
Rapport	
A expliquer continues	Oui
Variables à expliquer qualit	Oui
Etat	
Heure de création	15-04-12 13:04
ID d'exécution	80e52565-d819-46f4-958a

II.1 Méthodes des ensembles

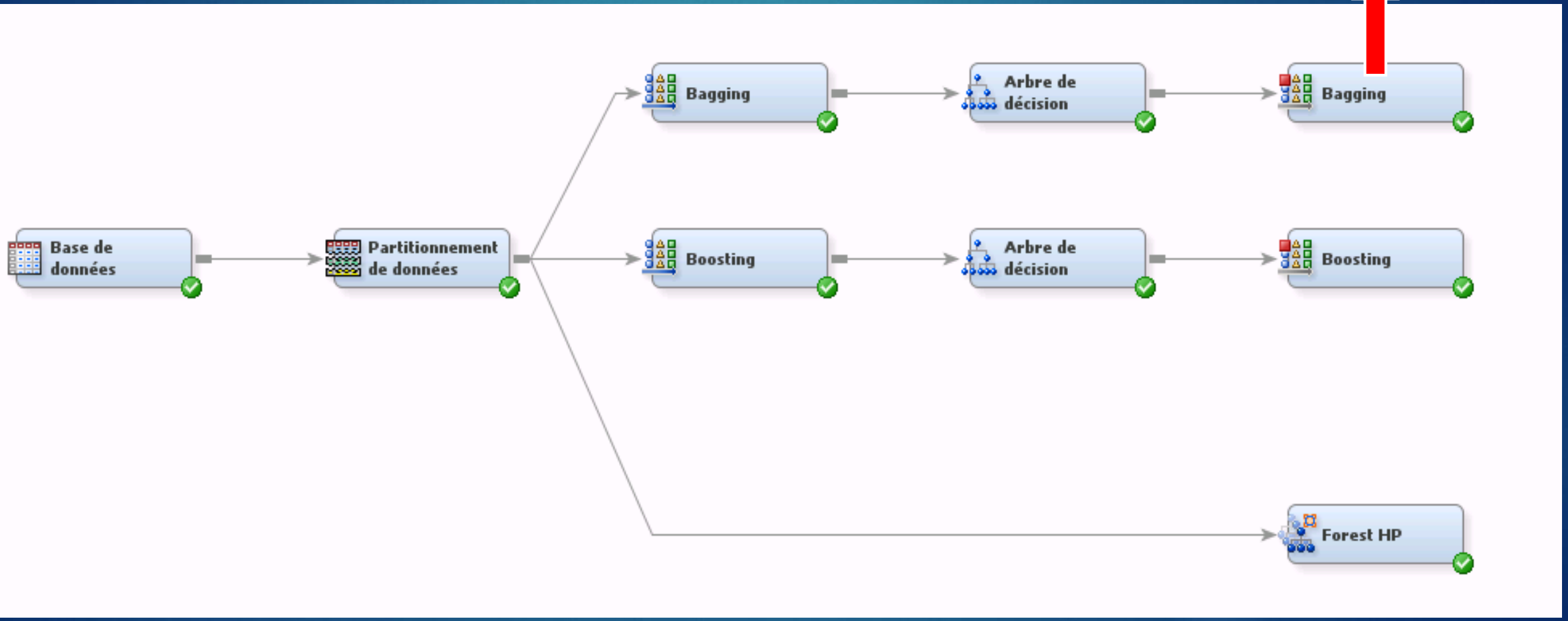


II.1 Méthodes des ensembles

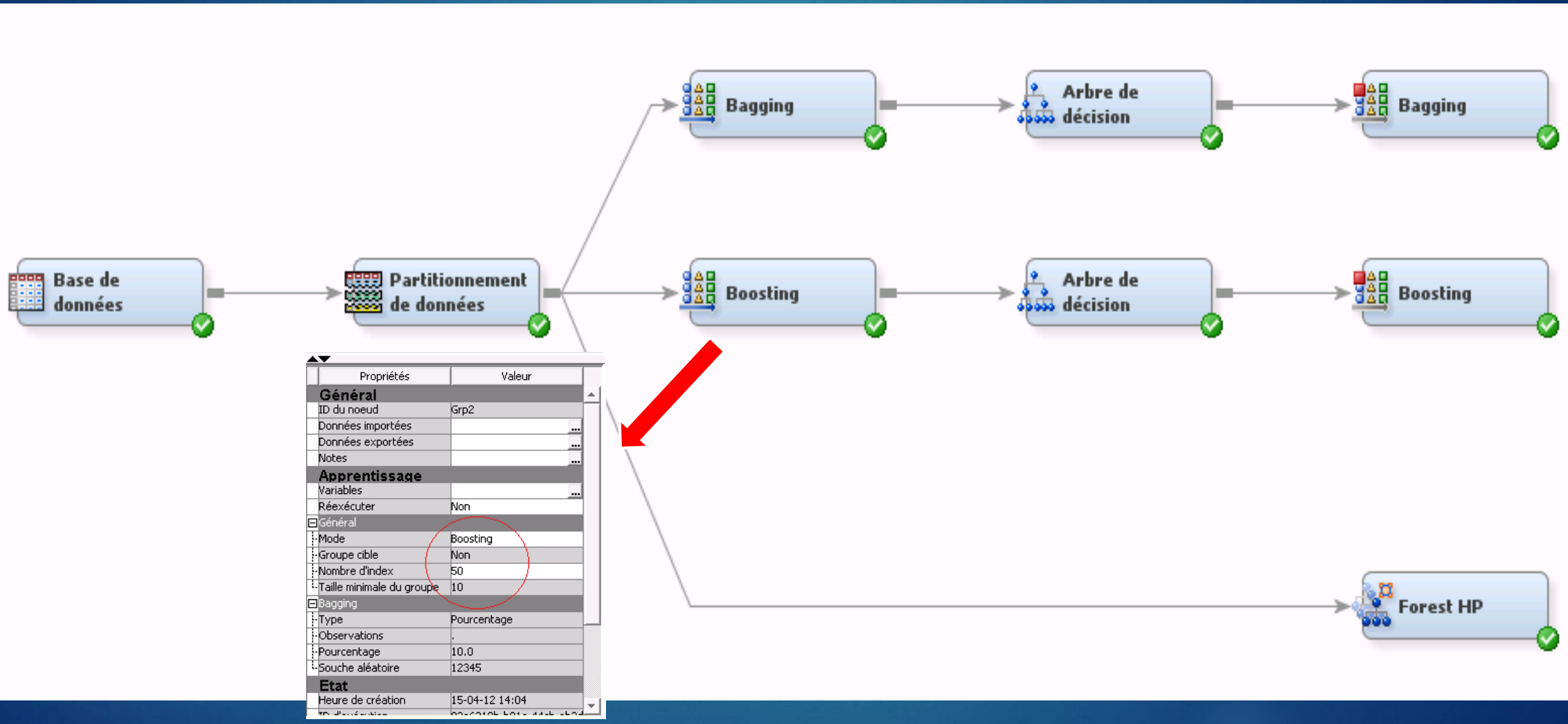


II.1 Méthodes des ensembles

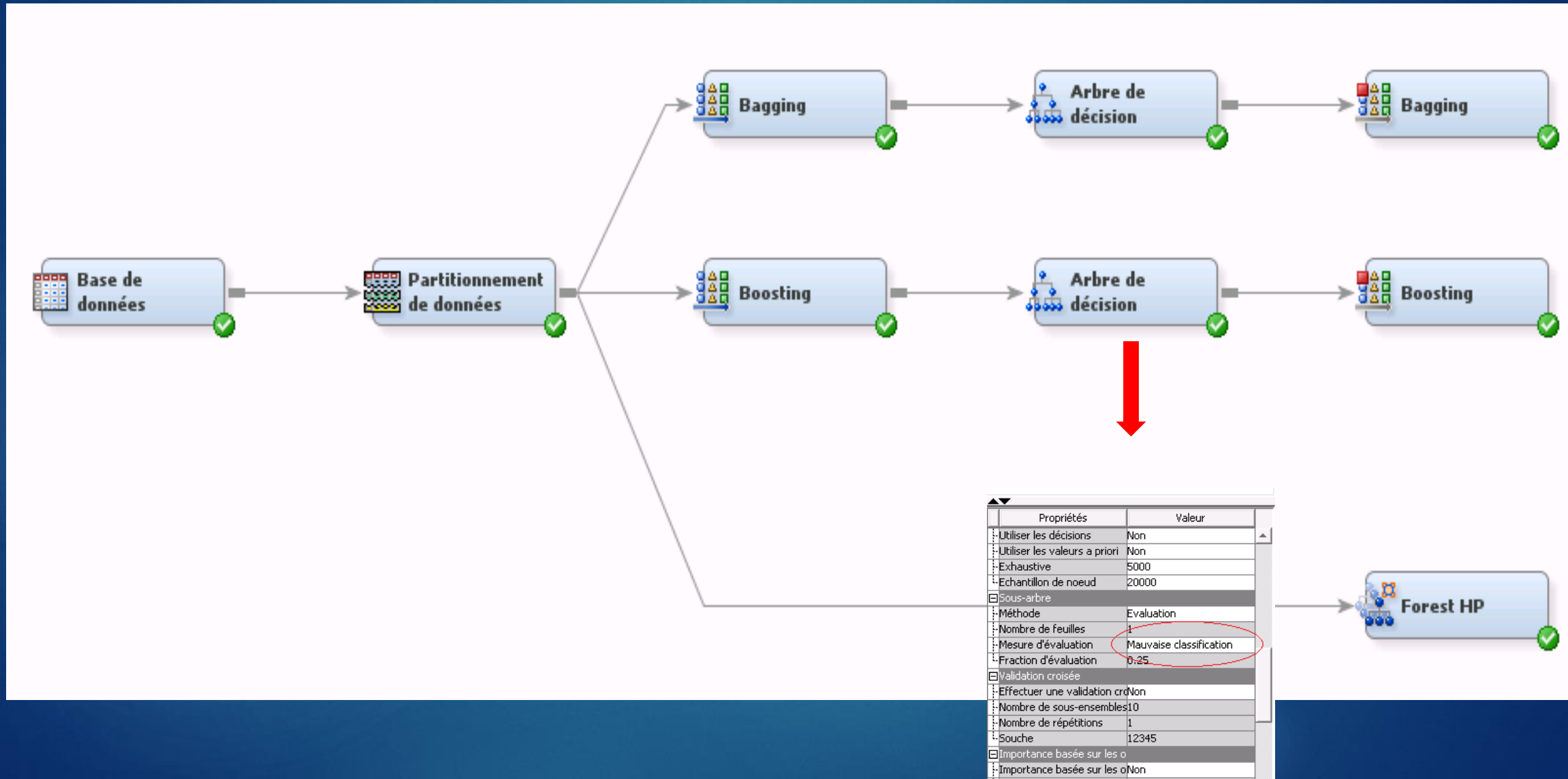
Exécuter la tache «Fin de Traitement en Groupe» pour fermer la boucle du Bagging



