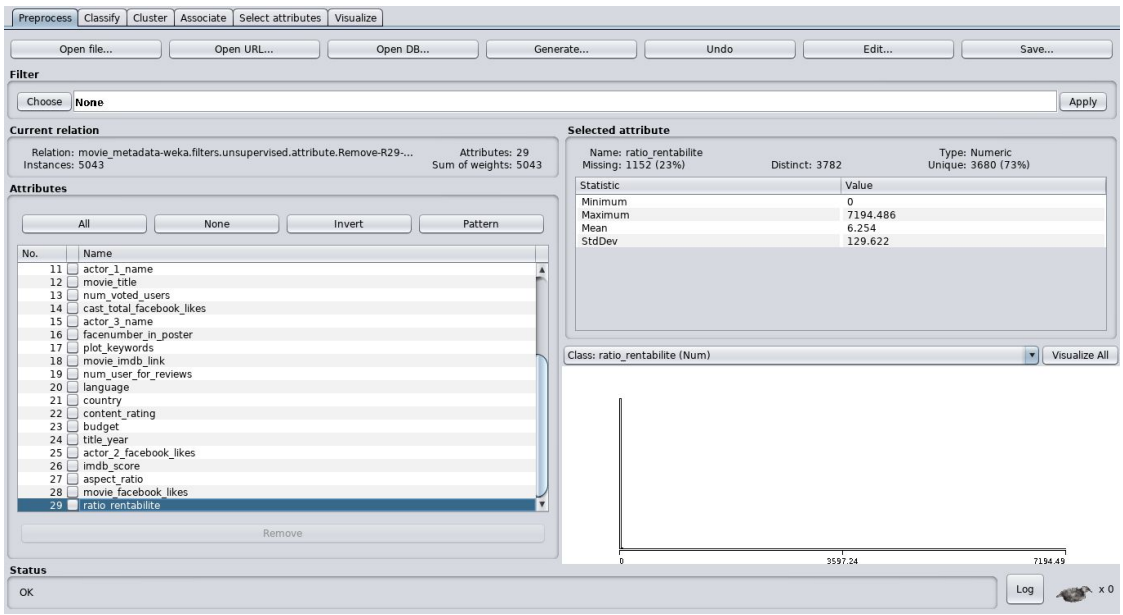
Projet Fouille de données

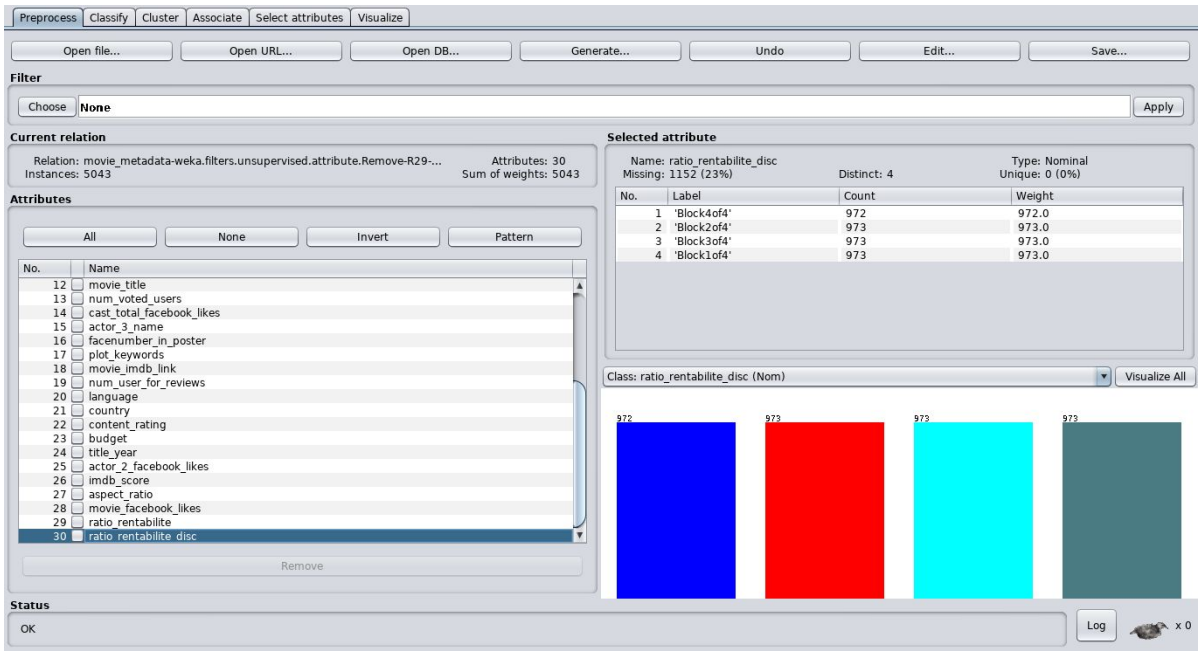
Partie 1 – Analyse descriptive des données:

Création d'attributs :

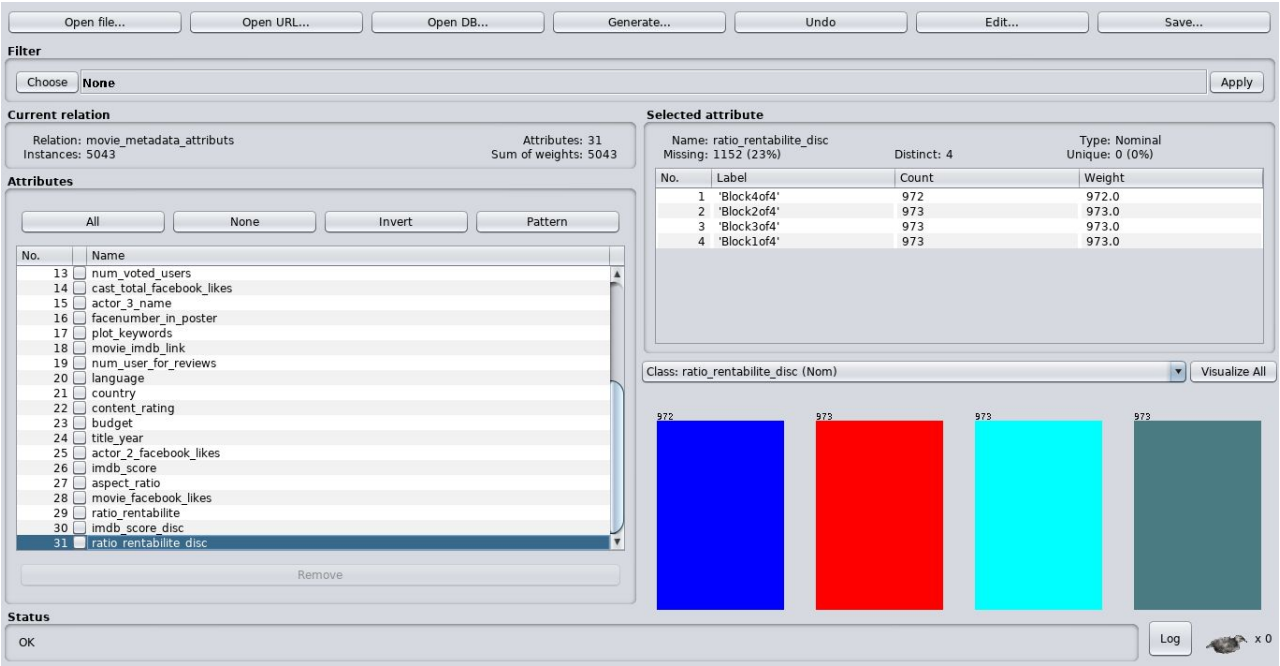
- Création un attribut « ratio_rentabilité » qui est un ratio entre ce qu'a rapporté le film et ce qu'il a couté.



- Création de l'attribut « ratio_rentabilitée_disc » qui discrétisé de façon pertinente le ratio de rentabilité précédemment calculé.



- Création de l'attribut « score_IMBD_disc » qui discrétisera le score IMDB.



Attributs et Statistique descriptive

1:L'attribute color → Type String, Couleur film (noir ou blanc).
Remarque : si le film il est en couleur il reçoit plus de j'aime.

2:L'attribut director_name → Type String, ça représente le réalisateur d'un film.

3:L'attribut num_critic_for_reviews→ Type Numérique, ça représente le nombre commentaires sur un film.

4:L'attribut duration→ Type Numérique, ça représente la durée du film par (heure, minutes , secondes).

5:L'attribut director_facebook_likes → Type Numérique, ça représente le nombre de j'aimes que le réalisateur a eu facebook a propos d'un film.

6:L'attribut actor_3_facebook_likes → Type Numérique, ça représente le nombre de j'aime que le 3éme acteur d'un film à eu sur facebook.

7:L'attribut actor_2_name →Type String, ça représente le nom du deuxième acteur dans un film.

8:L’attribut actor_1_facebook_likes → Type Numérique, ça représente le nombre de j’aime que le 1ér acteur d’un film à eu sur facebook.

9:L’attribut gross→ Type Numérique, ça représente la somme d’argent gagné par un film (les bénéfices d’un film).

L’attribut genres → Type String, ça représente le style d’un film (reportage , documentaire, film , série) d’un film à eu sur facebook.

