Projet de Bio-informatique
M1 IDIAG

ELABORE PAR :
ARFAOUI YOUSSEF
SECRAFI MOUNIR

A.U:2017/2018

Sommaire

Introduction

La bio-informatique est un champ de recherche multi-disciplinaire où travaillent de concert biologistes, médecins, informaticiens, mathématiciens, physiciens et bio-informaticiens, dans le but de résoudre un problème scientifique posé par la biologie. Le spécialiste qui travaille à mi-chemin entre ces sciences et l'informatique est appelé bio-informaticien ou bionaute.

Depuis quelques années les progrès de l'informatique (et en particulier de la bio-informatique mise au service de la biodiversité ou « Biodiversity informatics » pour les anglophones) dopent la biologie évolutive en offrant aux chercheurs un accès à un nombre croissant de données sur la diversité et les variations des gènes, ainsi que des génomes, des organismes et de l'environnement en général.

Tout cela peut être relié à la phylogénie (de l'étude des populations à celle de clades entiers), via de nouveaux protocoles et réseaux dans le domaine de l'informatique de la biodiversité

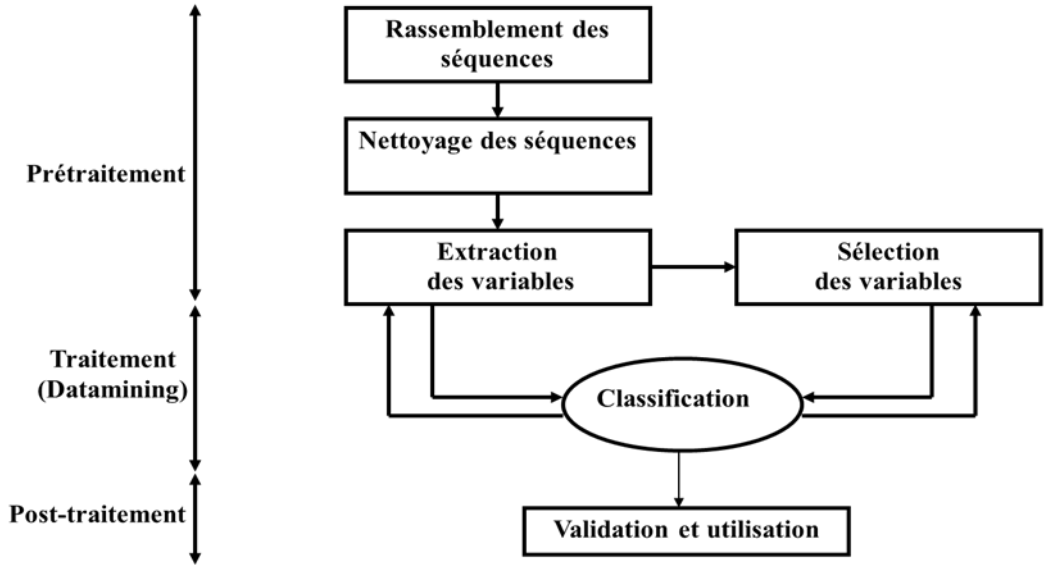
Le terme bio-informatique peut également décrire (par abus de langage) toutes les applications informatiques résultant de ces recherches Note 1. Plus généralement, la bio-informatique est l'application de la statistique et de l'informatique à la science biologique.

L'utilisation du terme bio-informatique est documentée pour la première fois en 1970 dans une publication1 de Paulien Hogeweg et Ben Hesper (université d'Utrecht, Pays-Bas), en référence à l'étude des processus d'information dans les systèmes biotiques.

Cela va de l'analyse du génome à la modélisation de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'image, l'assemblage de génome et la reconstruction d'arbres phylogénétiques (phylogénie). Cette discipline constitue la « biologie in silico », par analogie avec in vitro ou in vivo.

Processus général de la classification des protéines

Le projet consiste à classifier des séquences de protéines en se basant sur leurs structures primaires. La figure suivante explique les étapes du projet.



Processus général de la classification des protéines

Prétraitement

1. Extraction n-grammes

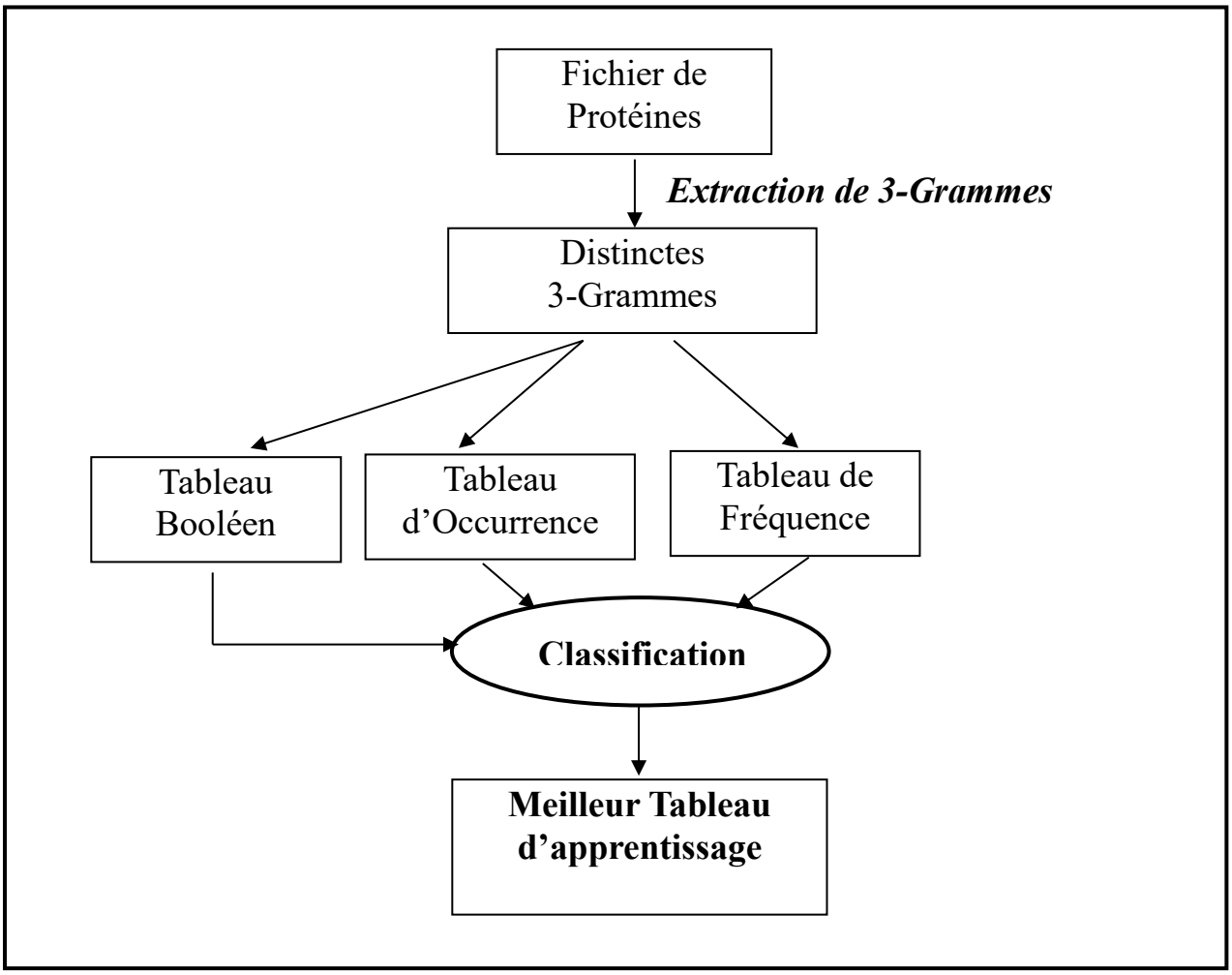
Un n-gramme est une sous-séquence de n éléments construite à partir d'une séquence donnée. L'idée semble provenir des travaux de Claude Shannon en théorie de l'information. Son idée était que, à partir d'une séquence de lettres donnée (par exemple « par exemple ») il est possible d'obtenir la fonction de vraisemblance de l'apparition de la lettre suivante. À partir d'un corpus d'apprentissage, il est facile de construire une distribution de probabilité pour la prochaine lettre avec un historique de taille n. Cette modélisation correspond en fait à un modèle de Markov d'ordre n où seules les n dernières observations sont utilisées pour la prédiction de la lettre suivante. Ainsi un bi gramme est un modèle de Markov d'ordre 2.

Exemple (K=3)