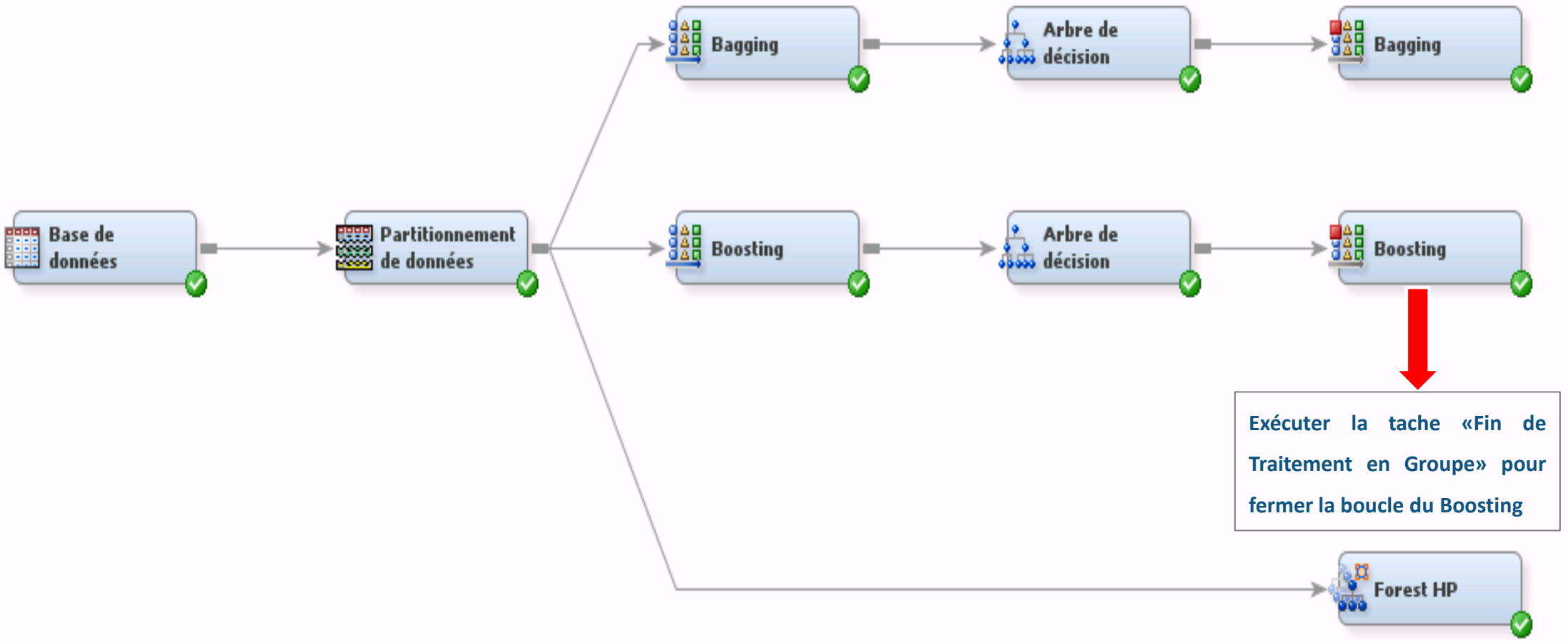
II.1 Méthodes des ensembles



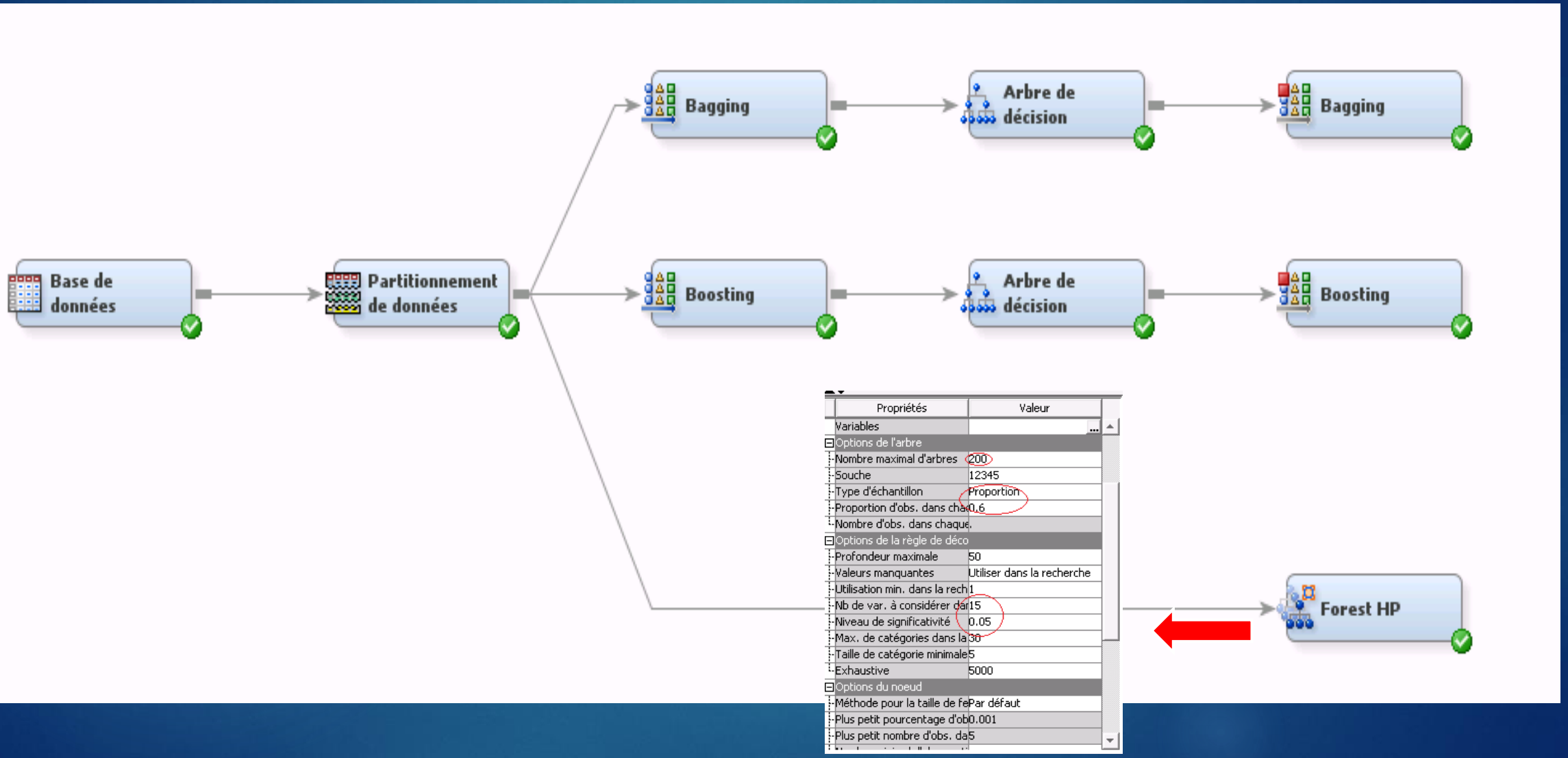
II.1 Méthodes des ensembles



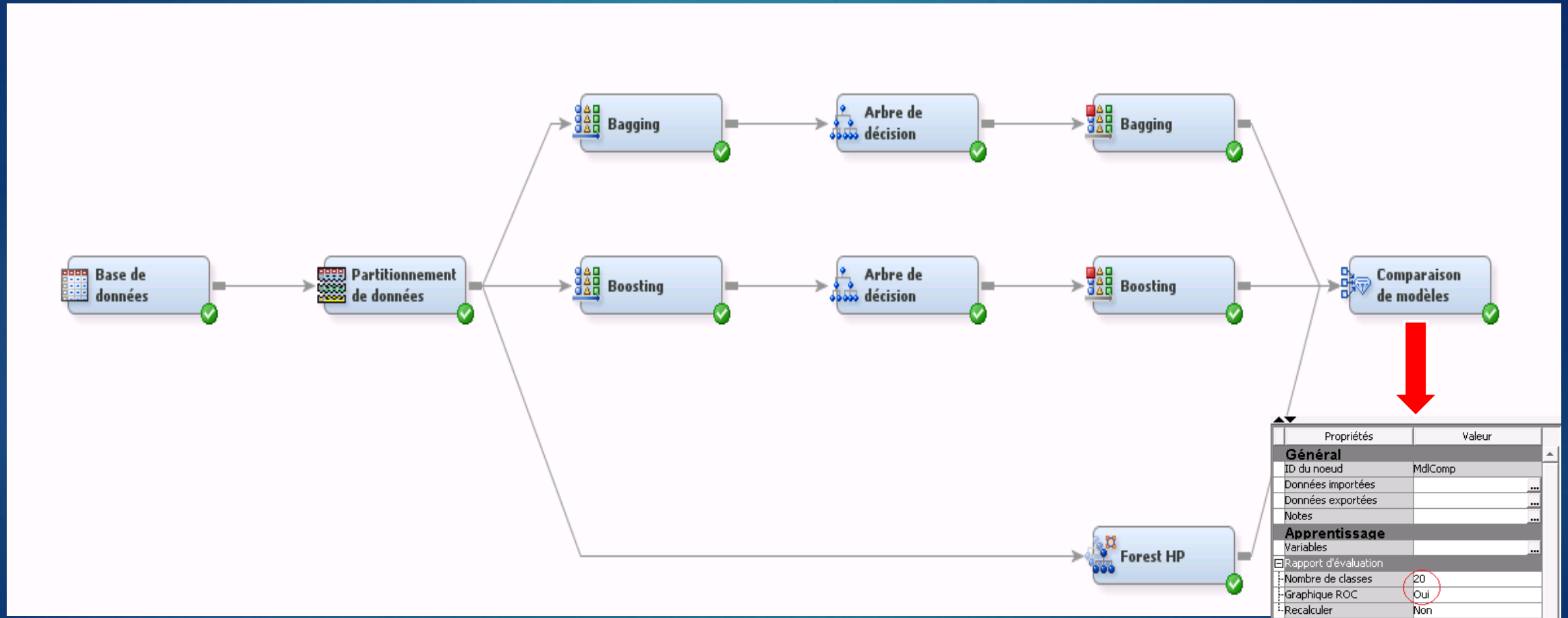
II.1 Méthodes des ensembles



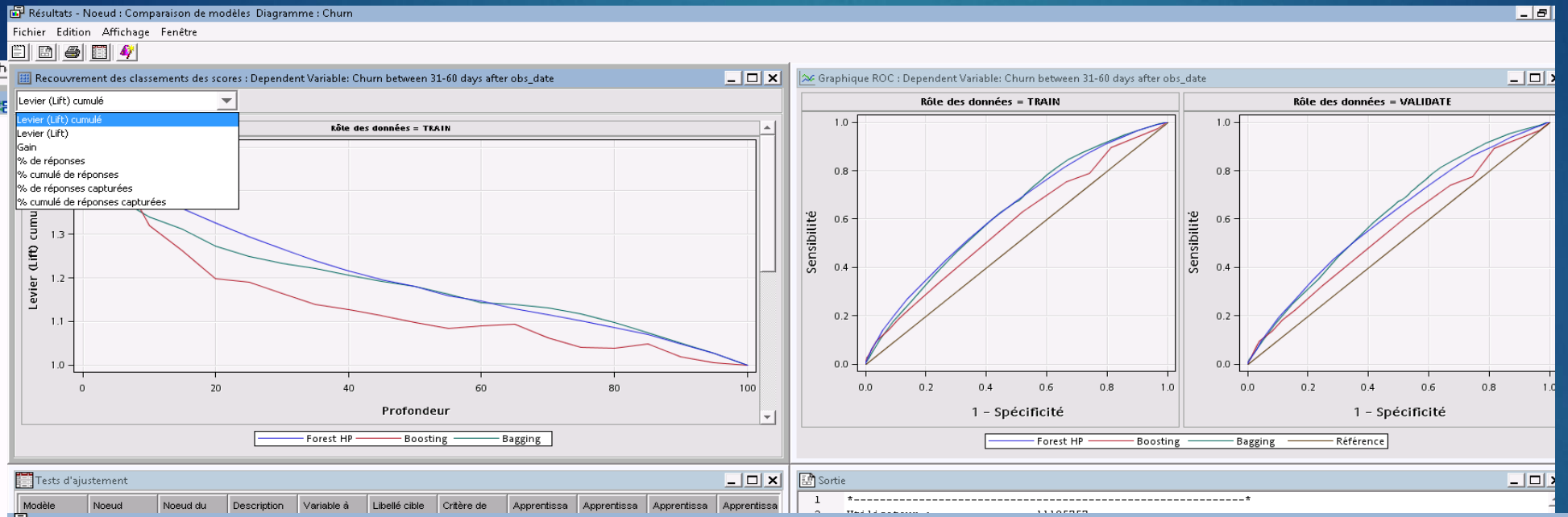
II.1 Méthodes des ensembles

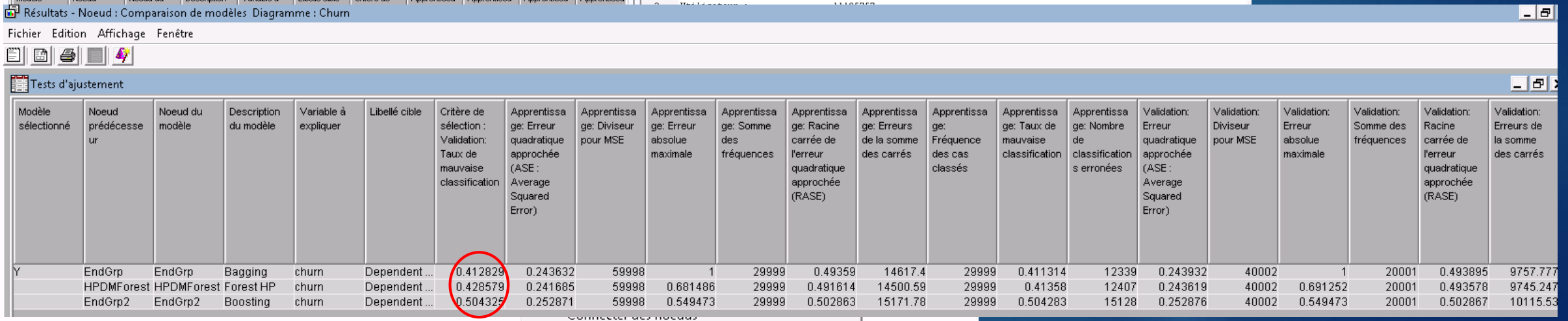


II.1 Méthodes des ensembles



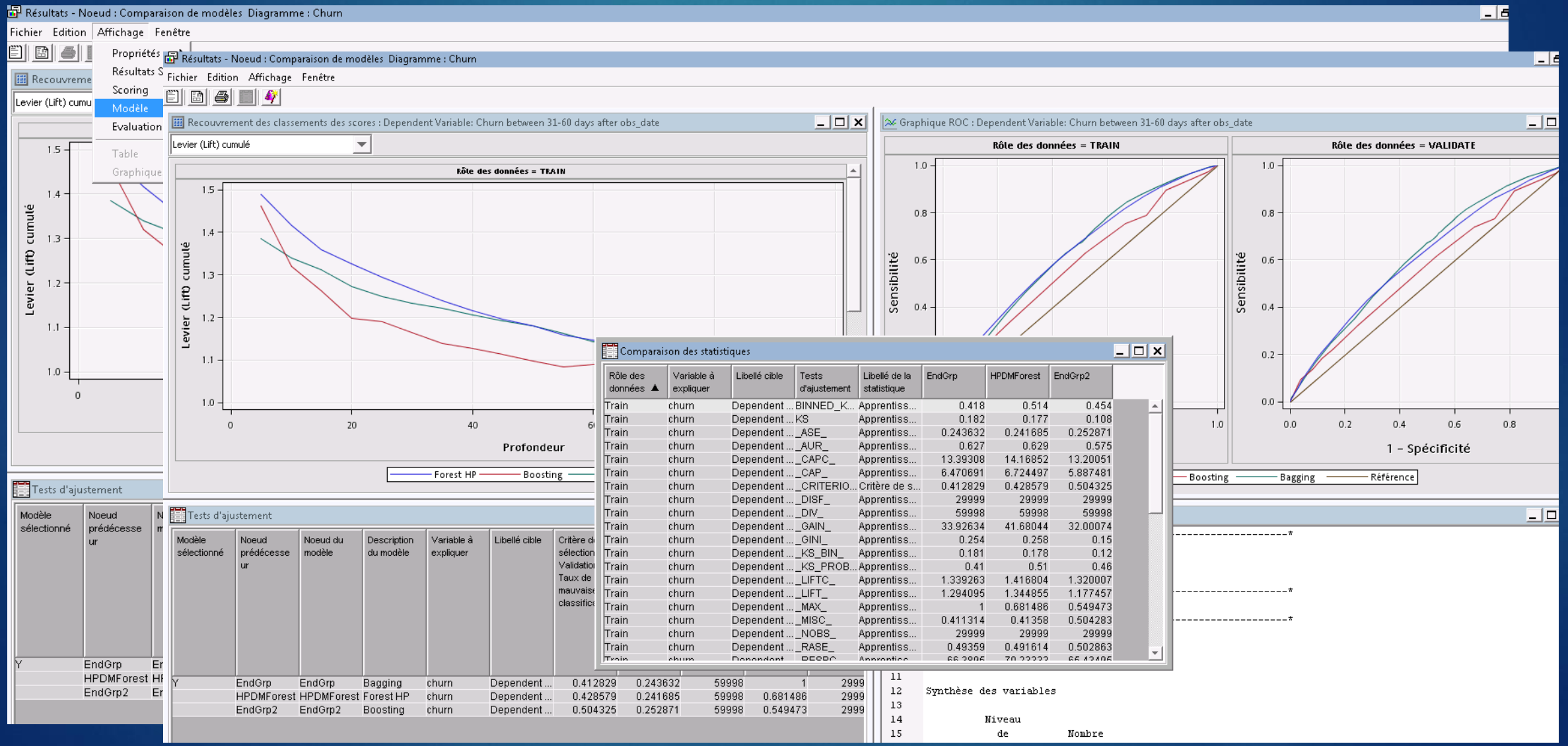
II.1 Méthodes des ensembles



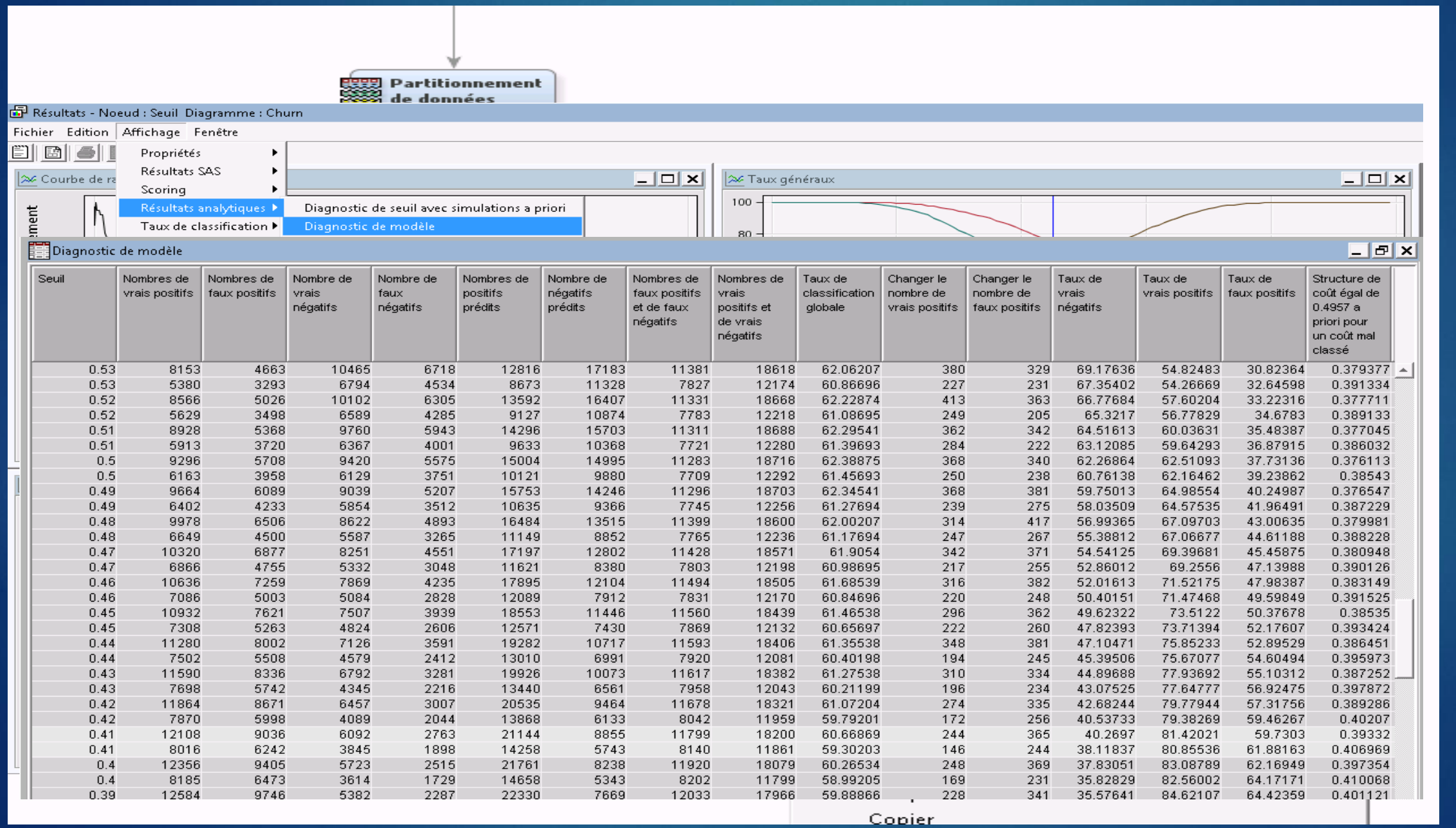


Modèle sélectionné	Noeud prédécesseur	Noeud du modèle	Description du modèle	Variable à expliquer	Libellé cible	Critère de sélection : Validation: Taux de mauvaise classification	Apprentissage: Erreur quadratique approchée (ASE : Average Squared Error)	Apprentissage: Diviseur pour MSE	Apprentissage: Erreur absolue maximale	Apprentissage: Somme des fréquences	Apprentissage: Racine carrée de l'erreur quadratique approchée (RASE)	Apprentissage: Erreurs de la somme des carrés	Apprentissage: Fréquence des cas classés	Apprentissage: Taux de mauvaise classification	Apprentissage: Nombre de classifications erronées	Validation: Erreur quadratique approchée (ASE : Average Squared Error)	Validation: Diviseur pour MSE	Validation: Erreur absolue maximale	Validation: Somme des fréquences	Validation: Racine carrée de l'erreur quadratique approchée (RASE)	Validation: Erreurs de la somme des carrés
Y	EndGrp	EndGrp	Bagging	churn	Dependent ...	0.412829	0.243632	59998	1	29999	0.49359	14617.4	29999	0.411314	12339	0.243932	40002	1	20001	0.493895	9757.777
	HPDMForest	HPDMForest	Forest HP	churn	Dependent ...	0.428579	0.241685	59998	0.681486	29999	0.491614	14500.59	29999	0.41358	12407	0.243619	40002	0.691252	20001	0.493578	9745.247
	EndGrp2	EndGrp2	Boosting	churn	Dependent ...	0.504325	0.252871	59998	0.549473	29999	0.502863	15171.78	29999	0.504283	15128	0.252876	40002	0.549473	20001	0.502867	10115.53