L’attribut actor_1_name → Type String, ça représente le nom du premier acteur du film.
L’attribut movie_title → Type String, ça représente le titre du film.

L’attribut num_voted_users → Type Numérique, ça représente le nombre de personne qui ont voté pour un film.

L’attribut cast_total_facebook_likes →Type numérique, ça représente le total de j’aimes du film parmi les autres films. (genre d’un classement).

L’attribut actor_3_name →Type String, ça représente le nom du 3éme acteur du film.

L’attribut facenumber_in_poster → Type numérique , ça représente l le nombre de visage sur le poster du film.

L’attribut plot_keywords →Type String, ça représente les mots clé pour rechercher un film.

L’attribut movie_imdb_link → Type String, ça représente le lien pour accéder à un film qui est dans le site imdb.

L’attribut num_user_for_reviews →Type numérique ça représente le nombre d’utilisateurs qui ont fait des critiques sur ce film.

L’attribut language →Type String, ça représente la langue dans laquelle le film est réalisé

L’attribut country →Type String, ça représente le pays dont le film à été tourné et réalisé.

L’attribut content_rating →Type String, le nombre d’étoile pour faire une évaluation sur ce film.

L’attribut budget →Type numérique , ça représente le budget total du film.

L’attribut title_year →Type numérique , ça représente l’année de la sortie du film.

L’attribut actor_2_facebook_likes →Type numérque , ça représente le le nombre de j’aimes que le deuxiéme acteur à récolté.

L’attribut imdb_score →Type numérique , ça représente le nombre du vote pour le film sur le site imdb.

L’attribut aspect_ratio →Type numérique , le format de la réalisation du film (exemple : 4 *3 ,16 * 9).

L’attribut movie_facebook_likes →Type numérique, le nombre de j’aimes que le film a pu récolté sur facebook.

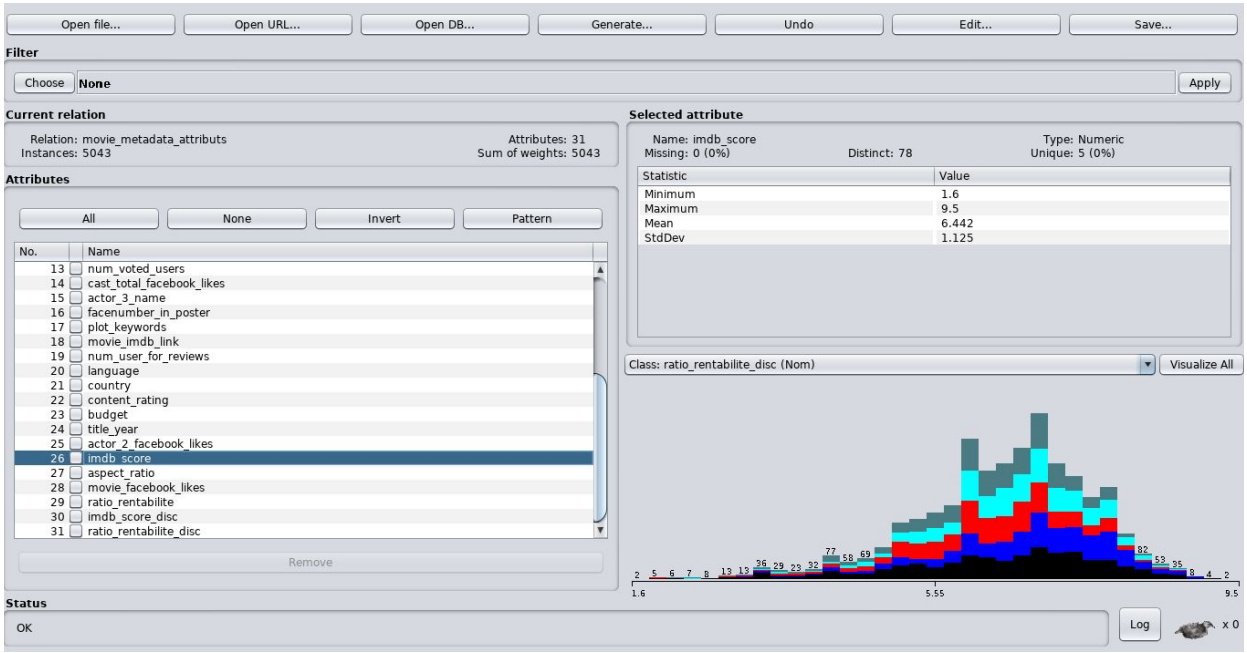
L’attribut ratio_rentabilite →Type numérique, il représente la rentabilité du film.

L’attribut ratio_rentabilite_disc →Type numérique, il représente un interval qui sert à regrouper des films par rapport a leurs rentabilité.

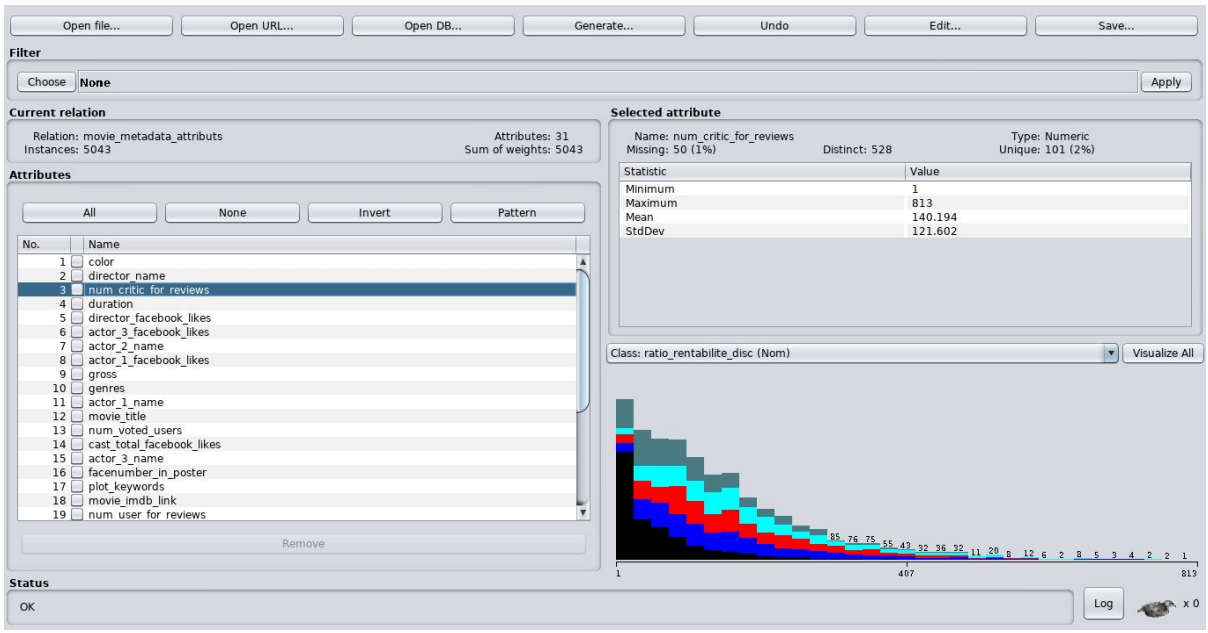
L’attribut score_IMBD_disc → Type numérique, il représente un interval qui sert à regrouper des films par rapport au score IMBD.

Statistique descriptive:

- On remarque que la rentabilité (resp score imbc) augmente avec l'augmentation du score imbc..



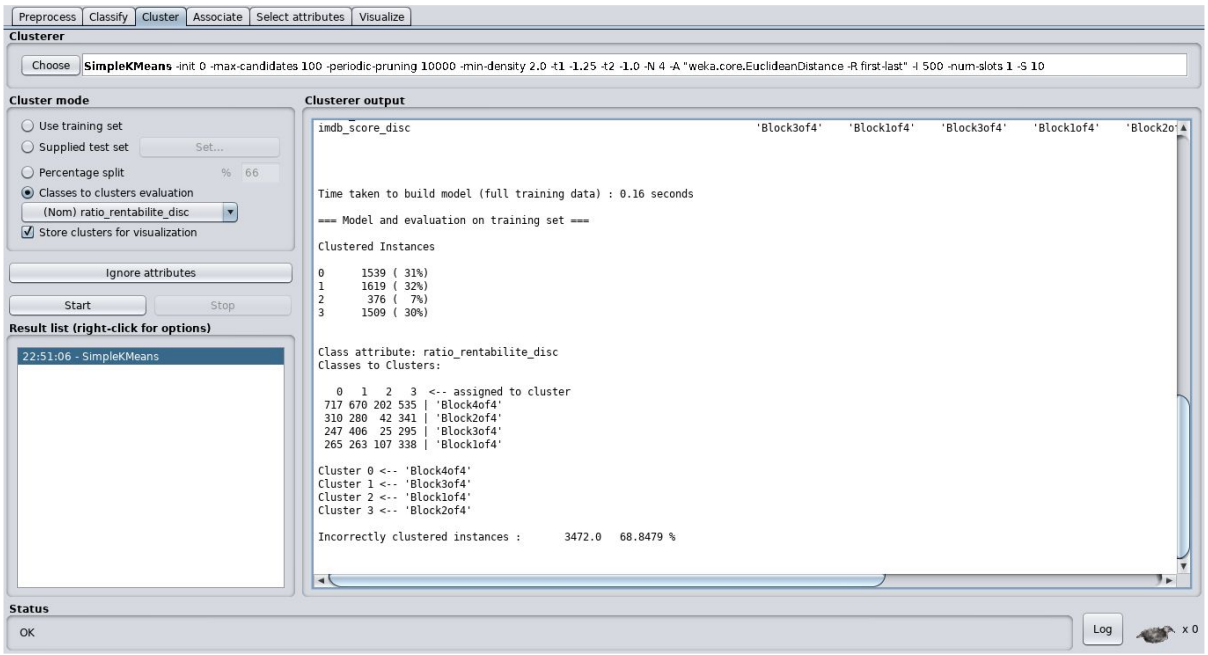
- On remarque que la rentabilité (resp score imbc) diminu avec l'augmentations des critiques.