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
```

ACA	CAA	AAG	AGA	GAT	ATG	TGC	GCC	CCA	CAT	ATT	TTG	TGT	GTC	TCC	CCC	CCG	CGG	GGC	CCT

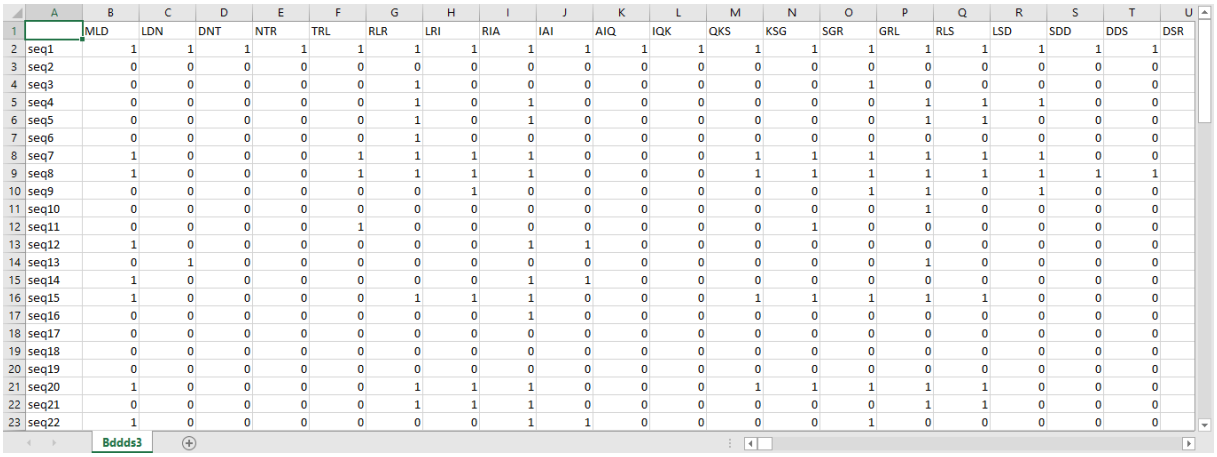
2. Construction du Tableau d'apprentissage (Booléen, Occurrence, Fréquence)



- **Fichier protéine**

Exemple N-grammes=3

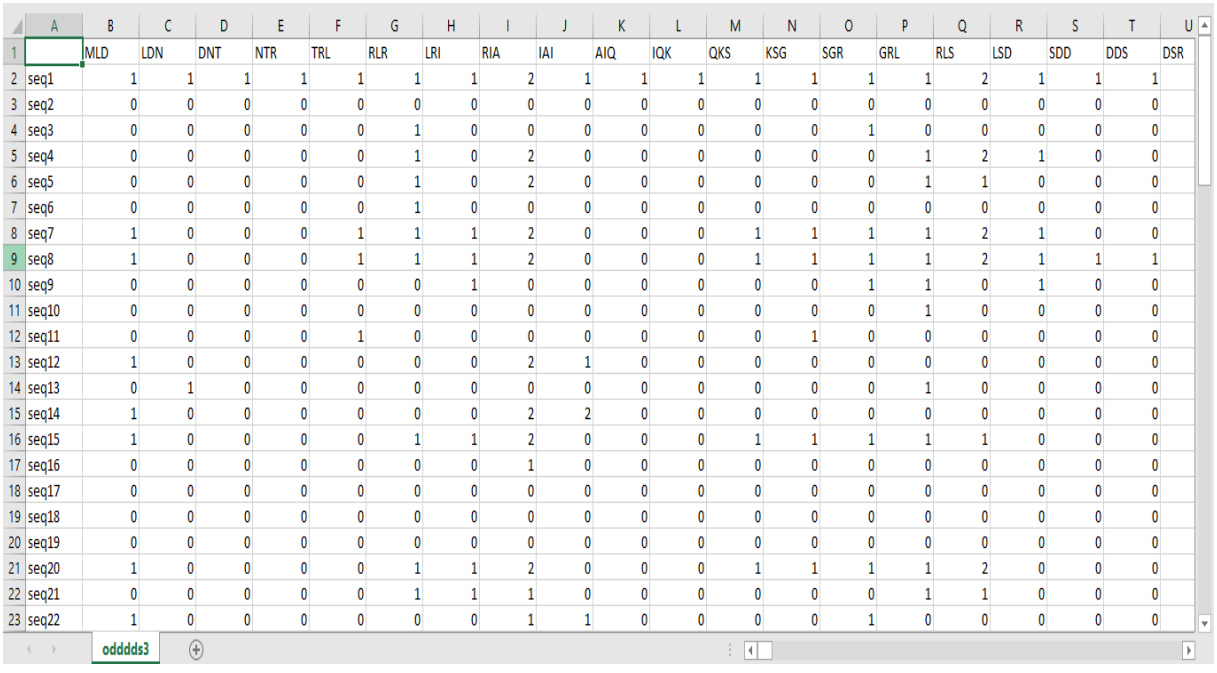
- Tableau Boolean **(**Ligne= **Ensemble de sequences**, Colonne =**variable(3-grammes))**



		MLD	LDN	DNT	NTR	TRL	RLR	LRI	RIA	IAI	AIQ	IQK	QKS	KSG	SGR	GRL	RLS	LSD	SDD	DDS	DSR
2	seq1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
3	seq2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	seq3	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	
5	seq4	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	1	0	0	
6	seq5	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	
7	seq6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	seq7	1	0	0	0	1	1	1	1	0	0	0	1	1	1	1	1	1	0	0	
9	seq8	1	0	0	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	
10	seq9	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1	0	0	
11	seq10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
12	seq11	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
13	seq12	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	
14	seq13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
15	seq14	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	
16	seq15	1	0	0	0	0	1	1	1	0	0	0	1	1	1	1	1	0	0	0	
17	seq16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
18	seq17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	seq18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	seq19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	seq20	1	0	0	0	0	1	1	1	0	0	0	1	1	1	1	1	0	0	0	
22	seq21	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0	
23	seq22	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	

Bddds3

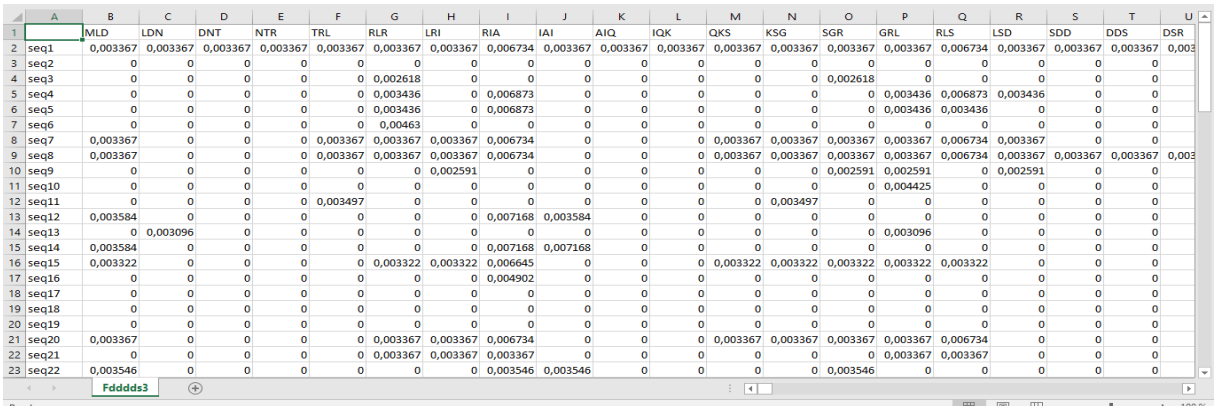
- Tableau Occurrence



		MLD	LDN	DNT	NTR	TRL	RLR	LRI	RIA	IAI	AIQ	IQK	QKS	KSG	SGR	GRL	RLS	LSD	SDD	DDS	DSR
2	seq1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	
