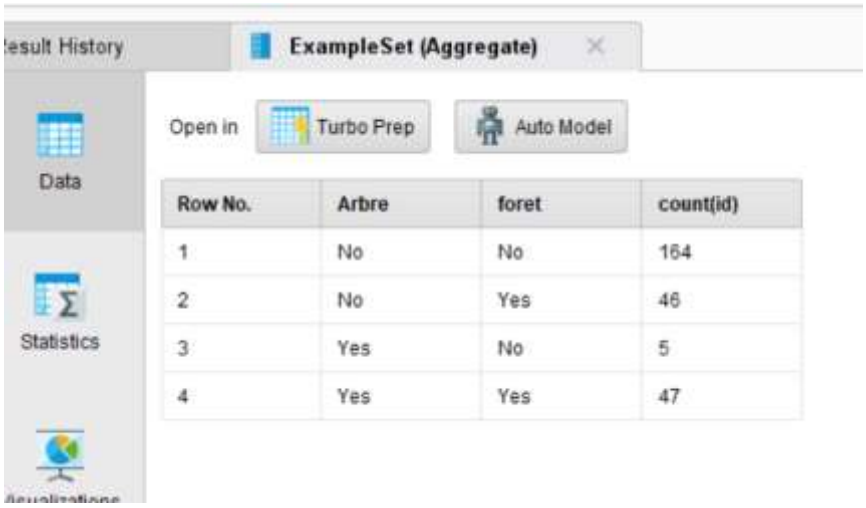
TD4

Classification sous RapidMiner

Le jeu de données Titanic disponible avec RapidMiner décrit les passagers du bateau Titanic avec différents critères. Pour chacun de ces passagers, nous disposons d'un attribut indiquant si ce dernier a survécu au naufrage ou pas.

1. Construire à partir du jeu de données Titanic un arbre de décision permettant visualiser les conditions qui ont fait qu'un passager ait survécu ou pas. C'est donc l'attribut « Survived » qu'on considère comme la Classe. Il faut donc que cet attribut ait le rôle « Label » (opérateur Set Role). Dans un premier temps, on ne touche à aucun des paramètres par défaut (utiliser l'opérateur Decision Tree).
2. L'arbre précédent est difficilement exploitable dans le sens où il contient « trop » de nœuds le rendant difficilement interprétable. Pour réduire la taille de l'arbre, on peut réduire sa hauteur maximale (par défaut c'est 10). Fixer cette hauteur à 4 et observer la taille du nouvel arbre.
3. Pour vérifier si l'arbre obtenu réalise de bonnes performances, on partitionne d'abord le jeu de données Titanic en deux parties : l'une de 80% pour apprendre l'arbre et l'autre de 20%, pour tester l'arbre (opérateur Split). Une fois le modèle construit à partir des 80%, on l'applique aux 20% (opérateur Apply Model). Observer le résultat de l'application du modèle.
4. On veut maintenant avoir un résumé de la performance du modèle (les mauvaises/bonnes classification). Pour cela, une fois que l'arbre appliqué au jeu de teste, on applique l'opérateur Performance (Classification) qui retourne un tableau de vérité.
5. Au lieu d'utiliser un arbre de décision, créer un classifieur Forêt aléatoire (opérateur « Random Forest »). Il s'agit d'un ensemble d'arbres de décision chacun créé à partir d'un sous ensemble du training set. La classification se fait en appliquant la règle de majorité : chaque arbre de la forêt prédit une classe et c'est la classe majoritaire qui sera retenue. Comparer ses performances à celles obtenues par l'arbre de décision.
6. On veut maintenant faire une analyse plus fine sur les 2 modèles (arbre vs forêt) : On veut compter le nombre d'enregistrements où les deux modèles prédisent OUI, le nombre d'enregistrements où ils prédisent tous les deux NON, le nombre d'enregistrements où l'arbre prédit OUI et l'autre prédit NON et inversement.



Row No.	Arbre	foret	count(id)
1	No	No	164
2	No	Yes	46
3	Yes	No	5
4	Yes	Yes	47