- On remarque que la rentabilité (resp score imbc) diminu avec l'augmentations des critiques.

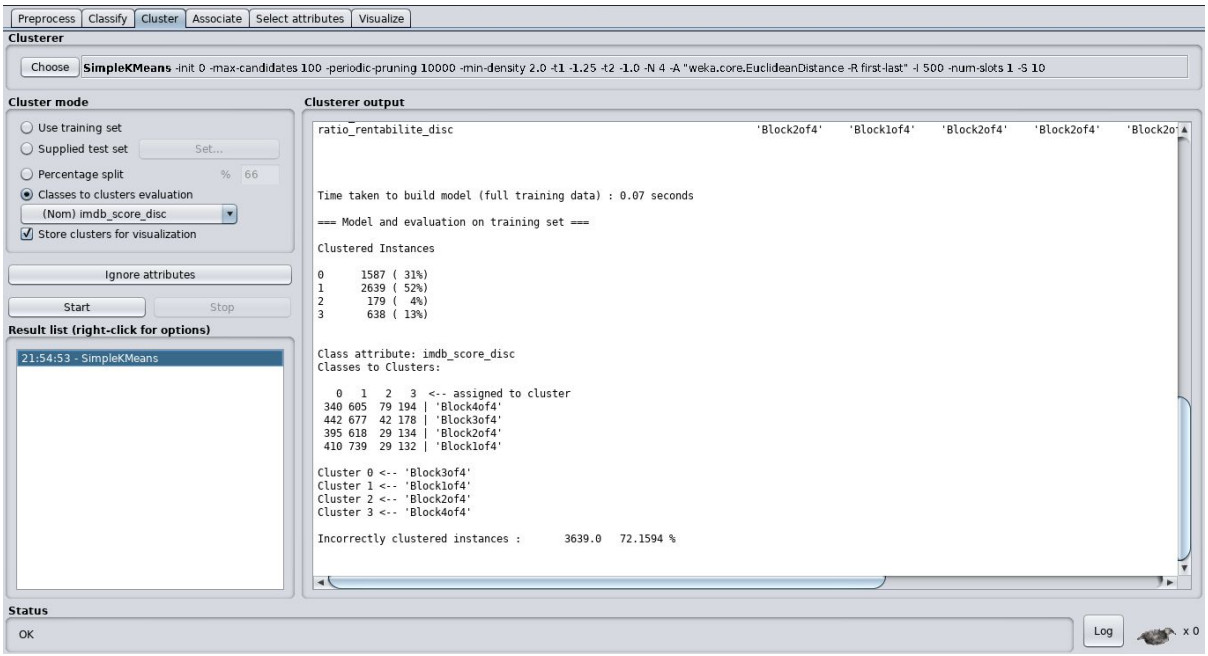
Partie 2 – Segmentation de films

SimpleKmeans avec 4 clusters sur rentabilite:



Après avoir tester plusieurs découpages, on a fini par choisir un découpage en 4 clusters.
On a obtenu un Incorrectly clustered instances de 3472 68.8479 %.

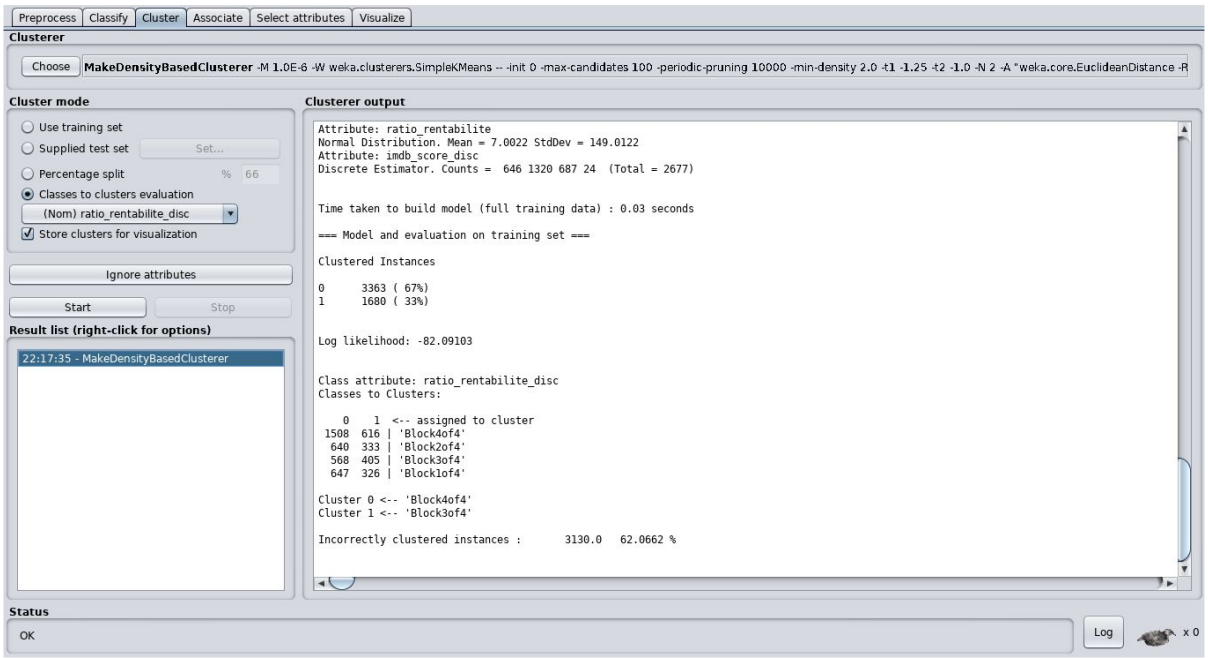
SimpleKmeans avec 4 clusters sur imdb score:



On a obtenu un Incorrectly clustered instances de 3639 72.1594 %.

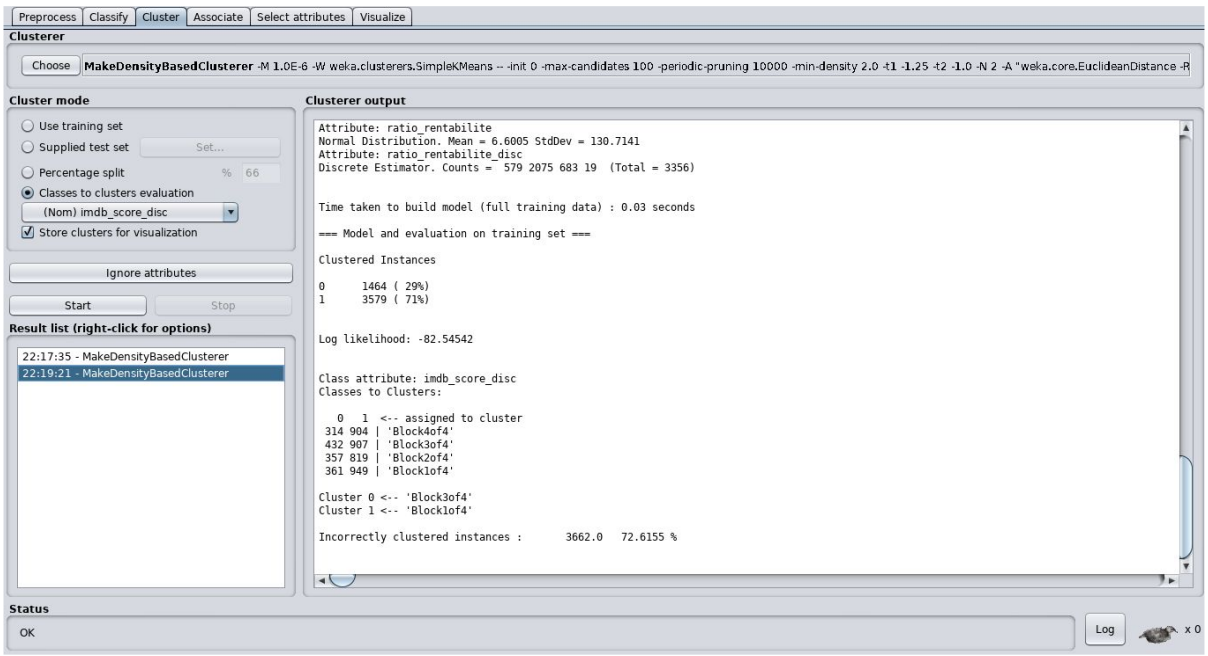
En utilisant SimpleKmeans: un meilleur clustering est obtenu avec le ratio rentabilité par rapport à au score imdb.