3	seq2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	seq3	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	
5	seq4	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	2	1	0	0	
6	seq5	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	1	0	0	0	
7	seq6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	seq7	1	0	0	0	1	1	1	2	0	0	0	1	1	1	1	2	1	0	0	
9	seq8	1	0	0	0	1	1	1	2	0	0	0	1	1	1	1	2	1	1	1	
10	seq9	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1	0	0	
11	seq10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
12	seq11	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
13	seq12	1	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	
14	seq13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
15	seq14	1	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	
16	seq15	1	0	0	0	0	1	1	2	0	0	0	1	1	1	1	1	0	0	0	
17	seq16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
18	seq17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	seq18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	seq19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	seq20	1	0	0	0	0	1	1	2	0	0	0	1	1	1	1	2	0	0	0	
22	seq21	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0	
23	seq22	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	

odddds3

➢ Tableau fréquence



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		MLD	LDN	DNT	NTR	TRL	RLR	LRI	RIA	IAI	AIQ	IQK	QKS	KSG	SGR	GRL	RLS	LSD	SDD	DDS	DSR
2	seq1	0,003367	0,003367	0,003367	0,003367	0,003367	0,003367	0,003367	0,006734	0,003367	0,003367	0,003367	0,003367	0,003367	0,003367	0,003367	0,006734	0,003367	0,003367	0,003367	0,003
3	seq2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	seq3	0	0	0	0	0	0,002618	0	0	0	0	0	0	0	0,002618	0	0	0	0	0	
5	seq4	0	0	0	0	0	0,003436	0	0,006873	0	0	0	0	0	0	0,003436	0,006873	0,003436	0	0	
6	seq5	0	0	0	0	0	0,003436	0	0,006873	0	0	0	0	0	0	0,003436	0,003436	0	0	0	
7	seq6	0	0	0	0	0	0,00463	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	seq7	0,003367	0	0	0	0,003367	0,003367	0,003367	0,006734	0	0	0	0,003367	0,003367	0,003367	0,003367	0,006734	0,003367	0	0	