II.1 Méthodes des ensembles

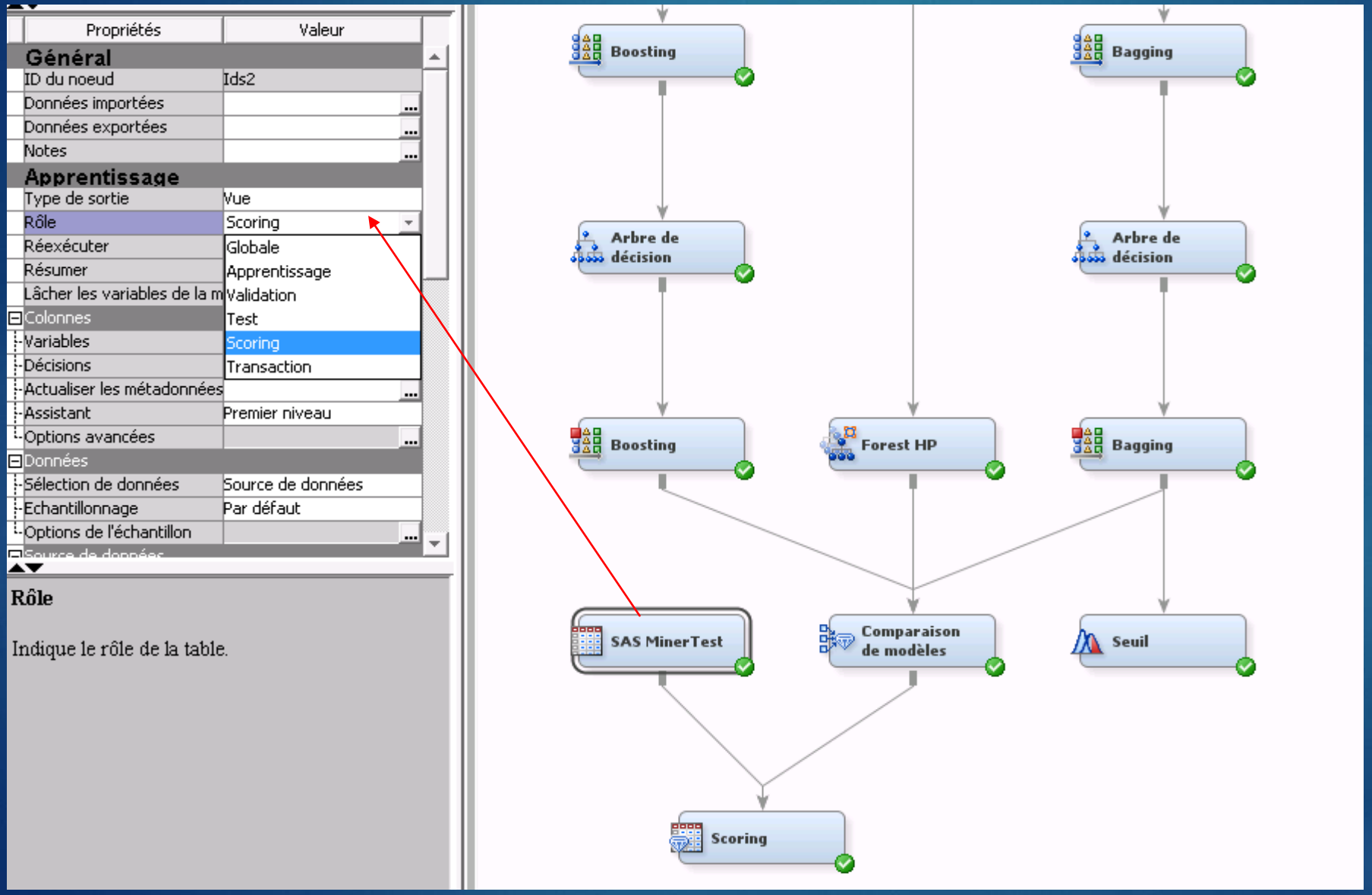


II.1 Méthodes des ensembles



Seuil	Nombres de vrais positifs	Nombres de faux positifs	Nombre de vrais négatifs	Nombre de faux négatifs	Nombres de positifs prédits	Nombre de négatifs prédits	Nombres de faux positifs et de faux négatifs	Nombres de vrais positifs et de vrais négatifs	Taux de classification globale	Changer le nombre de vrais positifs	Changer le nombre de faux positifs	Taux de vrais négatifs	Taux de vrais positifs	Taux de faux positifs	Structure de coût égal de 0.4957 a priori pour un coût mal classé
0.53	8153	4663	10465	6718	12816	17183	11381	18618	62.06207	380	329	69.17636	54.82483	30.82364	0.379377
0.53	5380	3293	6794	4534	8673	11328	7827	12174	60.86696	227	231	67.35402	54.26669	32.64598	0.391334
0.52	8566	5026	10102	6305	13592	16407	11331	18668	62.22874	413	363	66.77684	57.60204	33.22316	0.377711
0.52	5629	3498	6589	4285	9127	10874	7783	12218	61.08695	249	205	65.3217	56.77829	34.6783	0.389133
0.51	8928	5368	9760	5943	14296	15703	11311	18688	62.29541	362	342	64.51613	60.03631	35.48387	0.377045
0.51	5913	3720	6367	4001	9633	10368	7721	12280	61.39693	284	222	63.12085	59.64293	36.87915	0.386032
0.5	9296	5708	9420	5575	15004	14995	11283	18716	62.38875	368	340	62.26864	62.51093	37.73136	0.376113
0.5	6163	3958	6129	3751	10121	9880	7709	12292	61.45693	250	238	60.76138	62.16462	39.23862	0.38543
0.49	9664	6089	9039	5207	15753	14246	11296	18703	62.34541	368	381	59.75013	64.98554	40.24987	0.376547
0.49	6402	4233	5854	3512	10635	9366	7745	12256	61.27694	239	275	58.03509	64.57535	41.96491	0.387229
0.48	9978	6506	8622	4893	16484	13515	11399	18600	62.00207	314	417	56.99365	67.09703	43.00635	0.379981
0.48	6649	4500	5587	3265	11149	8852	7765	12236	61.17694	247	267	55.38812	67.06677	44.61188	0.388228
0.47	10320	6877	8251	4551	17197	12802	11428	18571	61.9054	342	371	54.54125	69.39681	45.45875	0.380948
0.47	6866	4755	5332	3048	11621	8380	7803	12198	60.98695	217	255	52.86012	69.2556	47.13988	0.390126
0.46	10636	7259	7869	4235	17895	12104	11494	18505	61.68539	316	382	52.01613	71.52175	47.98387	0.383149
0.46	7086	5003	5084	2828	12089	7912	7831	12170	60.84696	220	248	50.40151	71.47468	49.59849	0.391525
0.45	10932	7621	7507	3939	18553	11446	11560	18439	61.46538	296	362	49.62322	73.5122	50.37678	0.38535
0.45	7308	5263	4824	2606	12571	7430	7869	12132	60.65697	222	260	47.82393	73.71394	52.17607	0.393424
0.44	11280	8002	7126	3591	19282	10717	11593	18406	61.35538	348	381	47.10471	75.85233	52.89529	0.386451
0.44	7502	5508	4579	2412	13010	6991	7920	12081	60.40198	194	245	45.39506	75.67077	54.60494	0.395973
0.43	11590	8336	6792	3281	19926	10073	11617	18382	61.27538	310	334	44.89688	77.93692	55.10312	0.387252
0.43	7698	5742	4345	2216	13440	6561	7958	12043	60.21199	196	234	43.07525	77.64777	56.92475	0.397872
0.42	11864	8671	6457	3007	20535	9464	11678	18321	61.07204	274	335	42.68244	79.77944	57.31756	0.389286
0.42	7870	5998	4089	2044	13868	6133	8042	11959	59.79201	172	256	40.53733	79.38269	59.46267	0.40207
0.41	12108	9036	6092	2763	21144	8855	11799	18200	60.66869	244	365	40.2697	81.42021	59.7303	0.39332
0.41	8016	6242	3845	1898	14258	5743	8140	11861	59.30203	146	244	38.11837	80.85536	61.88163	0.406969
0.4	12356	9405	5723	2515	21761	8238	11920	18079	60.26534	248	369	37.83051	83.08789	62.16949	0.397354
0.4	8185	6473	3614	1729	14658	5343	8202	11799	58.99205	169	231	35.82829	82.56002	64.17171	0.410068
0.39	12584	9746	5382	2287	22330	7669	12033	17966	59.88866	228	341	35.57641	84.62107	64.42359	0.401121