MakeDensityBasedCluster avec rentabilité:



Incorrectly clustered instances : 3130.0 62.0662 %

MakeDensityBasedCluster avec score_imdb:



Incorrectly clustered instances : 3662.0 72.6155 %

En utilisant MakeDensityBasedCluster: un meilleur clustering est obtenu avec le ratio rentabilité par rapport à au score imdb.

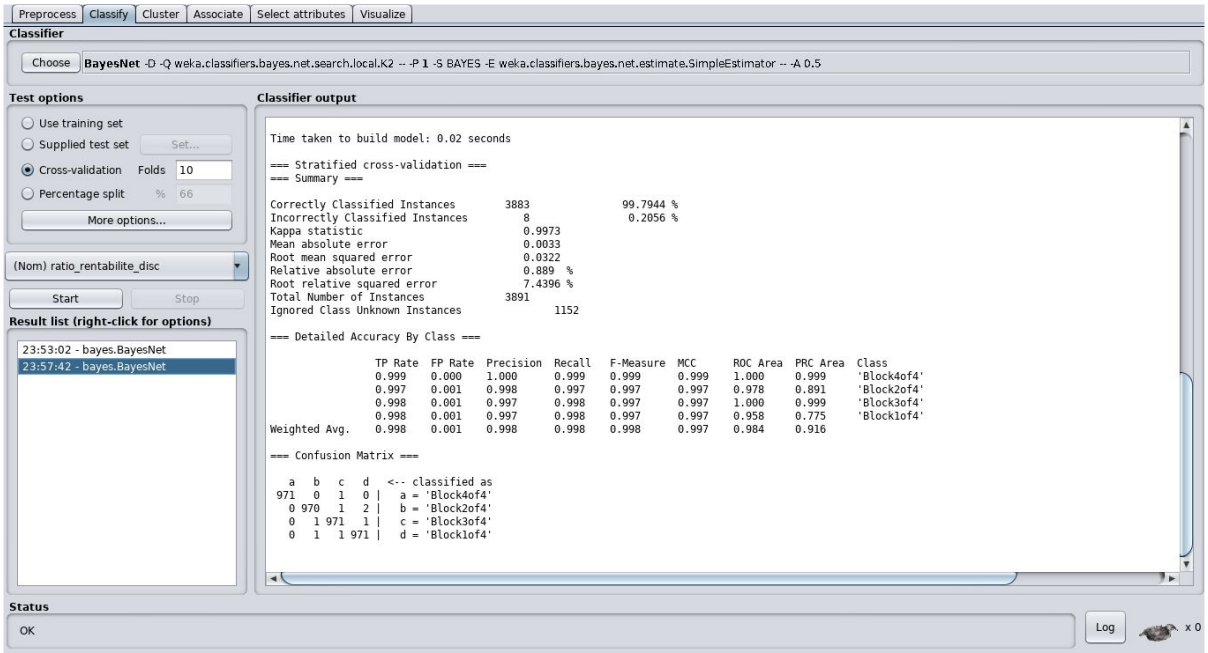
Partie 3 – Prédiction

L'objectif est de trouver le meilleur algorithme pour trouver le score IMBD puis le ratio de rentabilité.

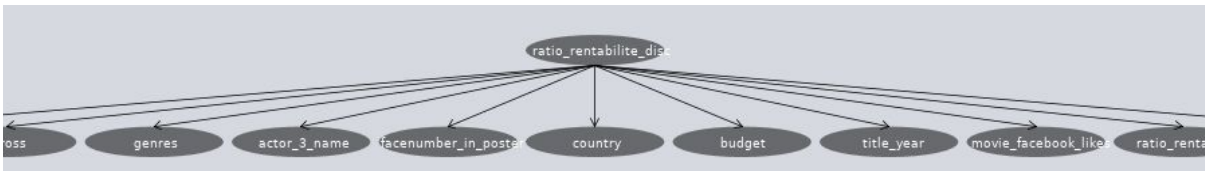
Score IMDB:

BayesNet:

BayesNet résultat:

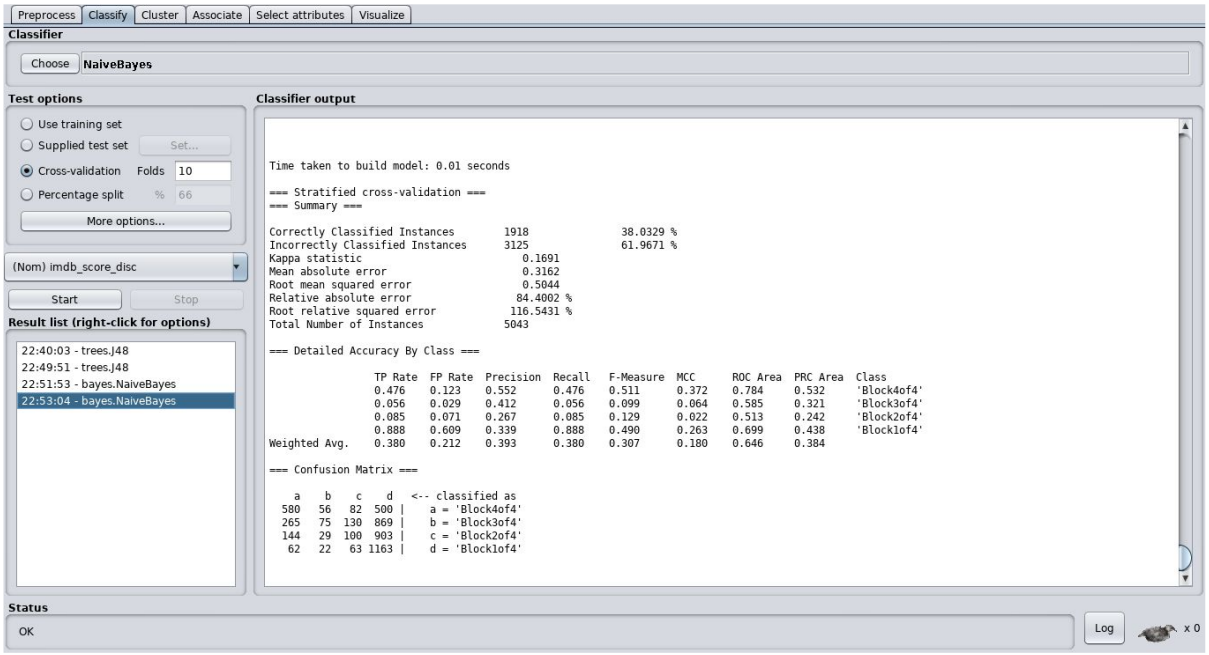


BayesNet Graphe:

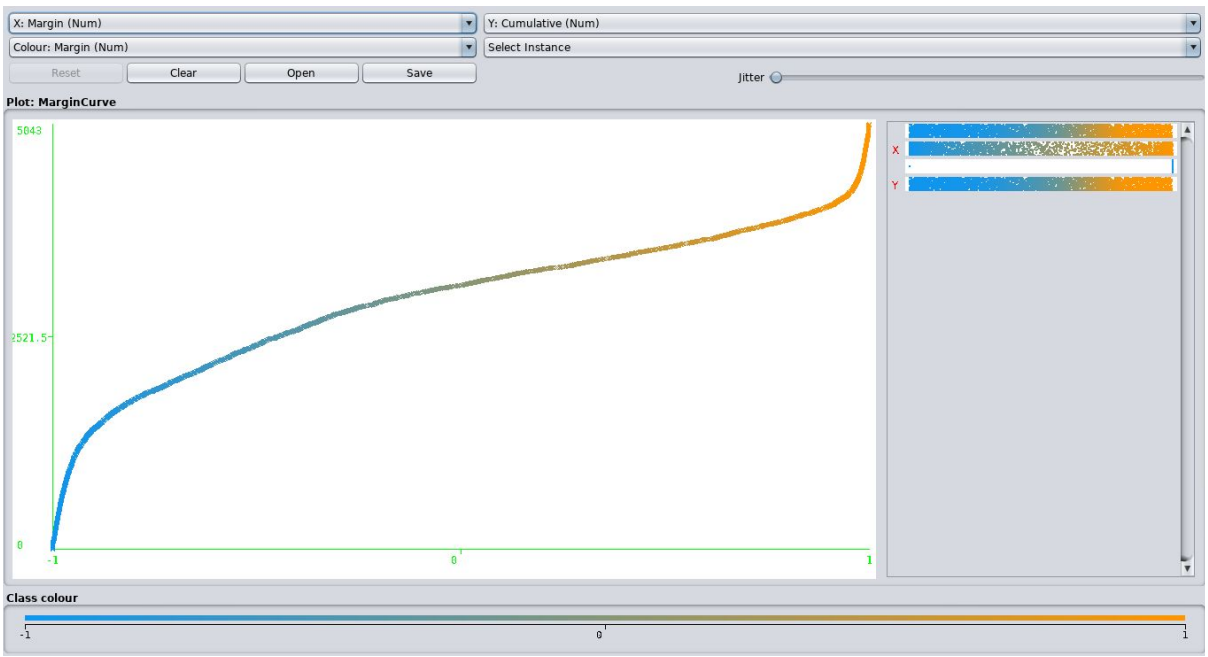


NaiveBayes:

NaiveBayes résultat:

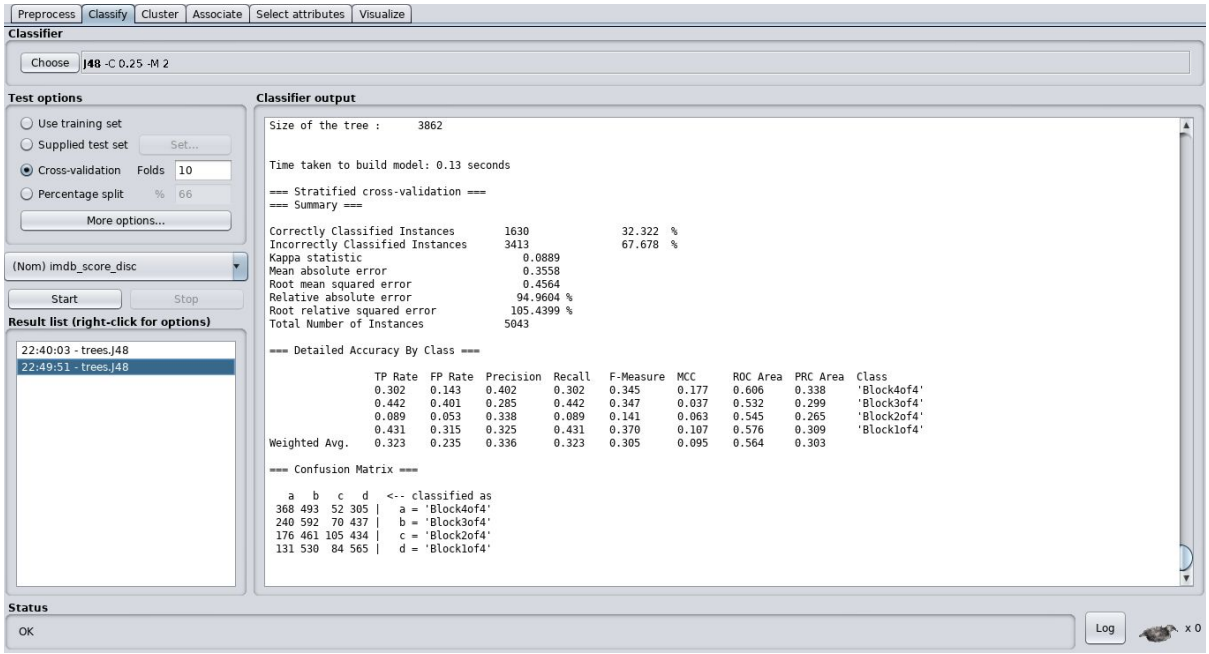


NaiveBayes curve:

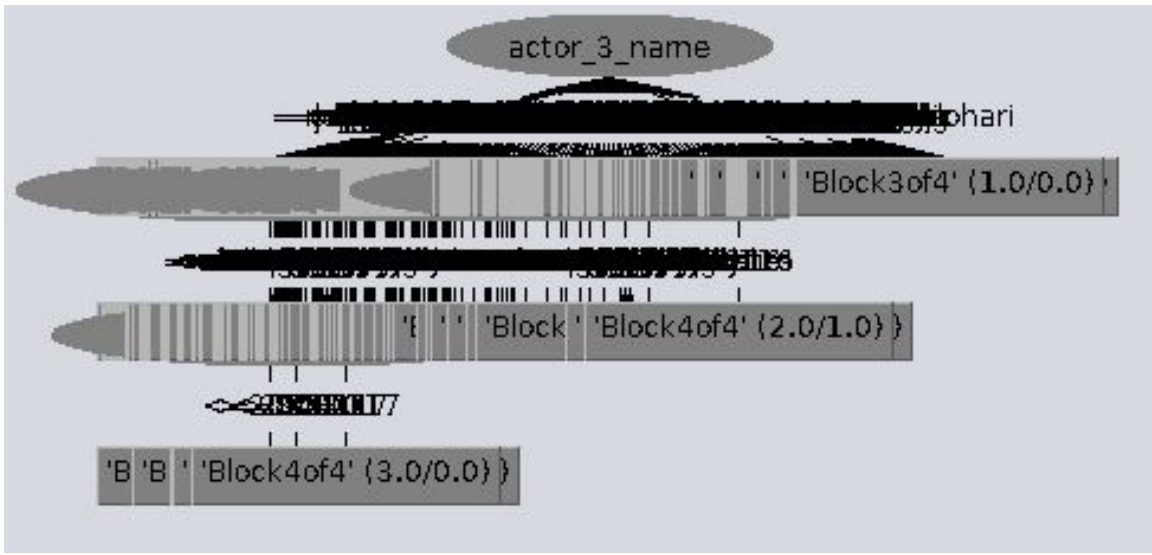


J48:

J48 résultat:

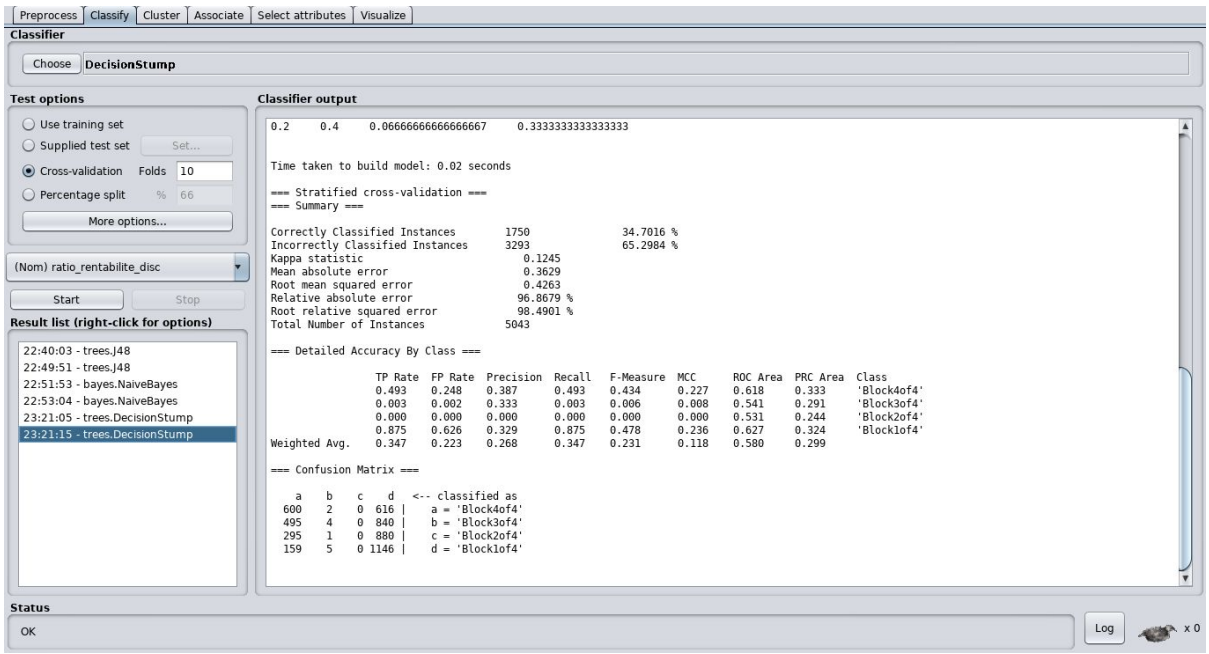


J48 arbre:



DecisionStump:

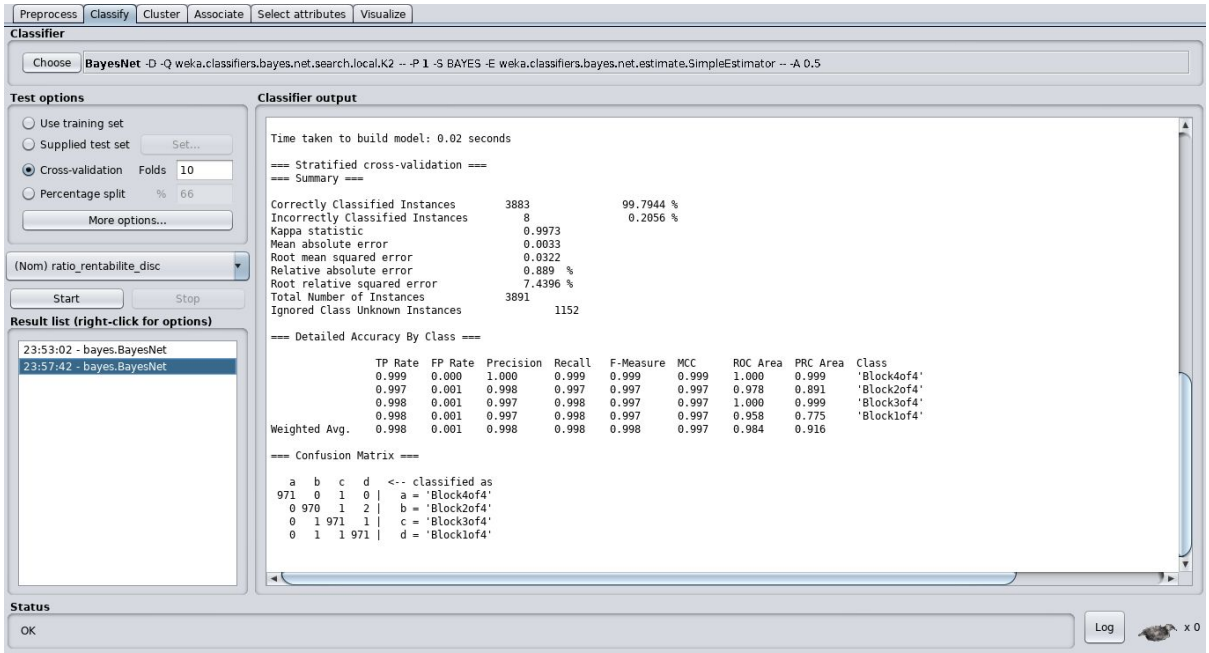
DecisionStump résultat:



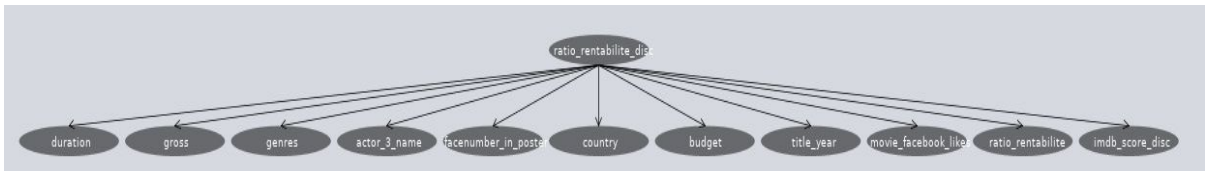
Score rentabilité:

BayesNet:

BayesNet résultat

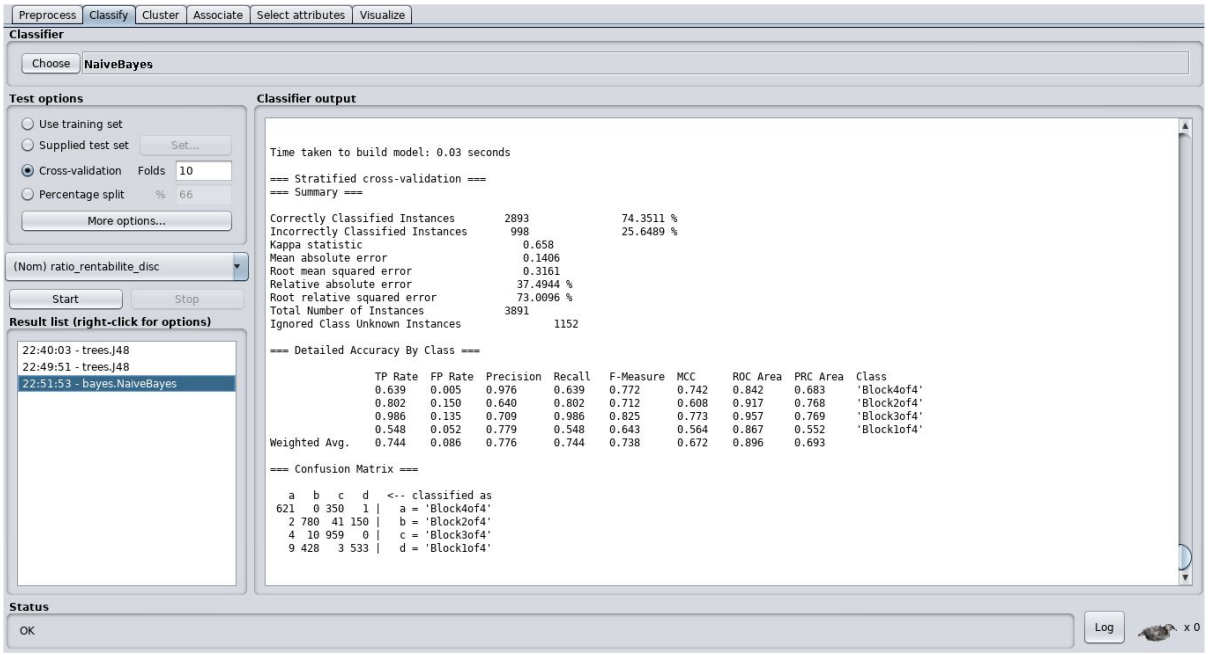


BayesNet Graphe:

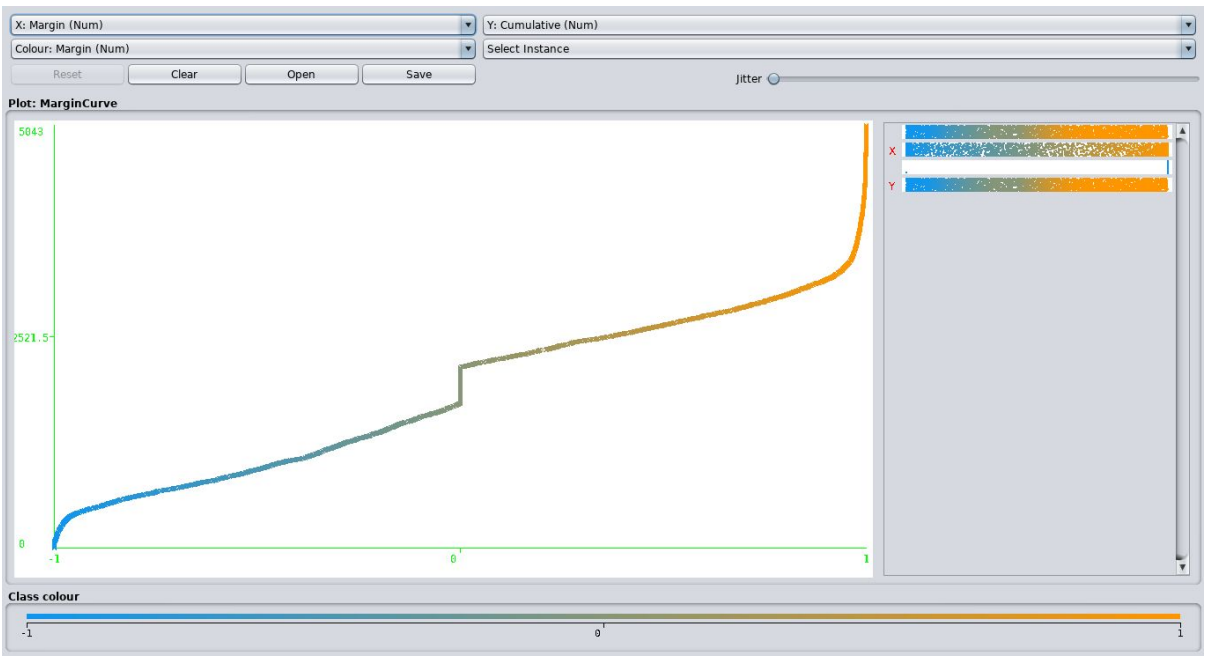


NaiveBayes:

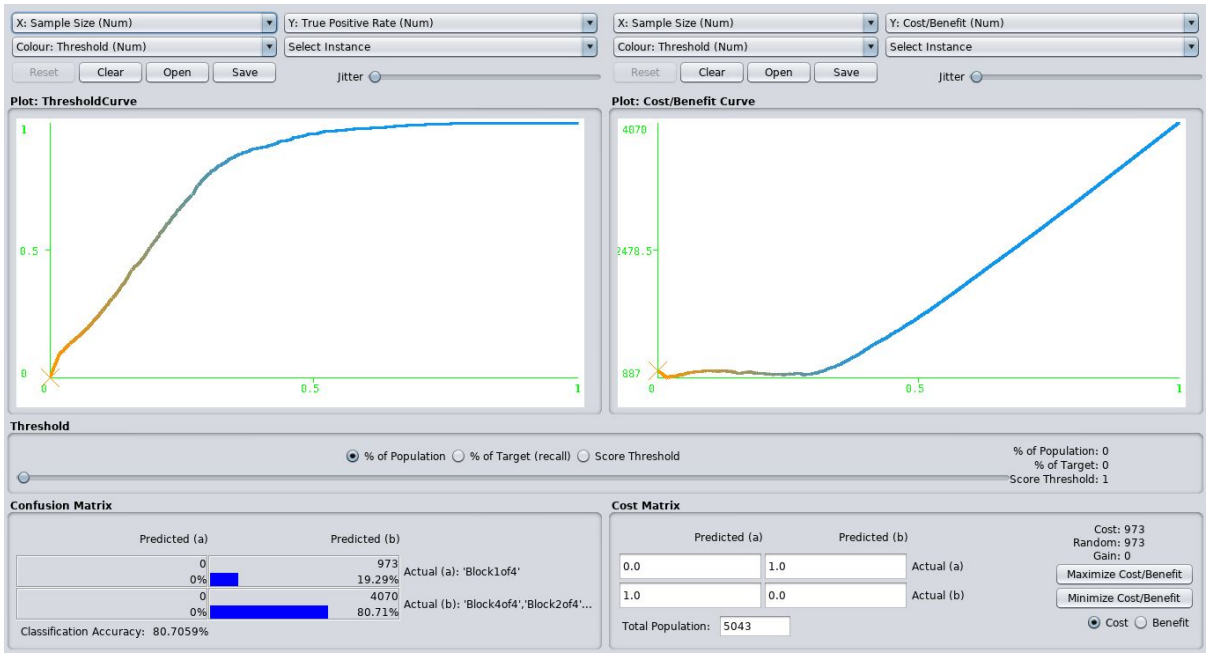
NaiveBayes résultat:



NaiveBayes curve:

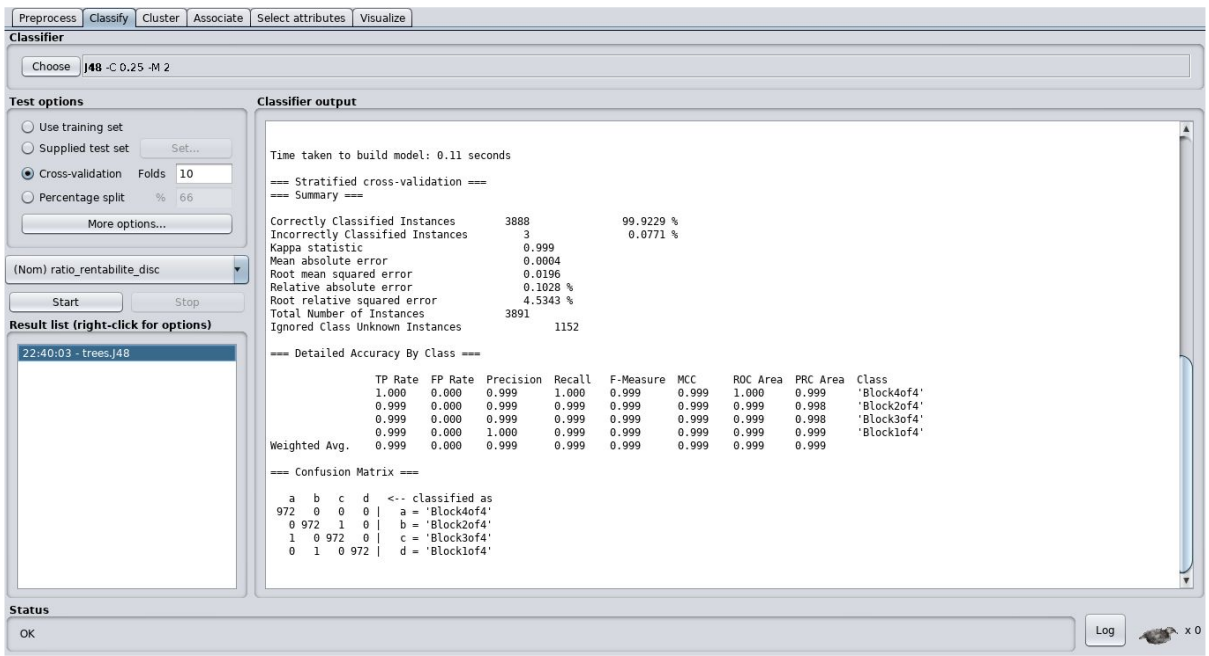


NaiveBayes coût / bénéfice:

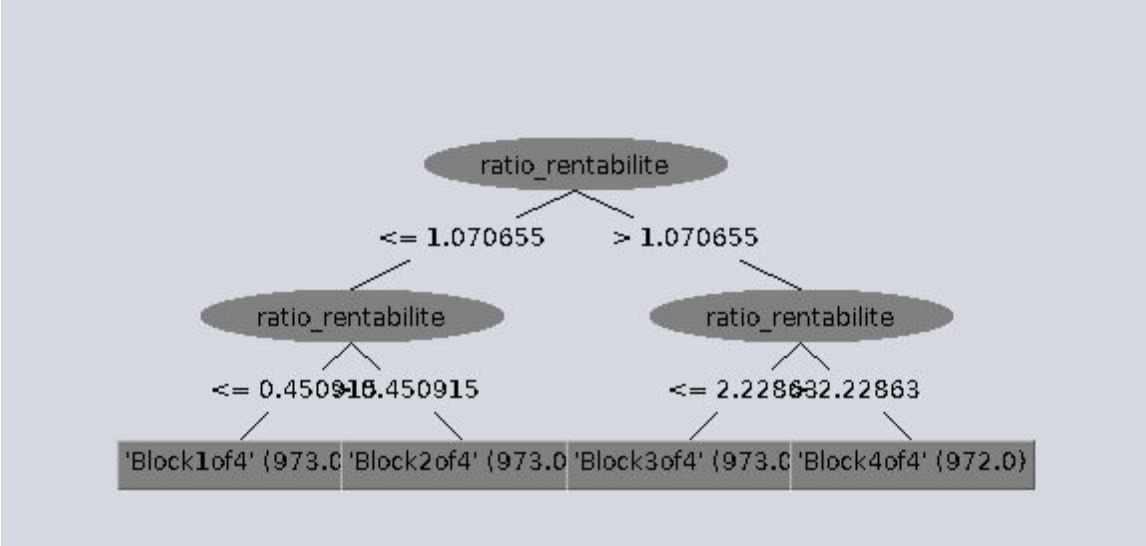


J48:

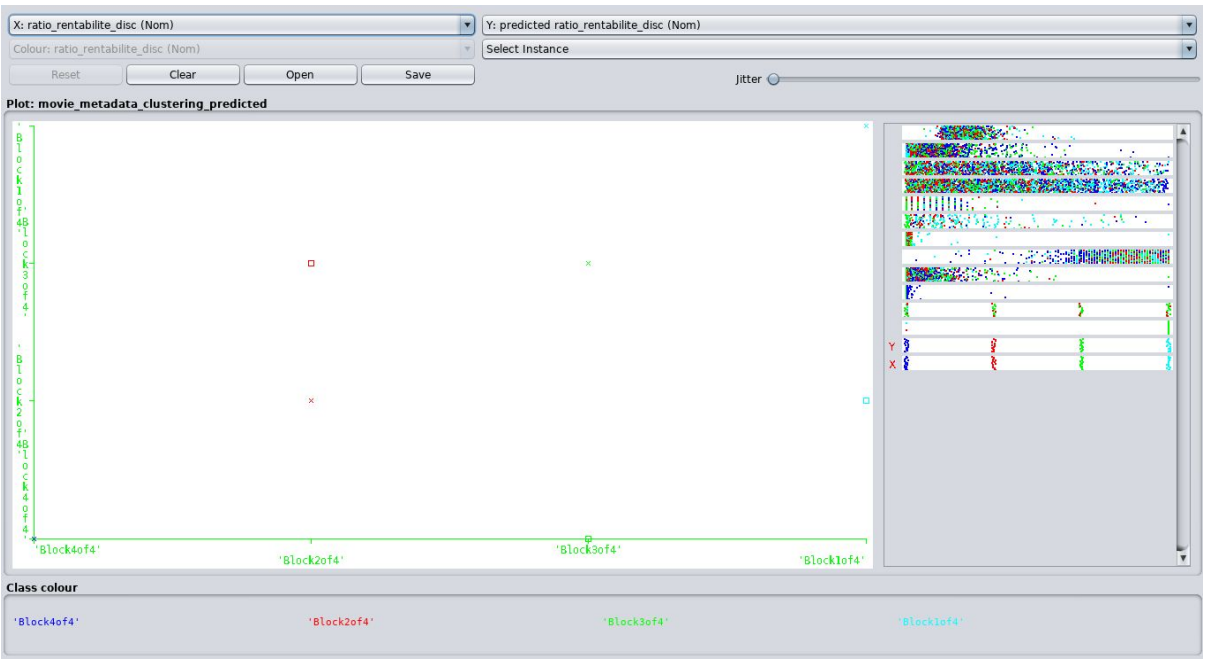
J48 résultat:



J48 arbre:

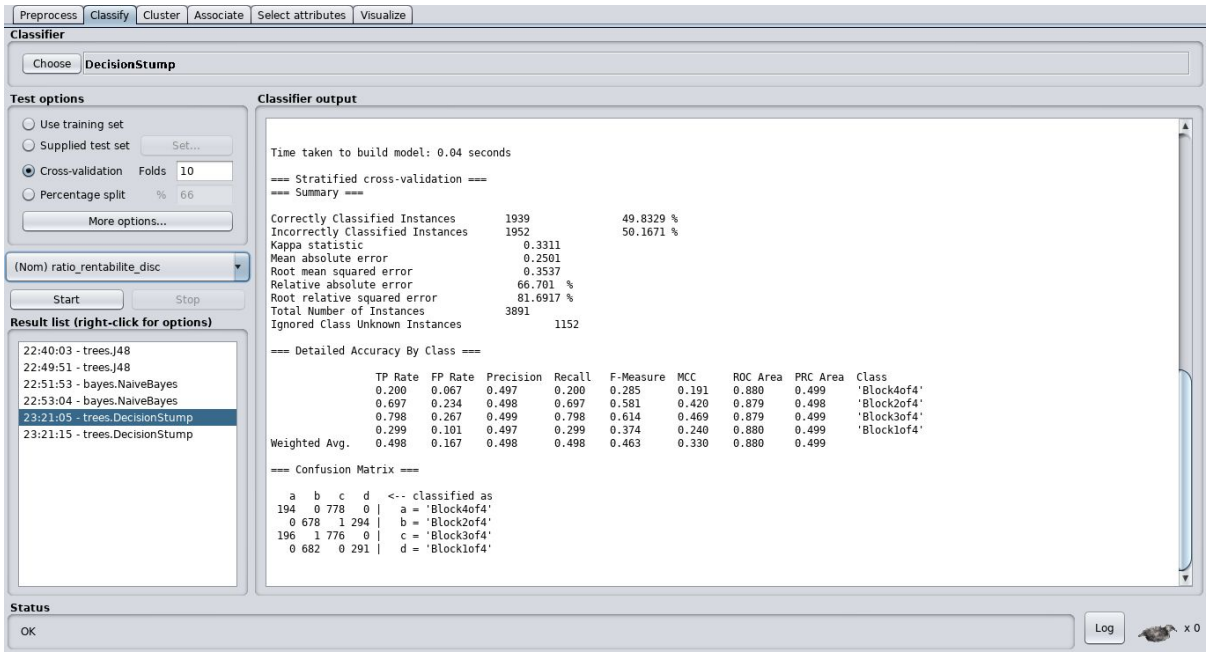


J48 erreurs:

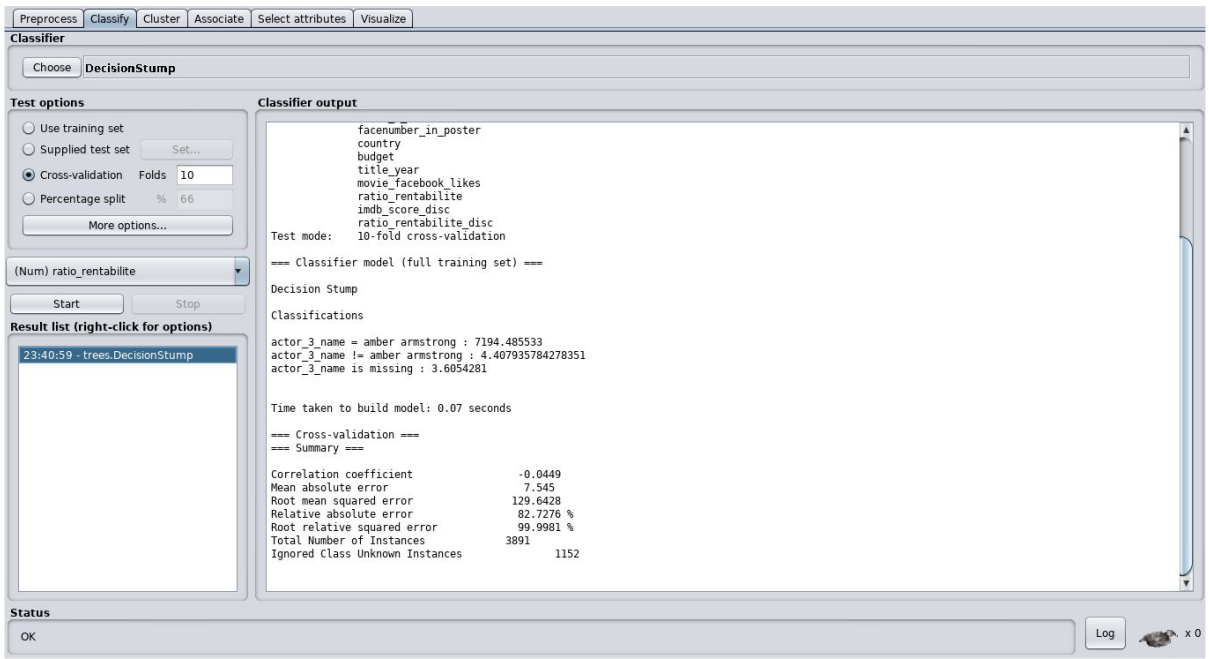


DecisionStump:

DecisionStump résultat:



DecisionStump rentabilité:



DecisionStump rentabilité error:

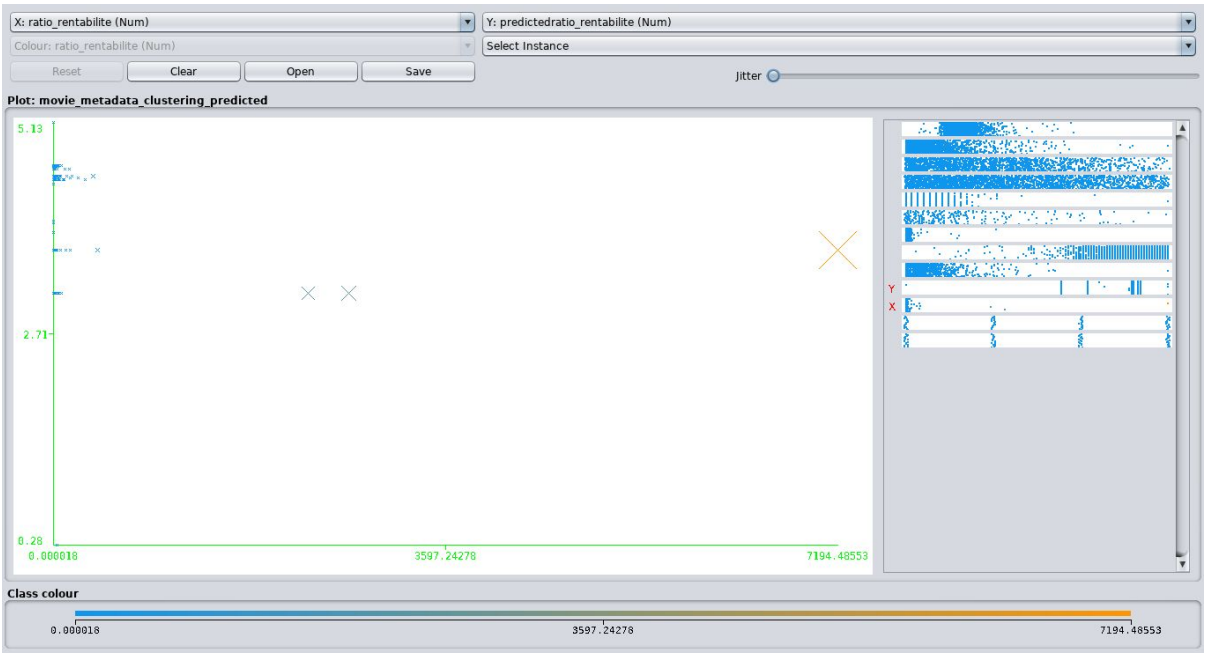


Tableau score IMDB:

	Temps de création du modèle		Taux d'erreur		Meilleur F-mesure	
	Données discrétisés	Données non discrétisés	Données discrétisés	Données non discrétisés	Données discrétisés	Données non discrétisés
Bayes Net	0.02 sec	Non fonctionnel	0.2056%	Non fonctionnel	0.999	Non fonctionnel
Naive Bayes	0.01 sec	Non fonctionnel	61.9671%	Non fonctionnel	0.511	Non fonctionnel
J48	0.13 sec	Non fonctionnel	67.678%	Non fonctionnel	0.402	Non fonctionnel
DecisionStump	0.02 sec	0 sec	65.2984%	Pas de Précision par classe	0.875	Pas de Précision par classe

Tableau rentabilité:

	Temps de création du modèle		Taux d'erreur		Meilleur F-mesure	
	Données discrétisés	Données non discrétisés	Données discrétisés	Données non discrétisés	Données discrétisés	Données non discrétisés
Bayes Net	0.12 sec	Non fonctionnel	0.2056%	Non fonctionnel	0.999	Non fonctionnel
Naive Bayes	0.03 sec	Non fonctionnel	25.6489%	Non fonctionnel	0.825	Non fonctionnel
J48	0.11 sec	Non fonctionnel	0.0771%	Non fonctionnel	0.999	Non fonctionnel
DecisionStump	0.04 sec	0.07 sec	50.1671%	Pas de Précision par classe	0.798	

Conclusion

Après analyse du tableau comparatif des résultats obtenus en appliquant chacun des algorithmes présents dans ce dernier.

Le meilleur algorithme pour prédire le **score IMBD** est :
BayesNet car il fournit la meilleur valeur de F-mesure (qui est donc une moyenne entre Precision et Recall) qui est 0.999.Il fournit le plus bas taux d'erreur avec un temps de création de modèle acceptable.

Le meilleur algorithme pour prédire le **ratio de rentabilité** est:
J48 car il fournit la meilleur valeur de F-mesure (qui est donc une moyenne entre Precision et Recall) qui est 0.999.Il fournit le plus bas taux d'erreur avec un temps de création de modèle acceptable.

On peut aussi choisir BayesNet car il est très proche en terme de performance de J48.