9	seq8	0,003367	0	0	0	0,003367	0,003367	0,003367	0,006734	0	0	0	0,003367	0,003367	0,003367	0,003367	0,006734	0,003367	0,003367	0,003367	0,003
10	seq9	0	0	0	0	0	0	0,002591	0	0	0	0	0	0	0,002591	0,002591	0	0,002591	0	0	
11	seq10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,004425	0	0	0	0	
12	seq11	0	0	0	0	0,003497	0	0	0	0	0	0	0	0,003497	0	0	0	0	0	0	
13	seq12	0,003584	0	0	0	0	0	0	0,007168	0,003584	0	0	0	0	0	0	0	0	0	0	
14	seq13	0	0,003096	0	0	0	0	0	0	0	0	0	0	0	0	0,003096	0	0	0	0	
15	seq14	0,003584	0	0	0	0	0	0	0,007168	0,007168	0	0	0	0	0	0	0	0	0	0	
16	seq15	0,003322	0	0	0	0	0,003322	0,003322	0,006645	0	0	0	0,003322	0,003322	0,003322	0,003322	0,003322	0	0	0	
17	seq16	0	0	0	0	0	0	0	0,004902	0	0	0	0	0	0	0	0	0	0	0	
18	seq17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	seq18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	seq19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	seq20	0,003367	0	0	0	0	0,003367	0,003367	0,006734	0	0	0	0,003367	0,003367	0,003367	0,003367	0,006734	0	0	0	
22	seq21	0	0	0	0	0	0,003367	0,003367	0,003367	0	0	0	0	0	0	0,003367	0,003367	0	0	0	
23	seq22	0,003546	0	0	0	0	0	0	0,003546	0,003546	0	0	0	0	0,003546	0	0	0	0	0	

Fdddds3

Protéines pairs	**2- grammes**	**3- grammes**	**2-3- grammes**	**4-grammes**	**[3-4]-grammes**	**5-grammes**
F1_2	400	6600	7000	23408	30008	28958

Tableau 1.Nombre de descripteurs en fonction de la longueur n de n-grammes

Traitement Datamining

➢ Data-Mining

Terme récent (1995) représentant un mélange d'idées et d'outils provenant de la Statistique, Science de l'information et l'Informatique.

• A évolué vers le data science

• Machine Learning,

• Big Data (explosion des données),

• Formalismes de stockage et de traitement distribués de données (NoSQL, NewSQL, Hadoop, MapReduce ...).

➢ Classification

La classification est la tâche la plus commune de la fouille de données qui semble être une tâche humaine primordiale.

Afin de comprendre notre vie quotidienne, nous sommes constamment obligés à classer, catégoriser et évaluer.

La classification consiste à étudier les caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie.