II.1 Méthodes des ensembles



II.1 Classes rares

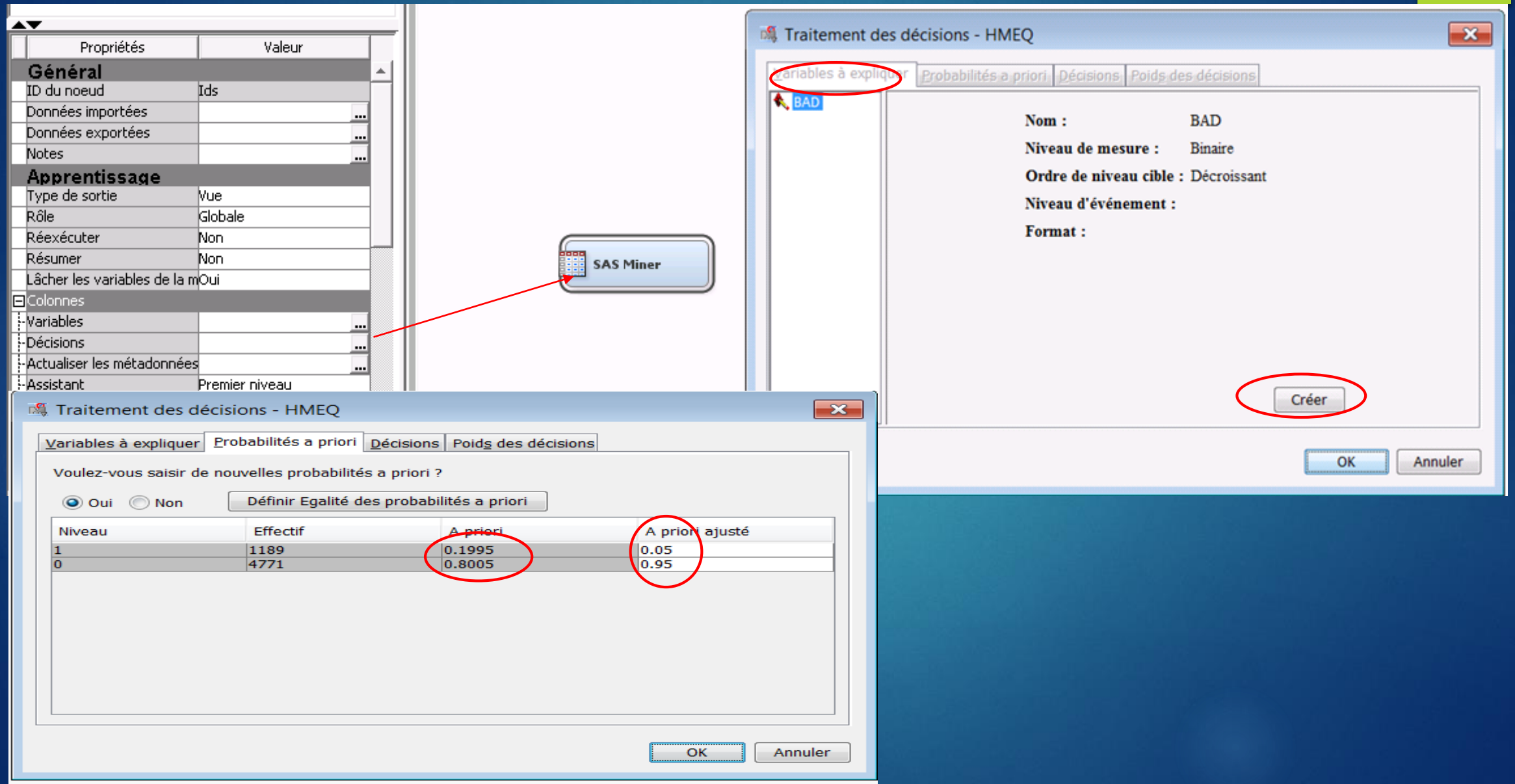
- ✓ Objectifs du modèle des classes rares :

1. Détecter ou prédire des événements relativement rares (<5%) ;
2. Éviter la réalisations des événements à conséquences coûteuses (écrasement d'avion, maladie rare, fraude, etc.);

- ✓ Méthodes de prévisions pour les classes rares

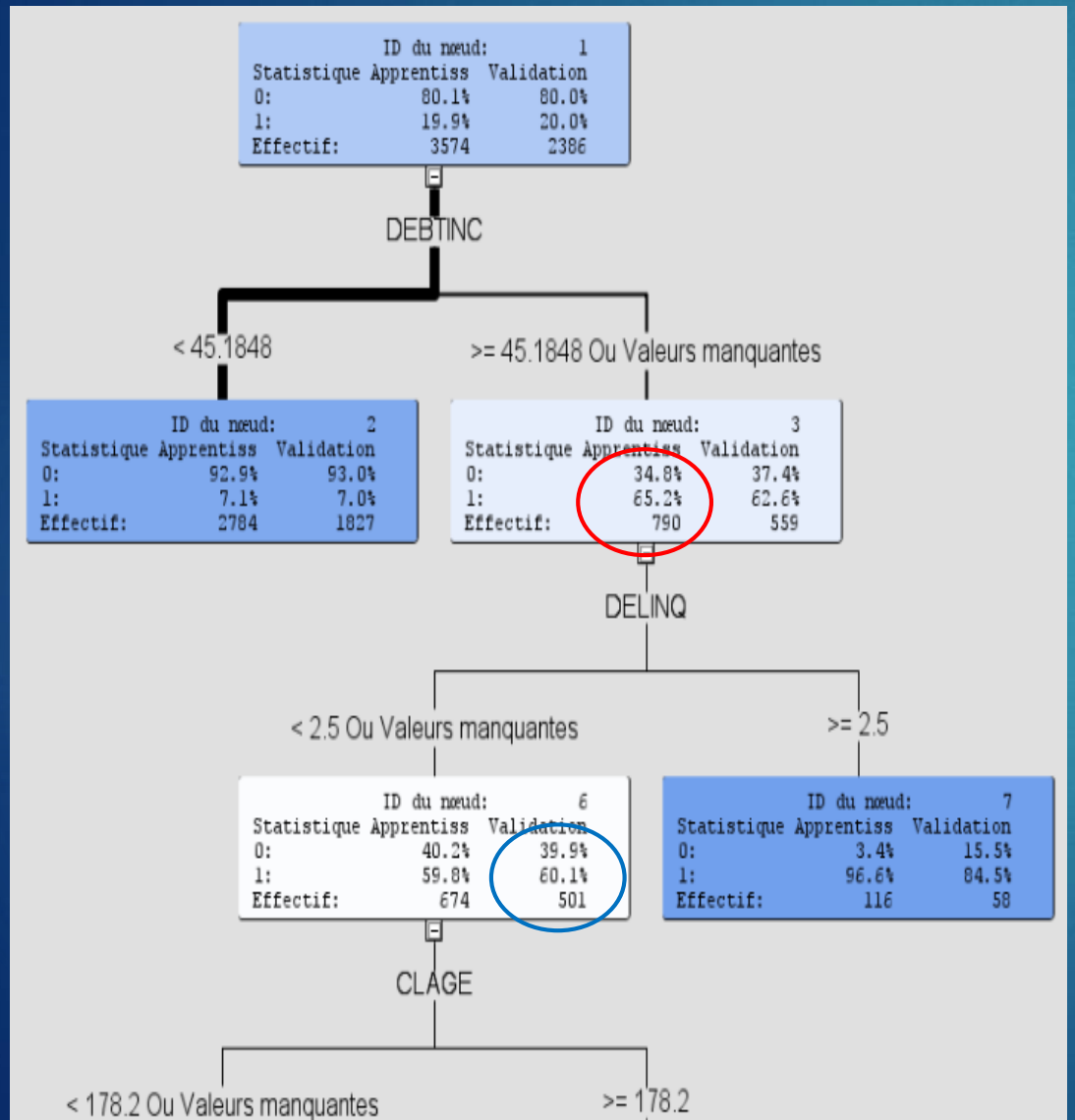
3. Un sur-échantillonnage de la classe rare (sous-échantillonnage de la classe majoritaire;
4. Méthodes basées sur le coût de classifier incorrectement les observations ;