Le fonctionnement de la classification se décompose en deux phases :

➢ **La première étant la phase d'apprentissage** :

Dans cette phase, les approches de classification utilisent un jeu d'apprentissage dans lequel tous les objets sont déjà associés aux classes de références connues.

L'algorithme de classification apprend du jeu d'apprentissage et construit un modèle.

➢ **La seconde phase est la phase de classification :**

Proprement dite, dans laquelle le modèle appris est employé pour classifier de nouveaux objets.

➢ Cross validation

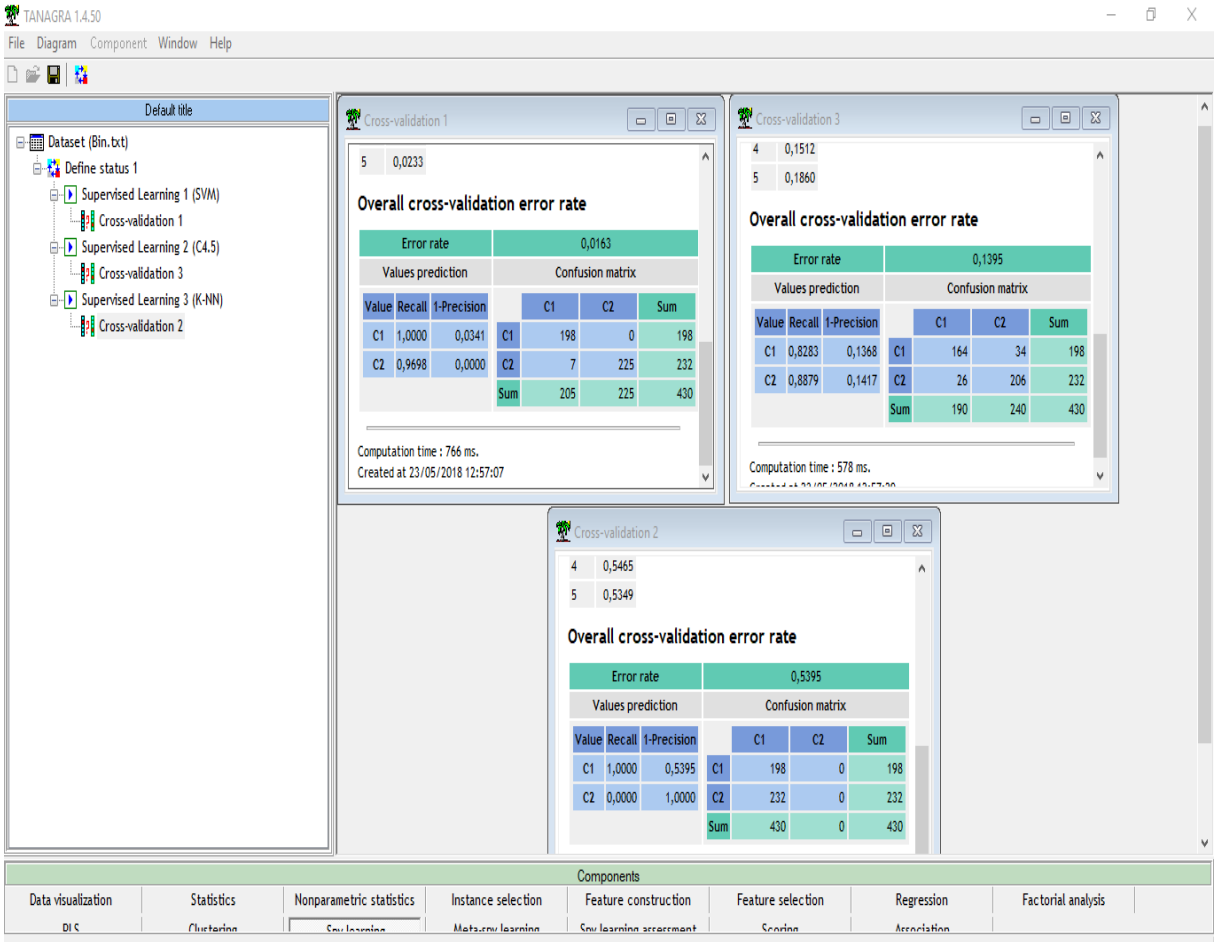
Dans l'apprentissage supervisé, il est généralement accepté de ne pas utiliser le même échantillon pour construire un modèle prédictif et estimer son taux d'erreur. L'erreur obtenue dans ces conditions - appelée taux d'erreur de resubstitution - est (très souvent) trop optimiste, laissant croire que le modèle présentera une excellente performance de prédiction.

Une approche typique consiste à diviser les données en deux parties (approche de retenue): un premier échantillon, le dit échantillon de train est utilisé pour construire le modèle; un second échantillon, dit échantillon d'essai, est utilisé pour mesurer sa performance. Le taux d'erreur mesuré reflète honnêtement le comportement du modèle en généralisation. Malheureusement, sur de petits ensembles de données, cette approche est problématique. En réduisant la quantité de données présentées à l'algorithme d'apprentissage, nous ne pouvons pas apprendre correctement la relation sous-jacente entre les descripteurs et l'attribut de classe. Dans le même temps, la partie consacrée aux tests reste limitée, l'erreur mesurée a une variance élevée.

Nous calculons le taux d'erreur en utilisant un ressemblant méthode. Nous avons choisi une validation croisée 5 X 2 que nous répéter un grand nombre de fois. Cette approche est privilégiée car cela permet d'obtenir des biais et des résultats relativement stables.
Dans notre contexte, de fort risque de sur apprentissage, nous cherchons à produire des résultats généraux et à éviter être trop dépendant de notre ensemble de données.

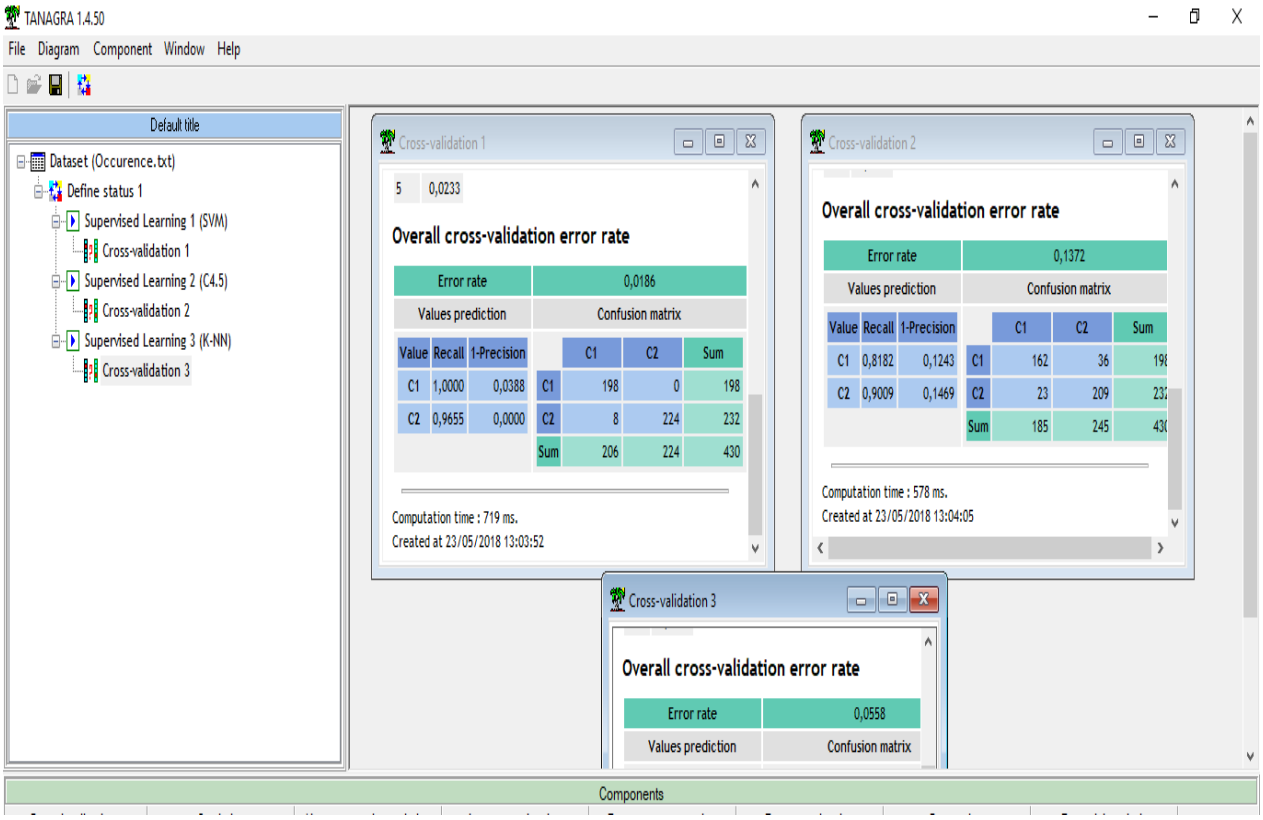
Application des techniques de datamining (SVM, KNN, C4.5) avec Tanagra

- Tableau Booléen(2-grammes)



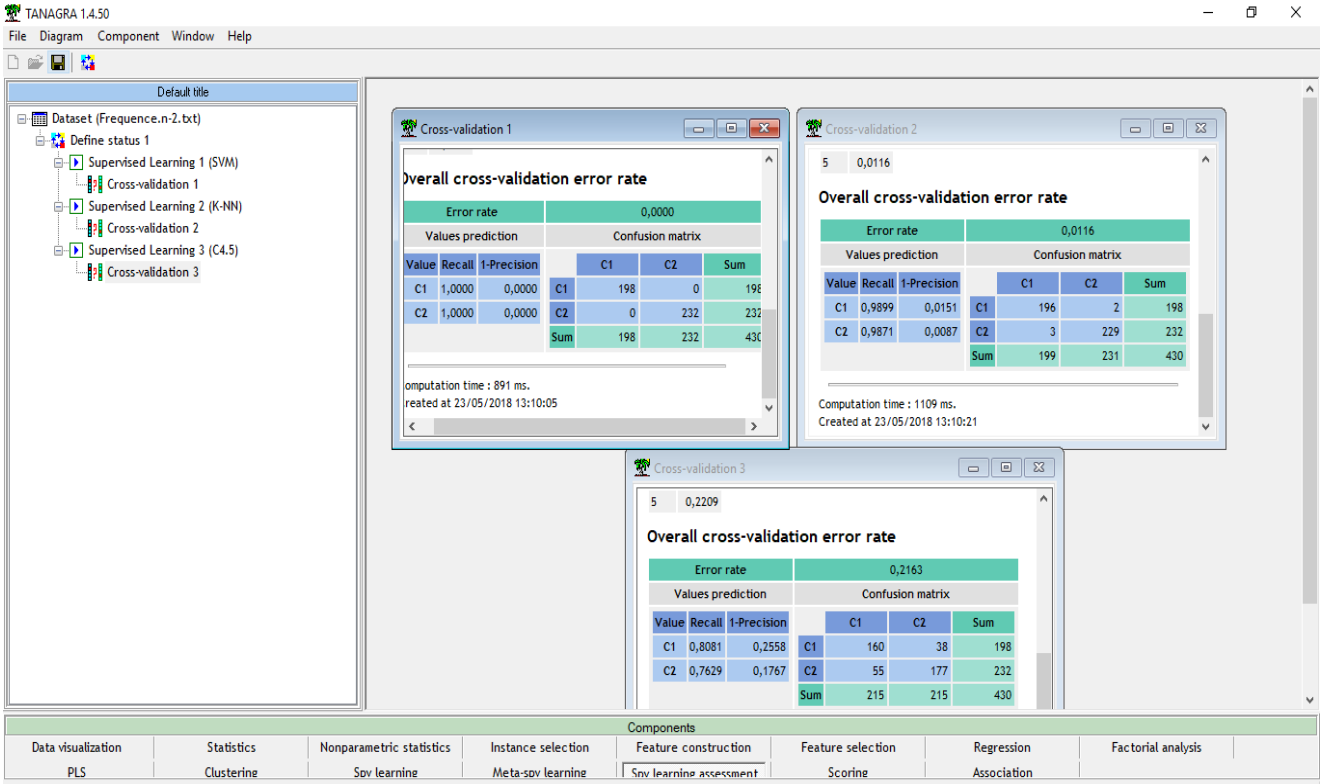
Algorithme de classification	*Taux d'erreur*
Knn	0.539
C4.5	0.139
SVM	0.016

➢ Tableau Occurrence(2-grammes)



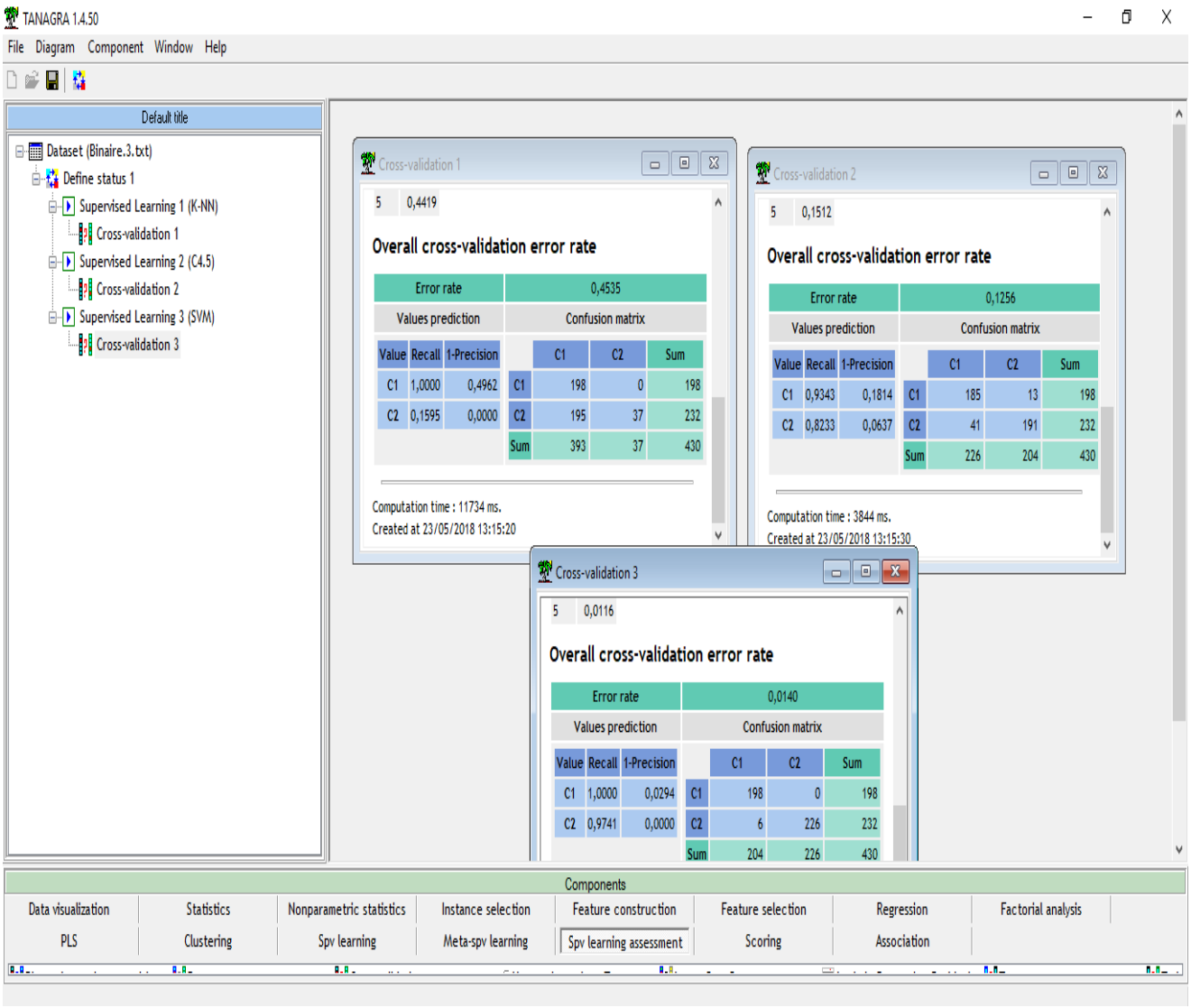
Algorithme de classification	*Taux d'erreur*
Knn	0.0558
C4.5	0.137
SVM	0.018

➢ Tableau Fréquences(2-grammes)



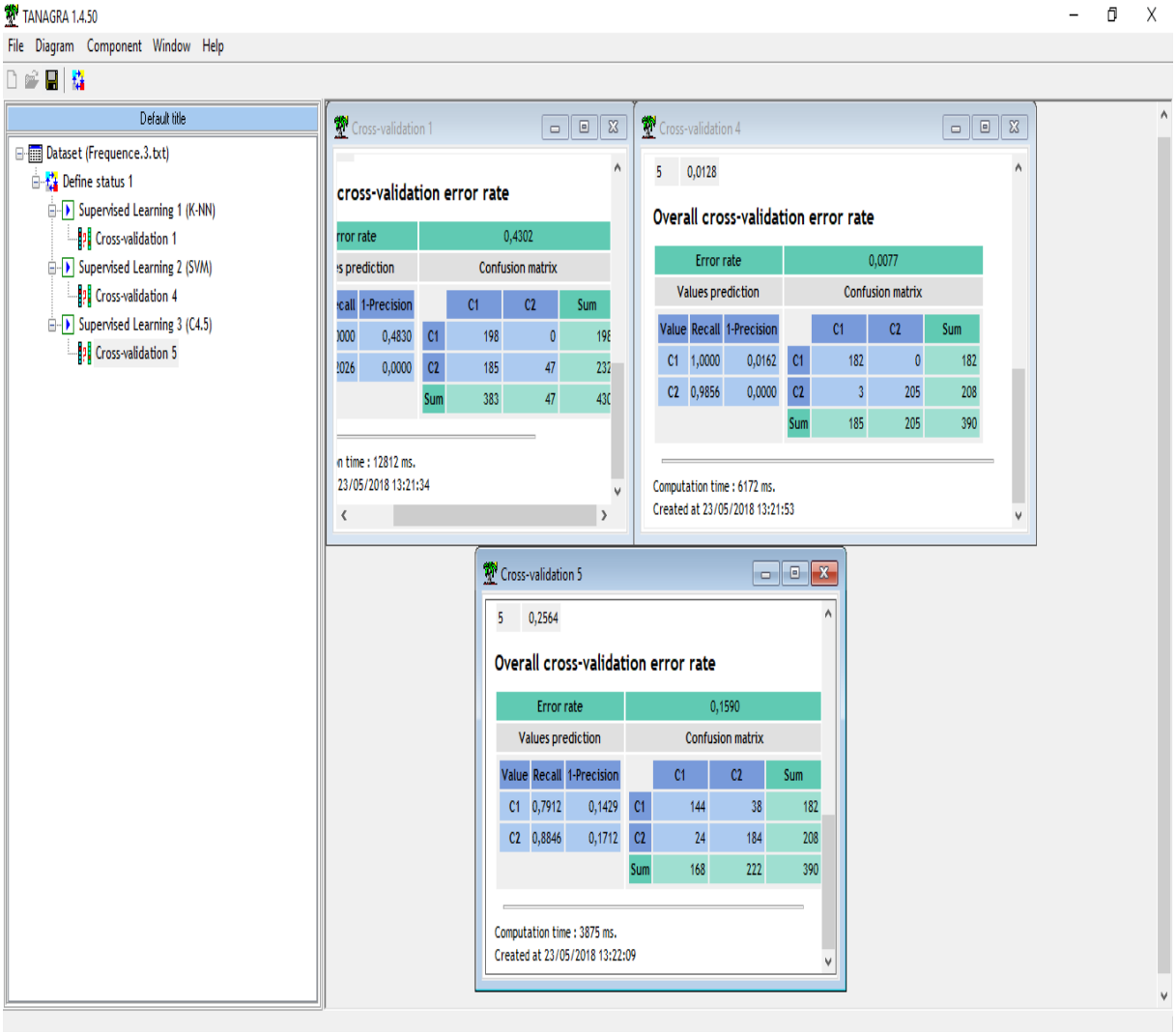
Algorithme de classification	*Taux d'erreur*
Knn	0.0116
C4.5	0.2163
SVM	0.00

➢ Tableau Booléen(3-grammes)



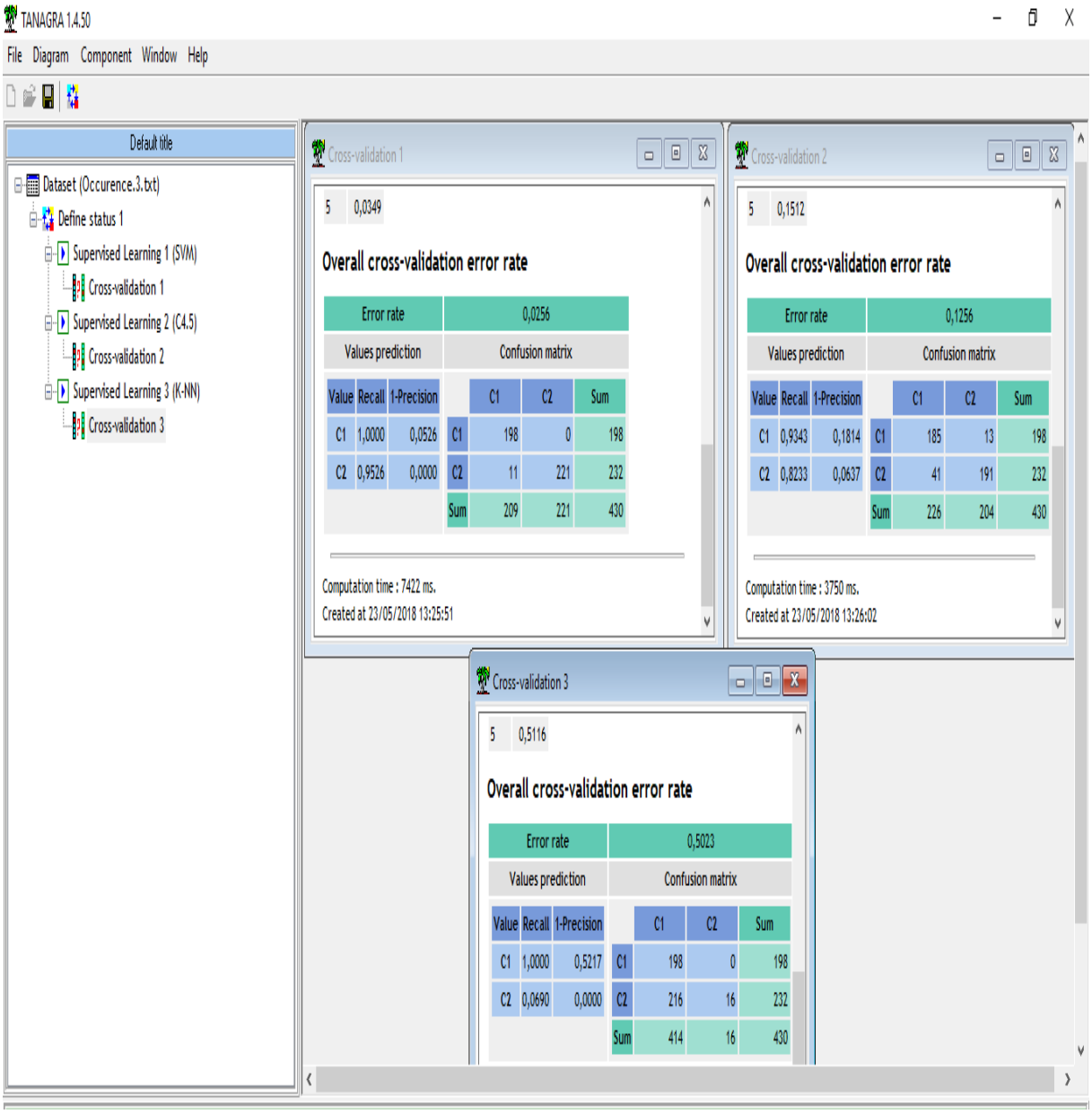
Algorithme de classification	*Taux d'erreur*
Knn	0.453
C4.5	0.125
SVM	0.014

➢ Tableau Fréquence(3-grammes)



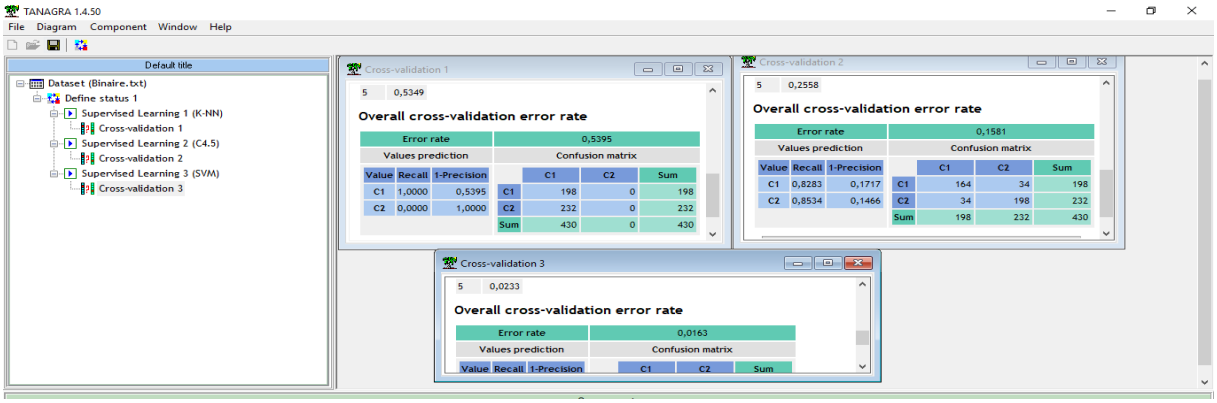
Algorithme de classification	*Taux d'erreur*
Knn	0.430
C4.5	0.159
SVM	0.007

➢ Tableau Occurrences(3-grammes)



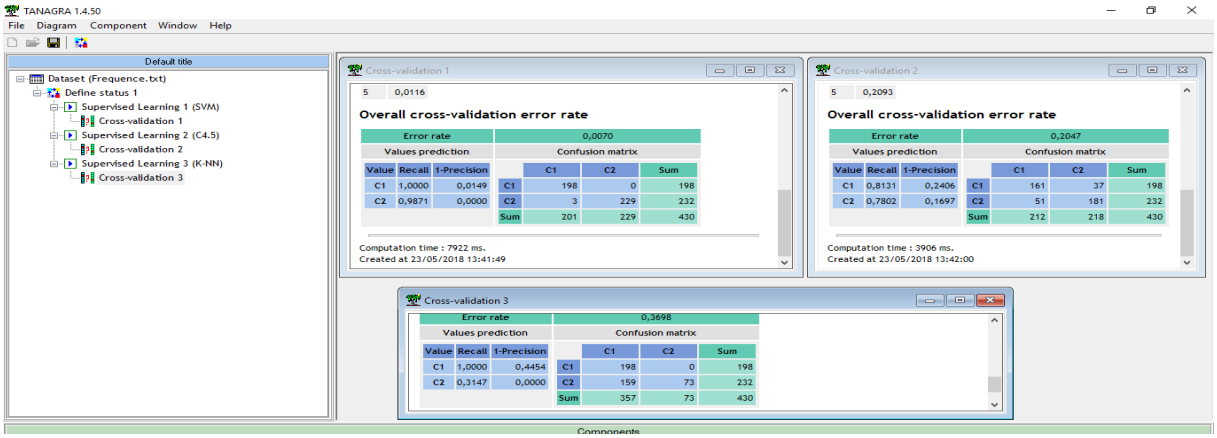
Algorithme de classification	*Taux d'erreur*
Knn	0.5023
C4.5	0.1256
SVM	0.0256

➢ Tableau Booléen([2-3]-grammes)



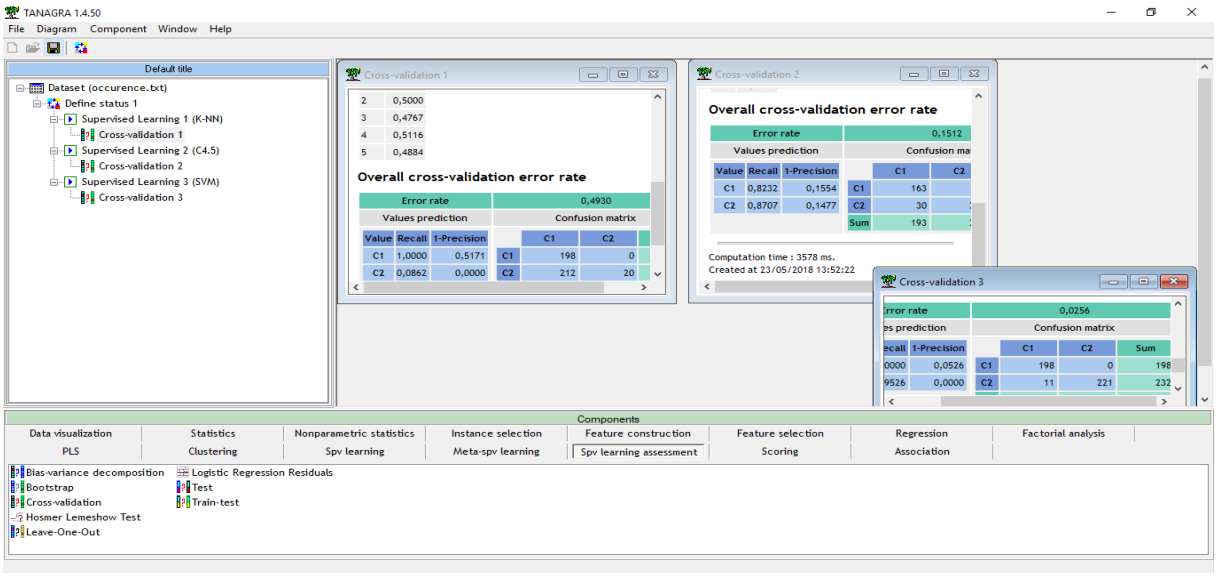
Algorithme de classification	*Taux d'erreur*
Knn	0.539
C4.5	0.158
SVM	0.016

➢ Tableau fréquences ([2-3]-grammes)



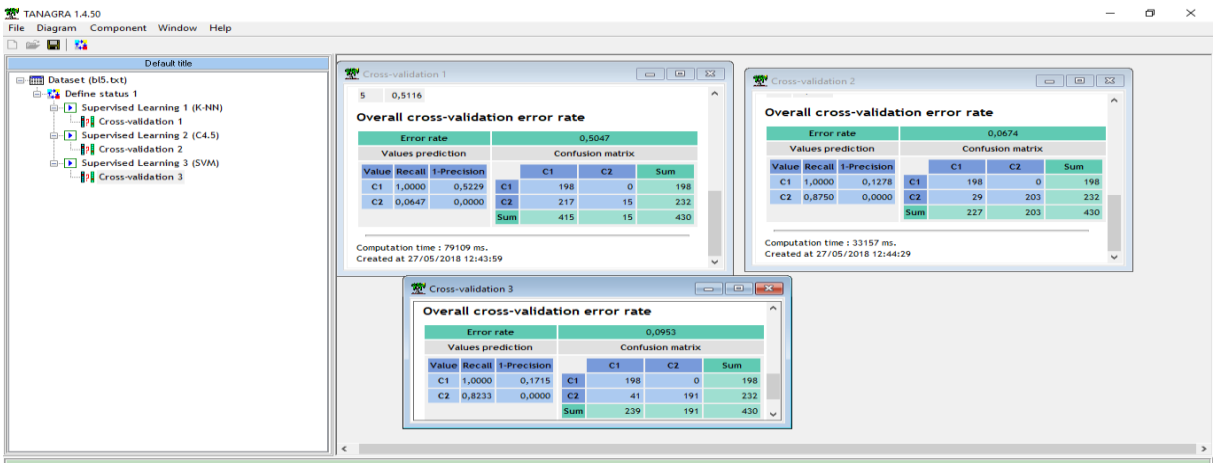
Algorithme de classification	*Taux d'erreur*
Knn	0.369
C4.5	0.204
SVM	0.007

➢ Tableau occurrences([2-3]-grammes)



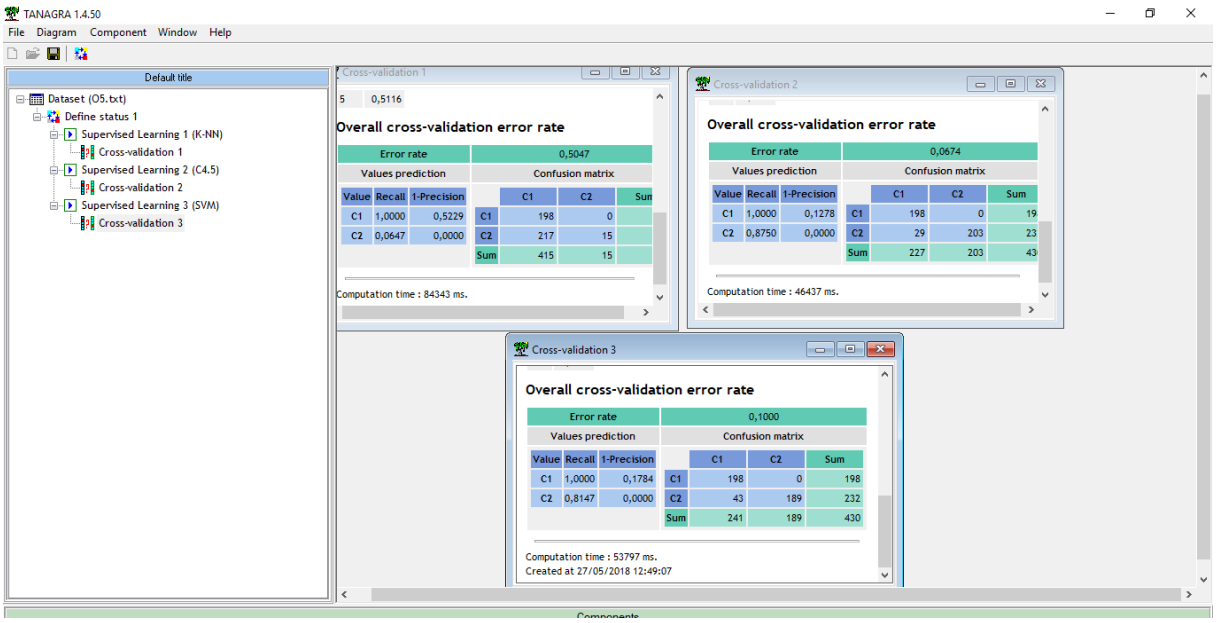
Algorithme de classification	*Taux d'erreur*
Knn	0.493
C4.5	0.151
SVM	0.025

➢ Tableau Booléen (4-grammes)



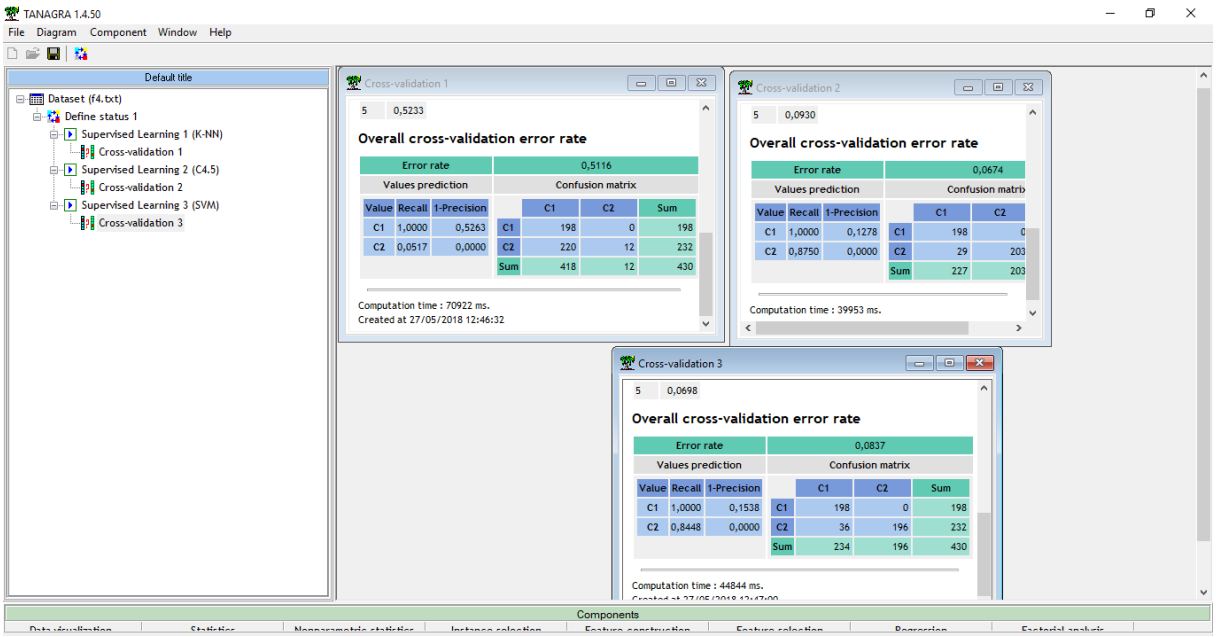
Algorithme de classification	*Taux d'erreur*
Knn	0.504
C4.5	0.064
SVM	0.095

- Tableau Occurrences (4-grammes)



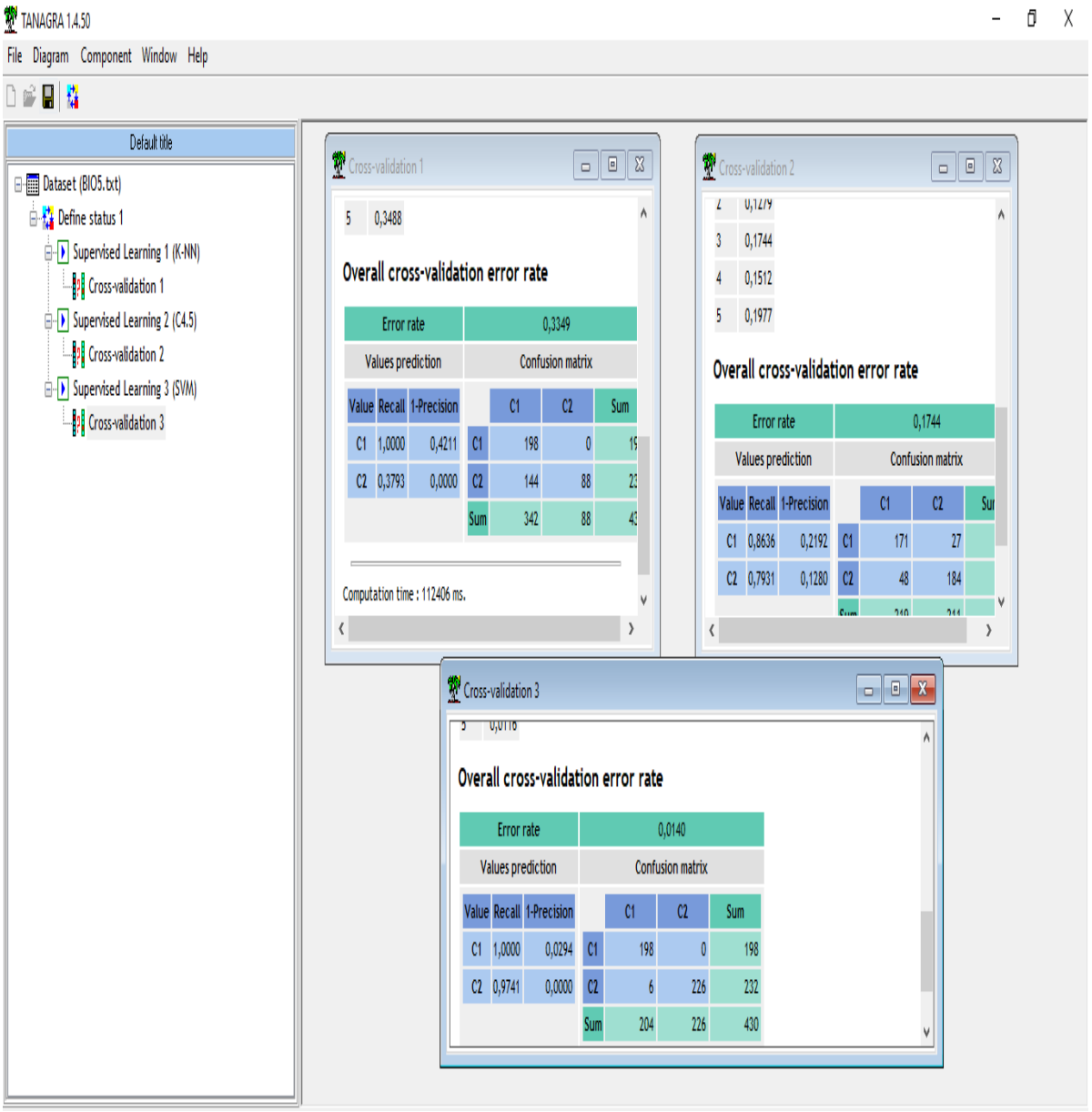
Algorithme de classification	*Taux d'erreur*
Knn	0.504
C4.5	0.067
SVM	0.100

➢ Tableau Fréquences(4-grammes)



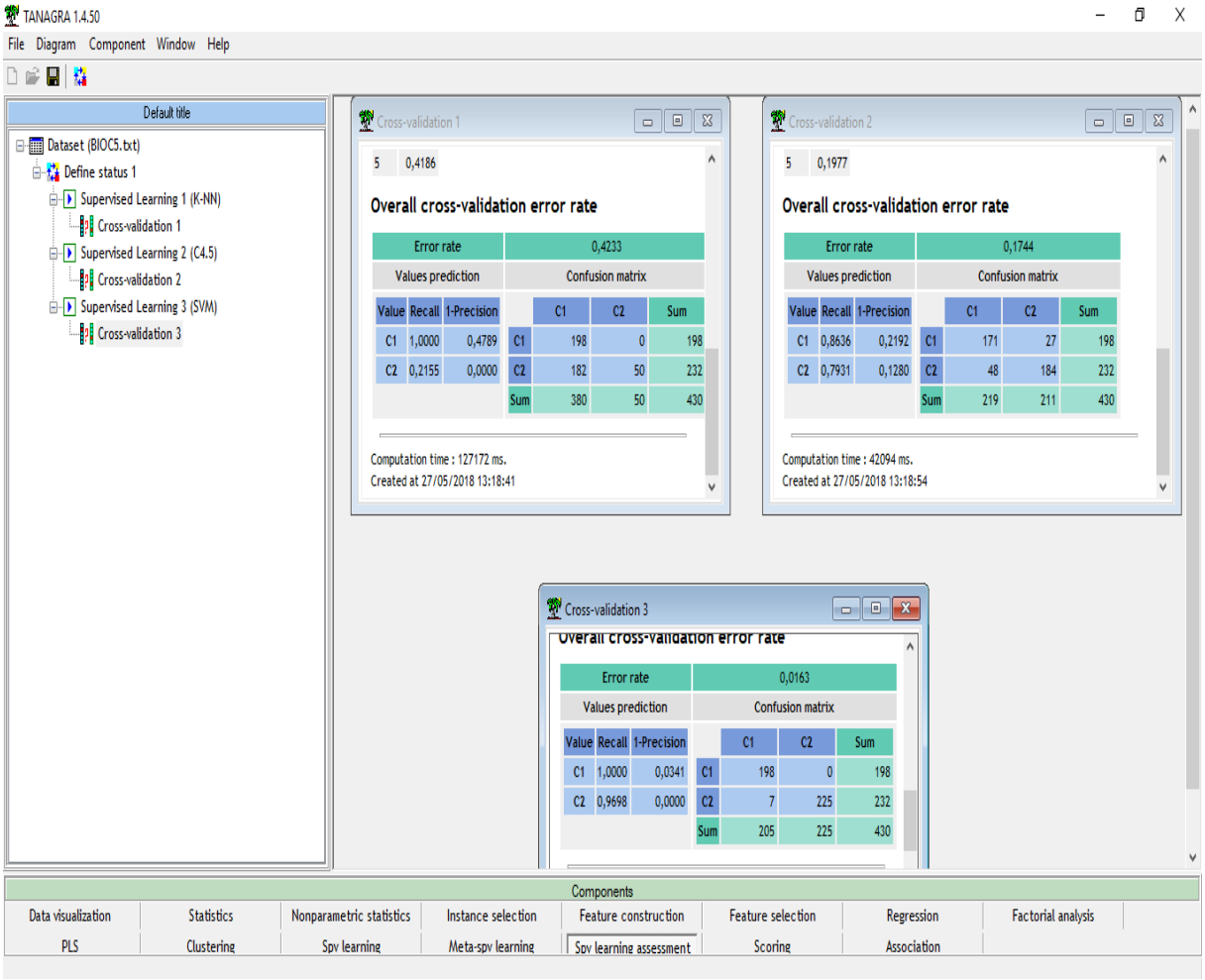
Algorithme de classification	*Taux d'erreur*
Knn	0.511
C4.5	0.067
SVM	0.083

- Tableau Booléen ([3-4]-grammes)



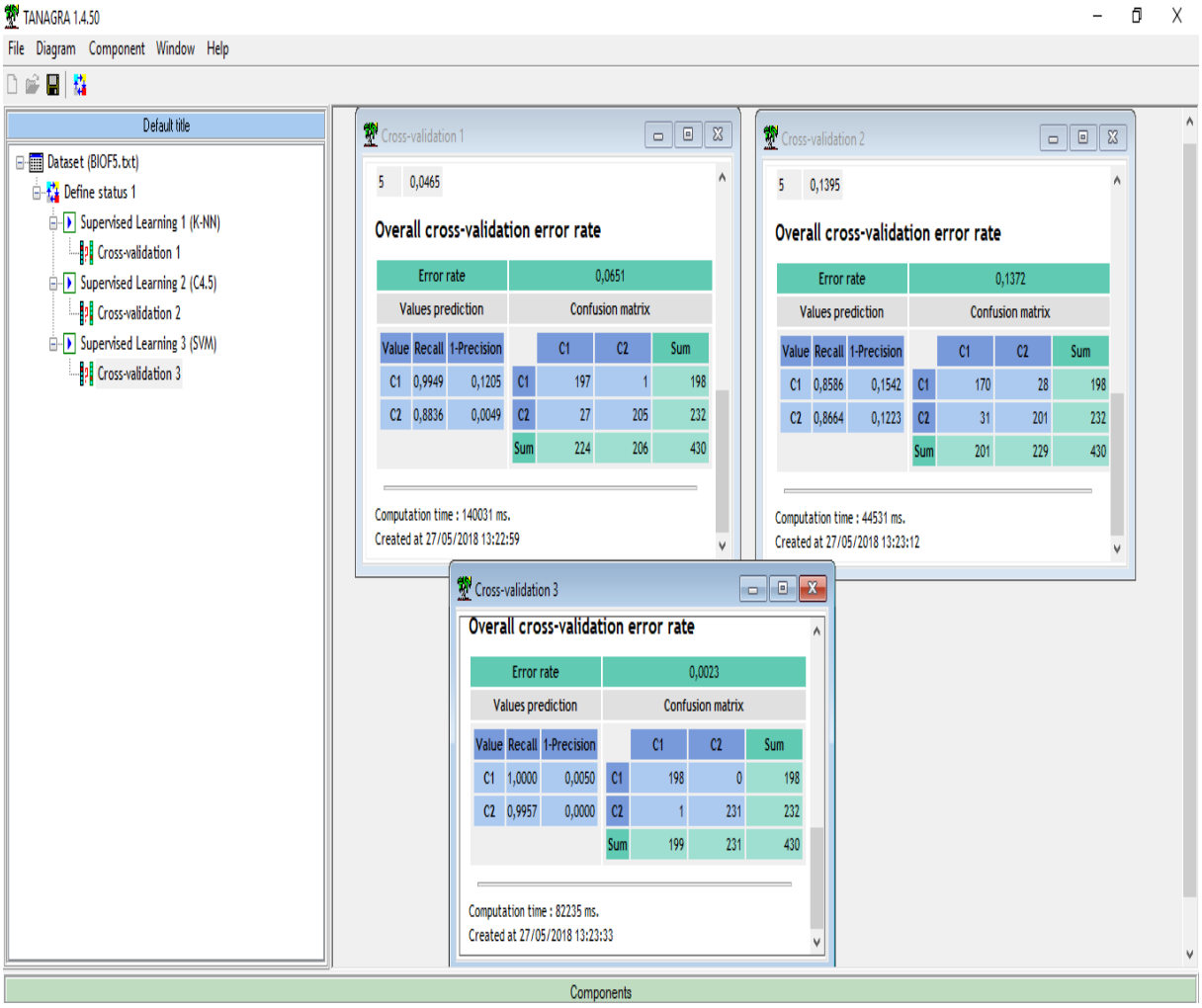
Algorithme de classification	*Taux d'erreur*