II.1 Classes rares

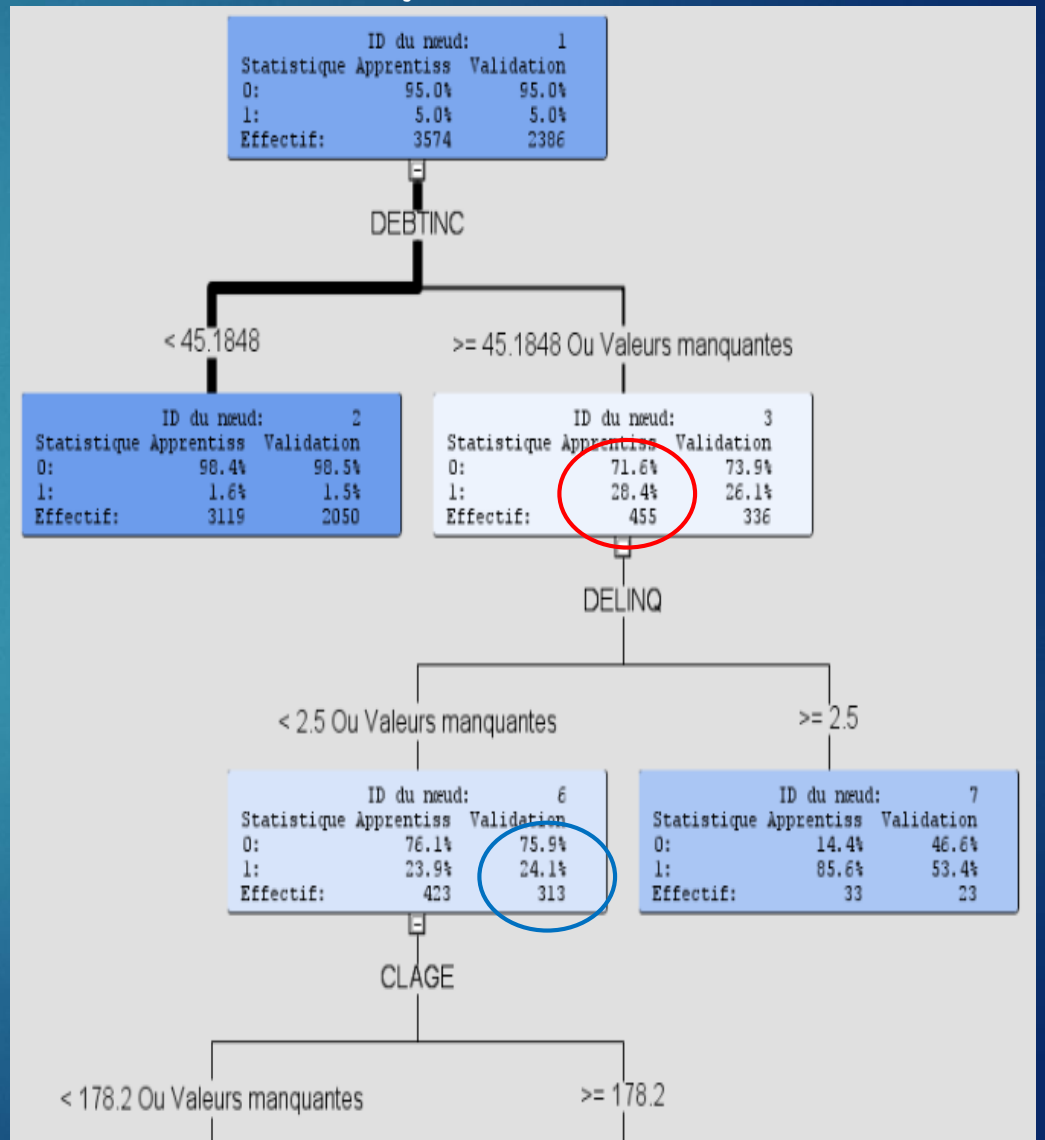


II.1 Classes rares

Sans ajustement

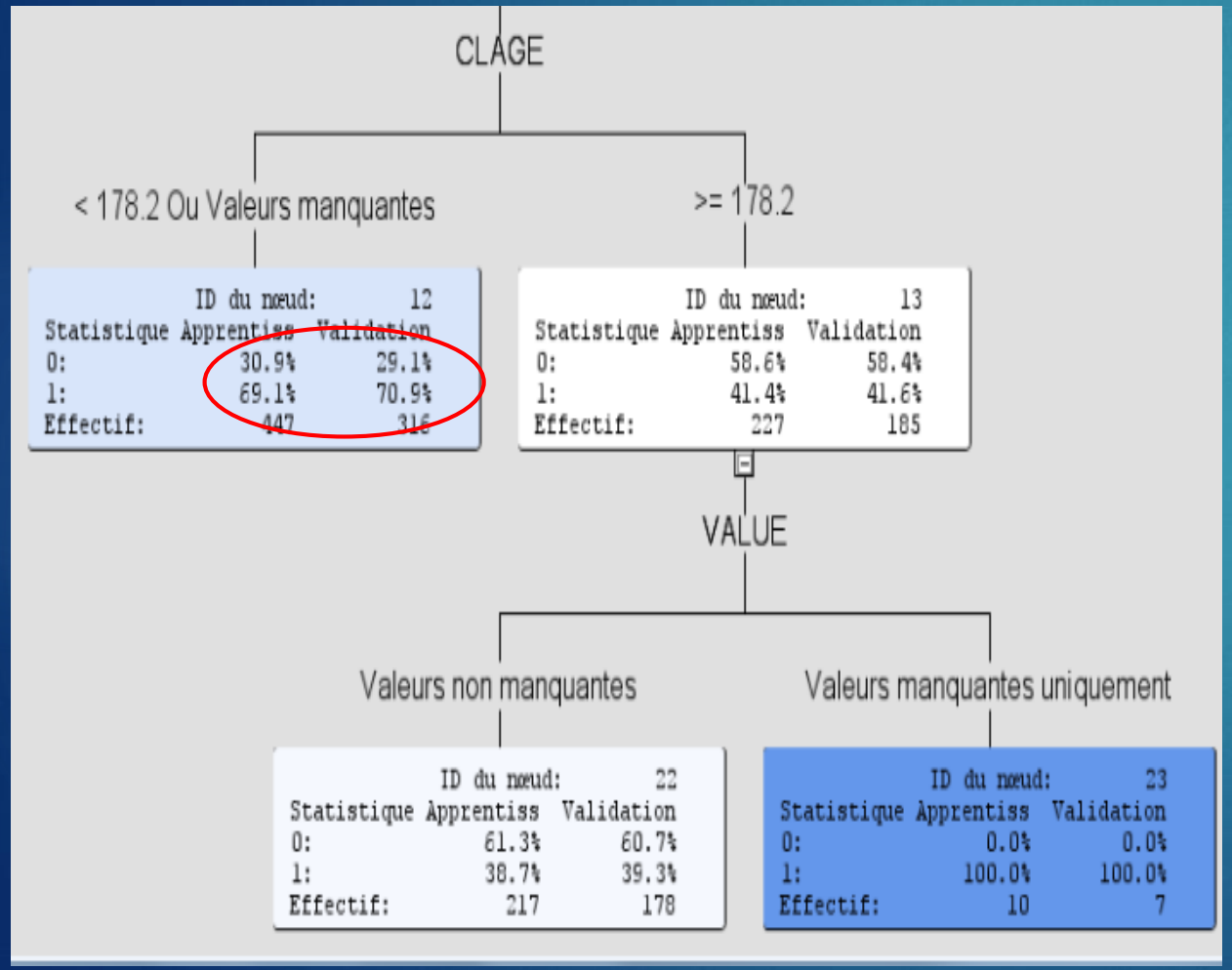


Avec ajustement

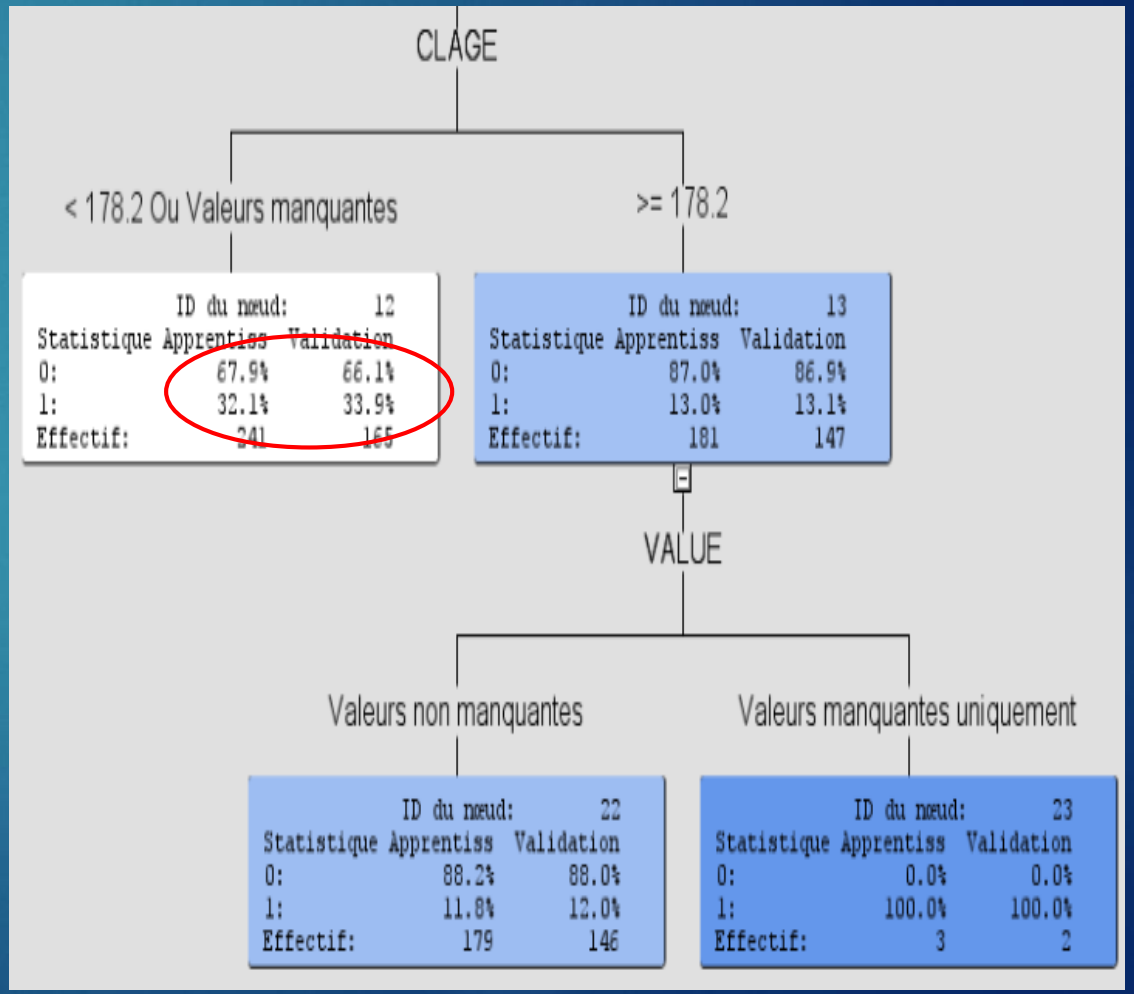


II.1 Classes rares

Sans ajustement



Avec ajustement

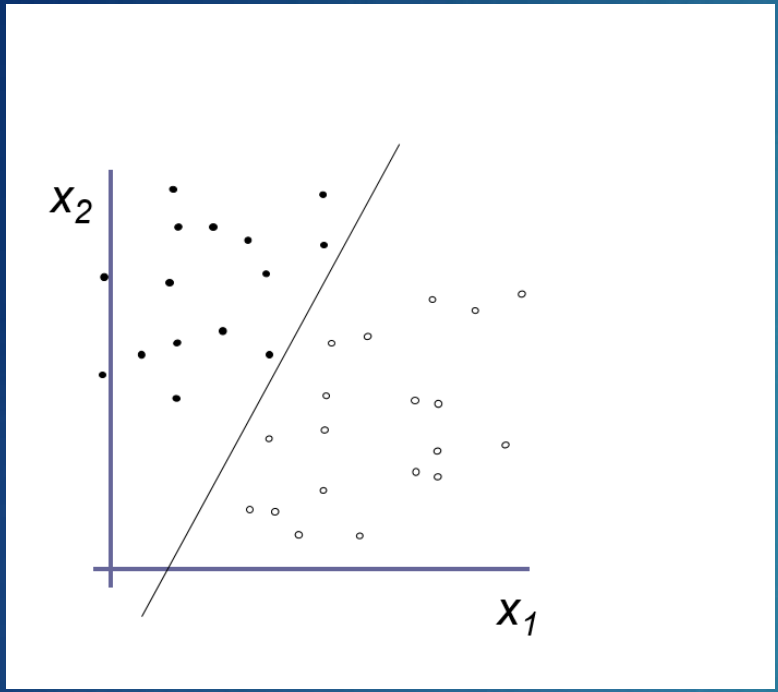


III.3 Machine à support vectoriel (SVM)

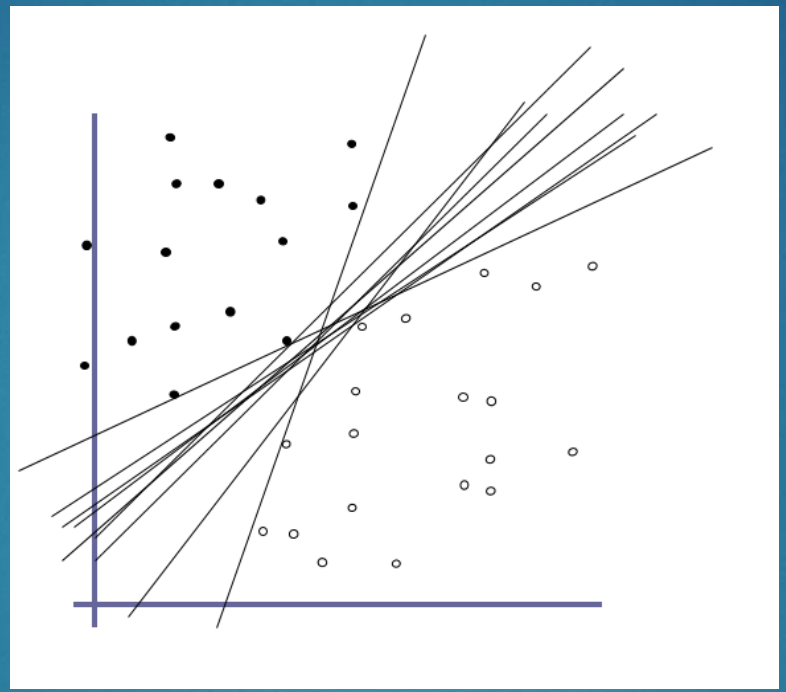
- ✓ Objectifs du modèle SVM:

1. Résoudre des problèmes de discrimination ou de régression ;
2. Modéliser des phénomènes non linéaires;
3. Une précision importante des prédictions dans certains problèmes (Bio-informatique, recherche informatique, etc.);

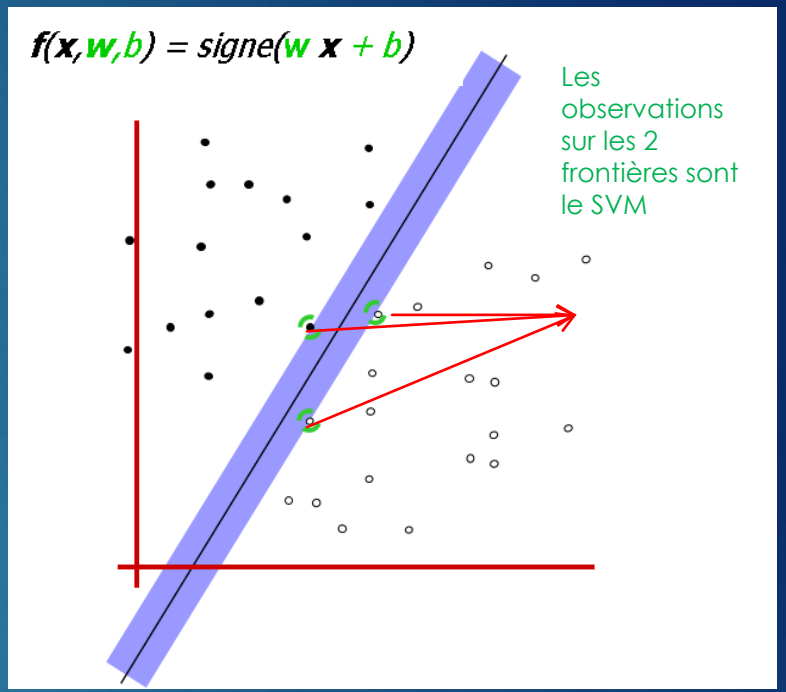
III.3 Machine à support vectoriel (SVM)



Comment nous pouvons classifier ces observations?