Knn	0.334
C4.5	0.174
SVM	0.014

- Tableau Occurrences ([3-4]-grammes)



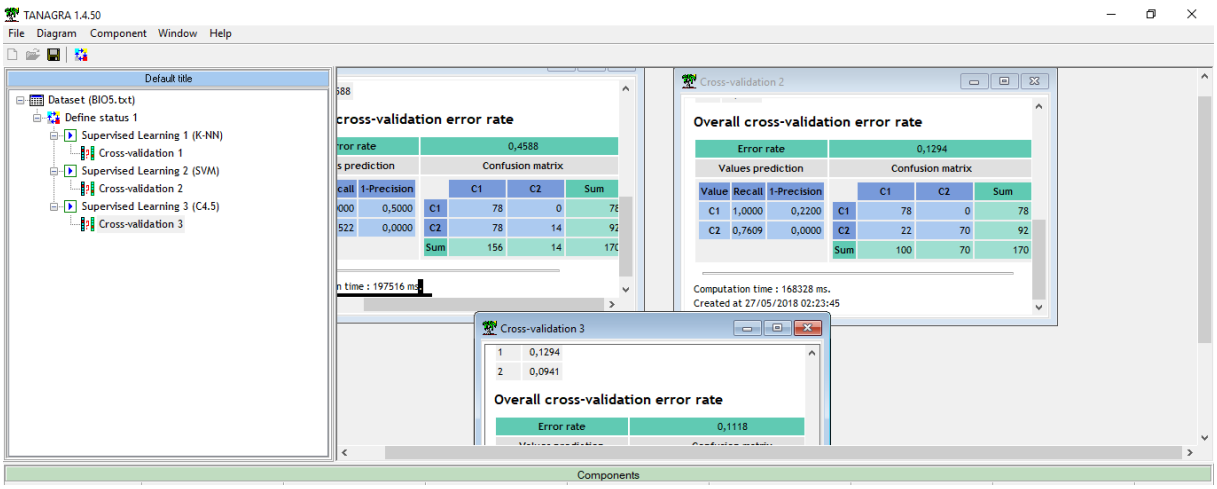
Algorithme de classification	*Taux d'erreur*
Knn	0.423
C4.5	0.174
SVM	0.0163

- Tableau Fréquences([3-4]-grammes)



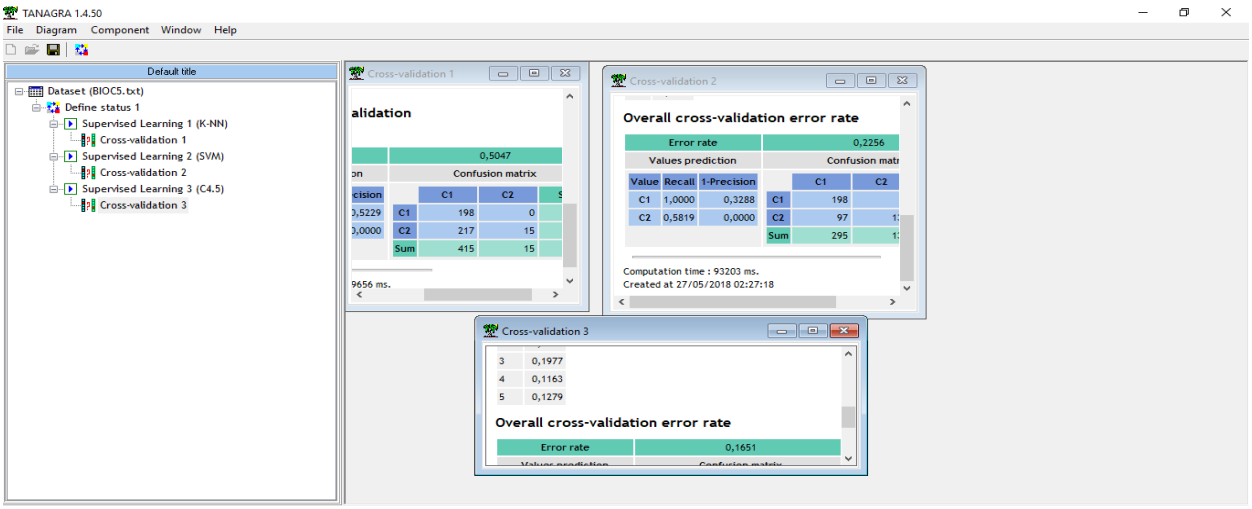
Algorithme de classification	*Taux d'erreur*
Knn	0.065
C4.5	0.137
SVM	0.002

➢ Tableau Booléen(5-grammes)



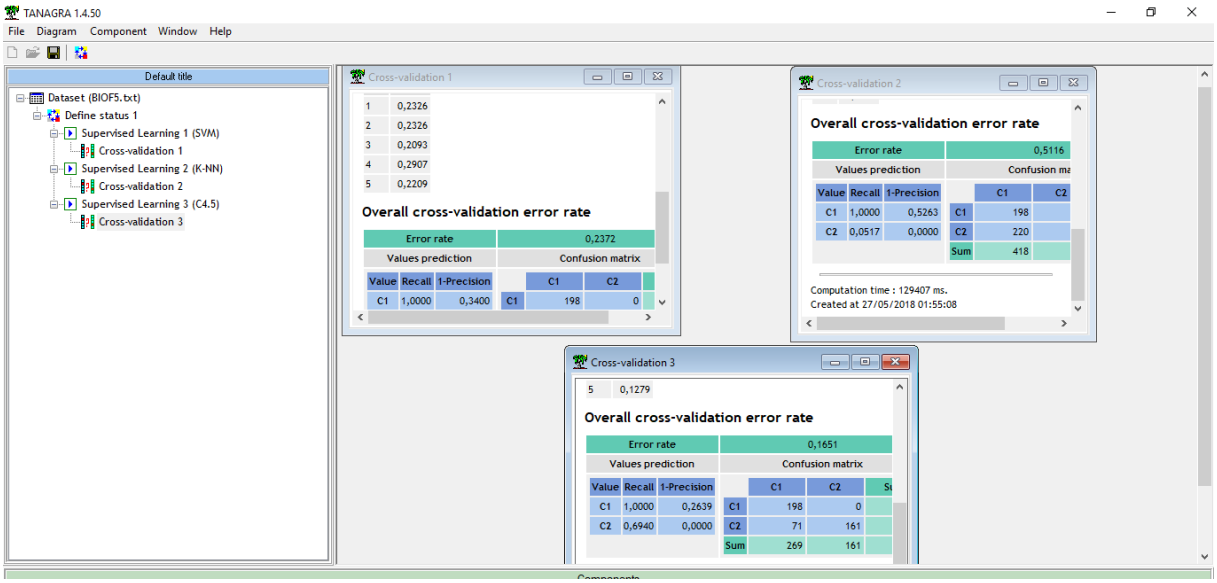
Algorithme de classification	*Taux d'erreur*
Knn	0.458
C4.5	0.111
SVM	0.129

➢ Tableau Occurrences (5-grammes)



Algorithme de classification	*Taux d'erreur*
Knn	0.504
C4.5	0.165
SVM	0.225

➢ Tableau Fréquences(5-grammes)



Algorithme de classification	*Taux d'erreur*
Knn	0.511
C4.5	0.165
SVM	0.237

Protéines pairs	[2]-*grams*	[3]-*grams*	[2-3]-*grams*	[4]-*grammes*	[3-4]-*grammes*	[5]-*grammes*
Occurrence	0.055	0.502	0.539	0.504	0.423	0.504
Booléen	0.539	0.453	0.369	0.504	0.334	0.458
Fréquences	0.011	0.430	0.369	0.511	0.065	0.511

Tableau 2.Résultats détaillés pour l'algorithme KNN

Protéines pairs	[2]-*grams*	[3]-*grams*	[2-3]-*grams*	[4]-*grammes*	[3-4]-*grammes*	[5]-*grammes*
Occurrence	0.018	0.025	0.025	0.100	0.016	**0.225**
Booléen	0.016	0.014	0.016	0.095	0.014	0.129
Fréquences	0.0	0.007	0.007	0.083	0.002	0.237

Tableau 3.Résultats détaillés pour l'algorithme SVM

➔**Meilleure résultat 2 et [3-4]-grammes Tableau Fréquences.**

⇨ **Meilleure Algorithme de classification SVM**

Protéines pairs	[2]-*grammes*	[3]-*grammes*	[2-3]-*grammes*	[4]-*grammes*	[3-4]-*grammes*	[5]-*grammes*
Occurrence	0.137	0.125	0.151	0.067	0.174	0.165
Booléen	0.139	0.125	0.158	0.064	0.174	0.111
Fréquences	0.216	0.159	0.204	0.067	0.137	0.165

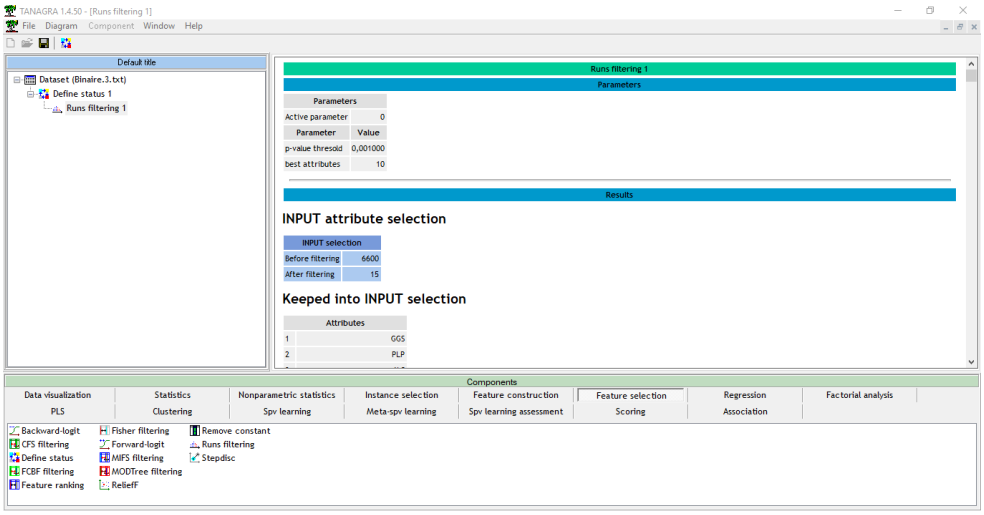
Tableau 4.Résultats détaillés pour l'algorithme C4.5

⇨ **Meilleure Algorithme de classification SVM et C4.5**

Sélection d’attributs

La sélection d’attributs consiste à réduire l’ensemble des attributs considérés et peut augmenter la précision d’un algorithme de regroupement/classification, améliorer la qualité des données, réduire le temps de calcul et/ou l’espace mémoire, en réduisant la dimension de l’espace des attributs par l’élimination des attributs redondants, non pertinents ou bruités.

Sélection d´attributs Avec Tanagra



N-grammes = 3

Fisher filtering

Nombre d’attributs avant sélection= 6600

Nombre d’attributs après sélection= 251

Logiciels et langages utilisées

Langage R

R est un langage de programmation et un logiciel libre dédié aux statistiques et à la

Science des données soutenu par la R Foundation for Statistical Computing. R fait partie

De la liste des paquets GNU et est écrit en C (langage), Fortran et R.

Le langage R est largement utilisé par les statisticiens et les data miner pour le

Développement de logiciels statistiques et l'analyse des données.

Logiciel Tanagra :

TANAGRA est un logiciel gratuit de DATA MINING destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

TANAGRA est un projet ouvert au sens qu'il est possible à tout chercheur d'accéder au code et d'ajouter ses propres algorithmes pour peu qu'il respecte la licence de distribution du logiciel.

L'objectif principal du projet TANAGRA est d'offrir aux chercheurs et aux étudiants une **plate-forme de Data Mining** facile d'accès, respectant les standards des logiciels du domaine, notamment en matière d'interface et de mode de fonctionnement, et permettant de **mener des études** sur des données réelles et/ou synthétiques.

Conclusion

Le travail réalisé dans ce document vise à présenter un processus optimal de classification des protéines.

Le processus tirer le maximum d'avantages, d'une part, en identifiant le meilleur couple de n-grammes et Classificateur. D'un autre côté, les processus de classification doit être réalisé dans un délai raisonnable. Par conséquent, nous avons varié la taille de n-grammes entre {2,3,2-3,3-4 ,4 ,5}, puis, évalué chaque type de n-grammes avec divers classificateurs, tels que KNN, Arbre de décision et SVM. Les résultats montrent que le couple optimal n'est pas le même habituellement. Mais le SVM et les classificateurs d'arbre de décision donnent les meilleurs résultats Respectivement avec 2 grammes ,3 grammes, [3-4]-grammes.