Toutes ces droites séparent bien les deux classes. Laquelle choisir ?



Seules les observations sur les frontières sont importantes

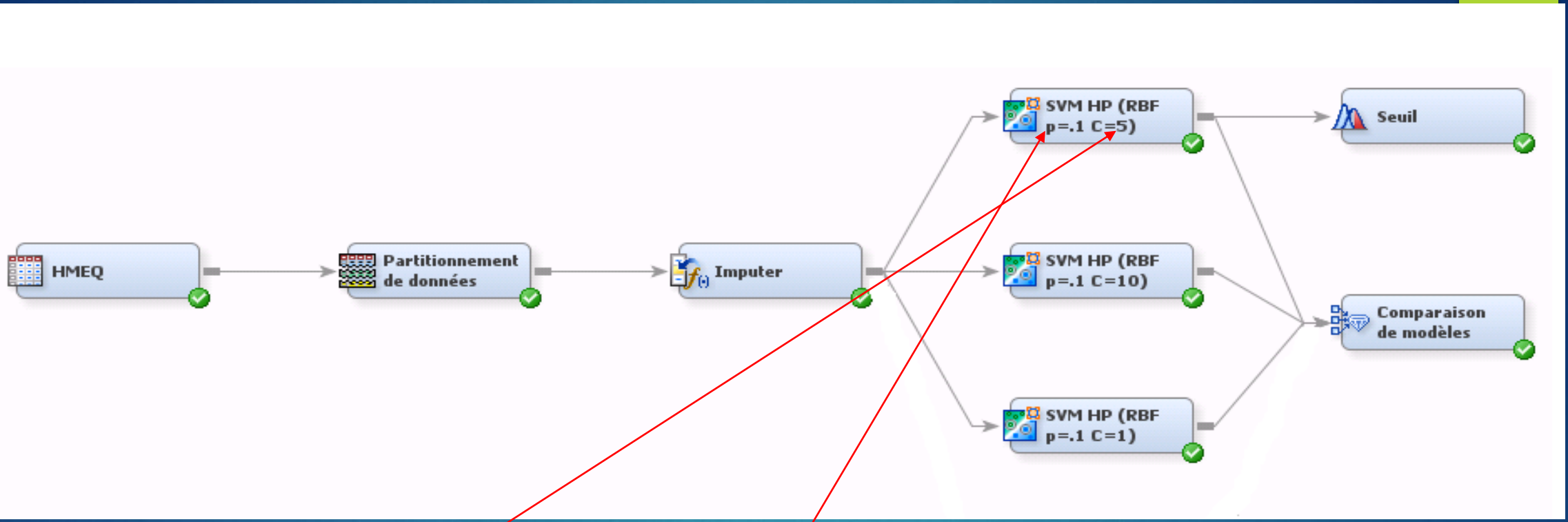
➔Le meilleur classificateur linéaire est celui qui a la marge maximale!!!

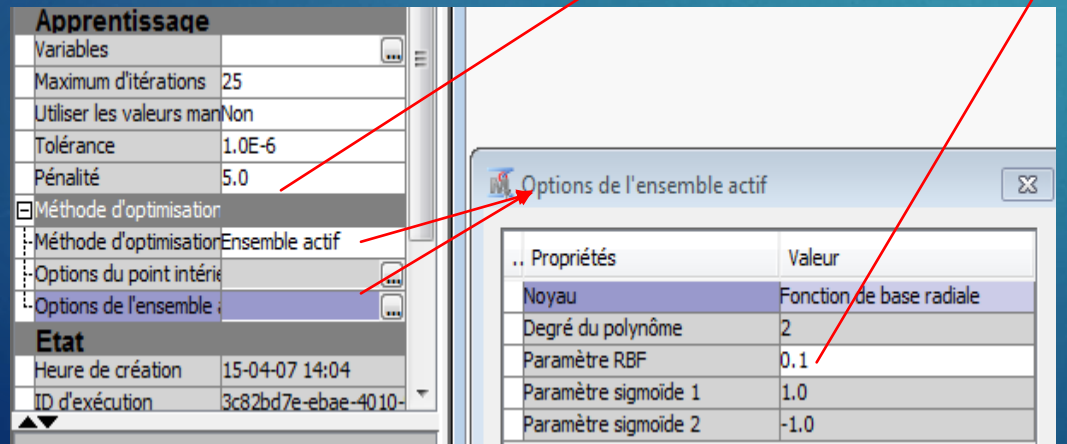
III.3 Machine à support vectoriel (SVM)

- ✓ Limites :

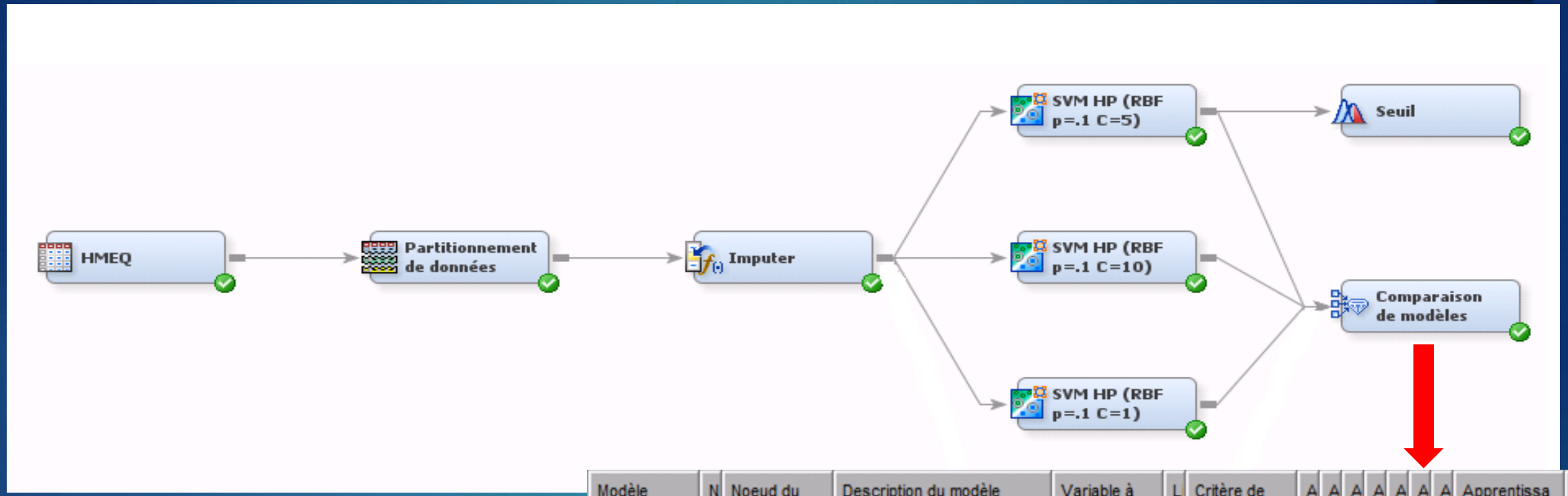
1. Le Nœud HP SVM de SAS EM peut seulement être utilisé pour les variables binaires ;
2. Les variables manquantes sont supprimées de l'analyse => imputer les valeurs manquantes;
3. Difficile à interpréter les modèles ;

III.3 Machine à support vectoriel (SVM)



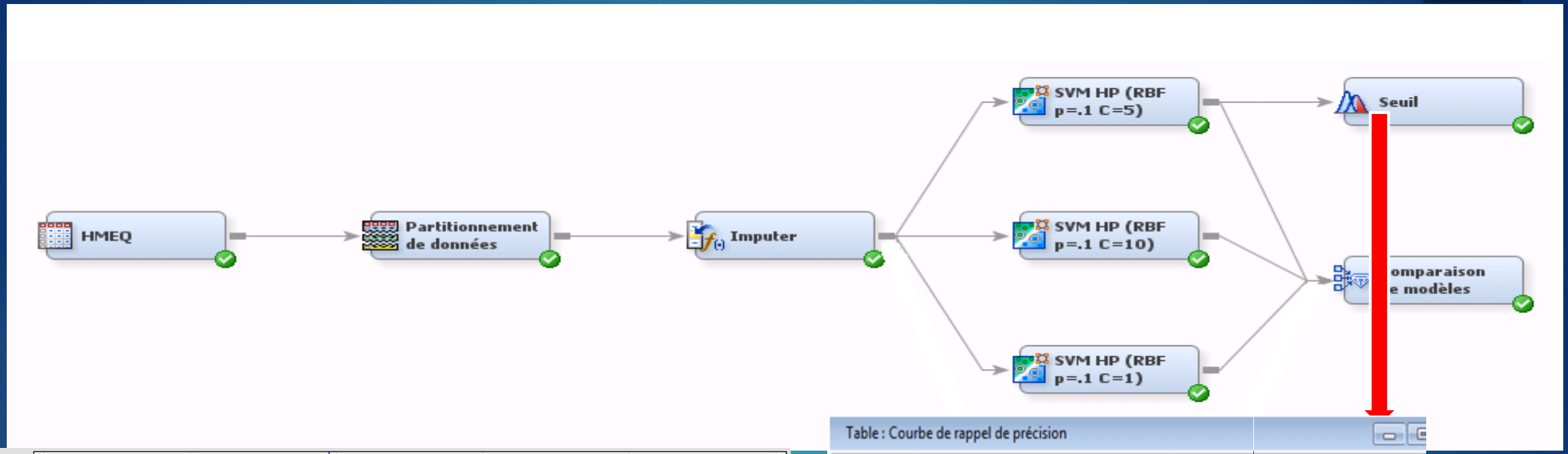


III.3 Machine à support vectoriel (SVM)



Modèle sélectionné	N	Noeud du modèle	Description du modèle	Variable à expliquer	L	Critère de sélection : Validation: Taux de mauvaise classification	A	A	A	A	A	A	A	Apprentissage: Taux de mauvaise classification
Y	...	HPSVM4	SVM HP (RBF p=.1 C=10)	BAD		0.046102	...	...	...	...	...	...	...	0
	...	HPSVM3	SVM HP (RBF p=.1 C=5)	BAD		0.04736	...	...	...	...	...	...	...	0.001959

III.3 Machine à support vectoriel (SVM)



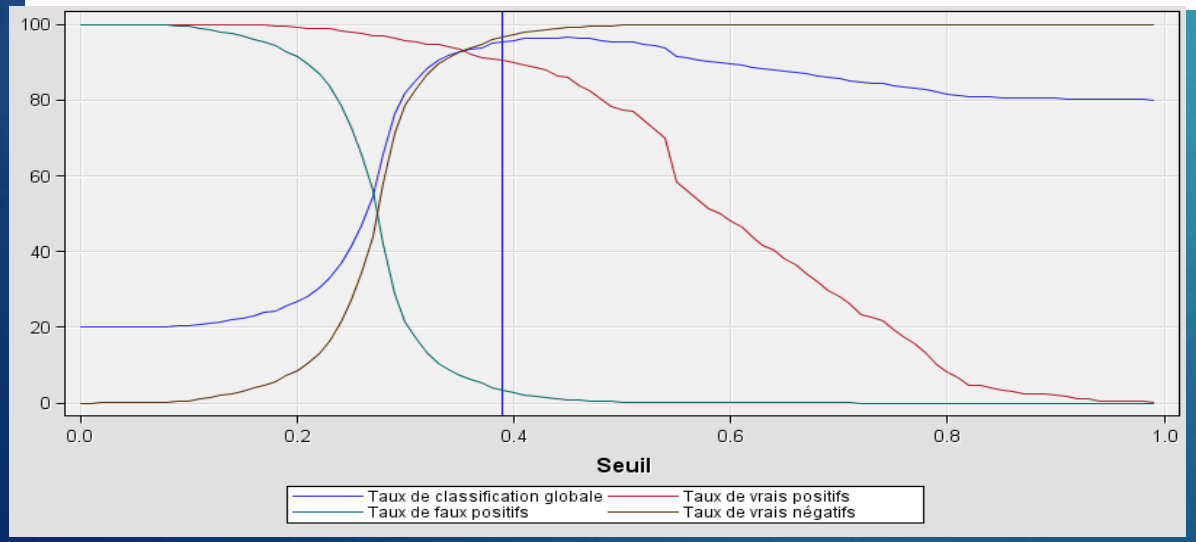


Table : Courbe de rappel de précision

Seuil	Nombres de vrais positifs	Nombres de faux positifs	Nombre de vrais négatifs	Nombre de faux négatifs	Taux de vrais positifs	Taux de vrais négatifs
0.42	710	2	2860	2	99.7191	99.93012
0.42	422	33	1876	55	88.4696	98.27135
0.41	710	2	2860	2	99.7191	99.93012
0.41	425	37	1872	52	89.09853	98.06181
0.4	710	2	2860	2	99.7191	99.93012
0.4	428	52	1857	49	89.72746	97.27606
0.39	711	2	2860	1	99.85955	99.93012
0.39	432	66	1843	45	90.56604	96.54269

Questions

Annexes

Échantillonnage :

Ajouter : Ajout des tables

Partitionnement des données : Partitionne les données en plusieurs tables (Apprentissage, validation et test)

Importation de fichier : Importer un fichier externe

Filtrer : Supprime des données en fonction des critères définis

Données en entrée : fournit des détails sur les variables contenues dans la table à étudier

Fusionner : Le nœud fusion crée de nouvelles tables ou vues en combinant des colonnes de plusieurs tables

Échantillonnage : Crée un échantillon de donnée

Annexes

Exploration :

Association : Effectue une étude d'association et recherche de séquence

Classe : Effectue une classification des observations pouvant être utilisée pour segmenter les bases de données (Analyse factorielle)

DMDB : Calcule les statistiques descriptives à l'aide de la procédure DMBD

Graphiques interactifs : Générer des rapports graphiques et des graphiques interactifs

Analyse des liens : Effectue une analyse de liens

Panier de consommation : Réalise une analyse du panier de consommation pour les données avec une potentielle taxinomie

Multiples graphiques statiques : Génère divers graphiques et courbes sur la totalité des données

Analyse de chemins : Analyse les données Web prétraitées journalisées

SOM/Kohonen : Effectue un apprentissage non supervisé en utilisant la quantification vectorielle de Kohonen (VQ), les cartes auto-organisatrices de Kohonen (SOM) ou un lot de cartes de Kohonen avec un lissage de Nadaraya-Watson ou un lissage linéaire local.

Statistiques exploratoires : Génère des statistiques univariées et bivariées.

Classification des variables : Divise un ensemble de variables explicatives en classes disjointes ou hiérarchiques

Sélection de variables : Offre un outil permettant de réduire le nombre de variables explicatives en utilisant les critères de sélection R^2 et Khi-2, etc.

Annexes

Modifier :

Supprimer : Supprime physiquement des colonnes de la table d'analyse.

Imputer : Impute les valeurs manquantes

Discrétisation interactive : Groupe les valeurs des variables dans des classes qui peuvent être utilisées comme entrées dans le modèle prédictif.

Composantes principales : Génère des composantes principales à utiliser comme entrées dans les noeuds successeurs.

Remplacement : Remplace les modalités spécifiques et inconnues pour les variables qualitatives

Générateur de règles : Vous permet de générer manuellement des règles conditionnelles ("if-then-else").

Transformer les variables : Applique des transformations aux variables de la table

Annexes

Modèle :

Neurones automatiques : Le réseau de neurones automatiques est un outil qui aide à trouver la configuration optimale d'un modèle de réseau de neurones.

Arbre de décision : Un arbre empirique représente une segmentation des données créée à l'aide d'une série de règles simples.

Régression Dmine : Calcule une régression ascendante pas-à-pas des moindres carrés incluant de manière facultative des interactions bidirectionnelles, des variables de groupe et les variables AOV16.

DMNeural : Crée un réseau de neurones sur les composantes principales découpées en classes.

Ensemble : Le noeud Ensemble crée un nouveau modèle en reprenant une fonction de probabilité a posteriori (pour les variables à expliquer qualitatives) ou les valeurs prédites (pour les variables à expliquer continues) de plusieurs modèles

Gradient Boosting : Crée une série d'arbres de décision en ajustant les résidus d'une prédiction à partir de l'arbre le plus récent de la série.

LARS : Least Angle Regressions

MBR : Le noeud Raisonnement à base de cas (MBR) crée un modèle de prévision des cibles nominales et binaires basé sur ses k plus proches voisins à partir d'une table d'apprentissage.

Importation des modèles : Vous permet d'importer et d'évaluer un modèle qui n'a pas été créé avec un des noeuds de modélisation d'Enterprise Miner.

Réseau de neurones : Les réseaux de neurones correspondent à une classe de modèles de régression non linéaires flexibles, de modèles discriminants et de modèles de réduction des données interconnectés dans un système dynamique non linéaire.

Moindres carrés partiels : Fournit plusieurs techniques de modélisation prédictive utilisant des variables latentes.

Annexes

Modèle :

Régression : Ajuste les modèles de régression linéaire et logistique.

Modélisation d'événement : Ce noeud crée des modèles de classification dans le but d'améliorer la classification des événements rares dans la variable à expliquer.

TwoStage : Modélise une variable qualitative et continue à expliquer. En général, la variable à expliquer continue est la valeur associée à un niveau de la variable à expliquer qualitative.

Évaluation :

Seuil : Nœud de seuil pour décisions cibles binaires

Décision : Utilisez le nœud Décisions pour créer ou modifier les données de décision nécessaires à la création de modèles basés sur la valeur des décisions et/ou les probabilités a priori.

Comparaison de modèle : Compare les modèles et prédictions de noeuds de modélisation précédents

Scoring : Le nœud Scoring applique le script de scoring à une table dont le rôle est SCORING.

Profil de segment : Utilisez le nœud Profil du segment pour examiner les données segmentées ou regroupées et identifier les facteurs qui différencient les segments de données de la population.

Annexes

Utilitaires :

Point de regroupement : Établit un point de regroupement au sein du diagramme.

Fin de traitement en groupe : Détermine une fin au traitement en groupe. A utiliser avec un noeud Début de traitement en groupe.

Ext Demo : tool that illustrates the various UI elements that can be used by extension nodes

Métadonnées : Le noeud Métadonnées permet de modifier les métadonnées des variables.

Intégration Open Source : Soumet les programmes écrits dans les langages "open source".

Inscrire un modèle : Inscrivez un modèle sur le SAS Metadata Server.

Annexes

Utilitaires :

Générateur de rapports : Génère un document pour les noeuds dans le flux de processus.

Code SAS : Exécute un programme SAS

Enregistrer les données : Enregistre les données sous forme de tables SAS ou dans d'autres formats à un emplacement spécifié.

Exportation du script de scoring : Le noeud Exportation du script de scoring extraie le script et les métadonnées de scoring dans un dossier. Il doit être précédé par un noeud Scoring.

Début de traitement en groupe : Vous pouvez utiliser le noeud Début de traitement en groupe pour exécuter un flux sur plusieurs groupes d'observations ou de variables. Cette propriété va de pair avec un noeud Fin de traitement en groupe.

Annexes

Applications:

Réponses incrémentale : Réponse incrémentale - Modélisation

Survie : Data Mining de survie

Série chronologique :

Corrélation TS : Analyse l'autocorrélation et la corrélation croisée des données de la série chronologique.

Préparation des données : Permet le nettoyage, l'agrégation, la transformation, la transposition, etc. des données de série chronologique.

Décomposition de la série chronologique : Calcule la décomposition saisonnière classique des données de la série chronologique.

Réduction de la dimension TS : Réduit la dimension des séries chronologiques à l'aide de la transformation par ondelettes discrète (DWT), la transformation de Fourier discrète (TFD), la décomposition de la valeur singulière ou l'approximation du segment de droite.

Lissage exponentiel TS : Génère des prévisions à l'aide de modèles de lissage exponentiel avec des constantes de lissage optimisées pour un grand nombre de séries chronologiques.

Similarité des séries chronologiques : Calcule les mesures de similarité associées aux données horodatées ou de série chronologique.

Annexes

Text Minig :

Classification du texte : Le noeud Classification du texte est utilisé pour réaliser une analyse de classification sur une collection de documents. Le noeud Classification du texte doit être précédé de noeuds Décomposition du texte et Filtre du texte, et peut également être précédé d'un noeud Thèmes du texte.

Filtre du texte : Appliquer des filtres pour réduire le nombre de termes ou de documents qui seront analysés. Le noeud Filtre du texte doit être précédé du noeud Décomposition du texte et peut être suivi d'un autre noeud Filtre du texte et d'un noeud Thèmes du texte.

Importation du texte : Extrait le texte à partir de documents contenus dans un répertoire et crée une table de résultats. Le noeud Importation de texte peut également analyser le Web en commençant à partir d'une URL indiquée, et extraire les pages Web qu'il trouve.

Décomposition du texte : Décomposer une collection de documents afin de quantifier l'information concernant les termes présents dans les documents. Le nœud de décomposition s'applique à des données textuelles telles que les courriers électroniques, articles de presse, pages Web, publications et enquêtes.

Profil du texte : Le noeud Profil du texte est utilisé pour associer les termes descriptifs avec différents niveaux d'une variable à expliquer.

Générateur de règles : Le nœud Générateur de règles génère un ensemble ordonné de règles booléennes qui serviront à décrire et prédire une variable cible. Le nœud doit être précédé des nœuds Décomposition du texte et Filtre du texte.

Thème du texte : Le noeud Thèmes du texte est utilisé pour détecter automatiquement des thèmes à partir d'une collection de documents. Pour chaque thème créé, une variable est ajoutée à la table d'apprentissage que le noeud exporte. Le noeud Thèmes du texte doit être précédé du noeud Décomposition du texte et peut être suivi d'un ou plusieurs autres noeuds Filtre du texte.