

Hakiri Siwar 22113801

Mazeline Robin 22203804

L2 MIASHS

M.Sallaberry

Science des données 1

Classification supervisé pour l'analyse de la Cirrhose

Et

Classification non supervisé pour l'analyse de textes ayant pour thème la technologie

Table des matières

Introduction.....	3
1 - Apprentissage non supervisé.....	4
1.1 - Jeu de données.....	4
1.2 – Nettoyage et prétraitement des données.....	4
1.3 – Clustering.....	6
2- Apprentissage Supervisé.....	10
2.1-Introduction:.....	10
2.2- Collecte de jeux de données:.....	10
2.3-Nettoyage et prétraitement des données:.....	10
2.4- Modèles d'apprentissages:.....	10
2.5-Prédictions:.....	18
2.6-Discussion:.....	19
2.7-Conclusion:.....	19
Conclusion.....	20

Table des figures

Figure 1: Chaîne pour visualiser les textes.....	4
Figure 2: Sac de mots non traités.....	4
Figure 3: Figure 3: Chaîne de prétraitement.....	5
Figure 4: Une partie de la liste de mots d'arrêt.....	5
Figure 5: Interface de « Preprocess Text ».....	5
Figure 6: Sac de mots traités.....	6
Figure 7: Chaîne de traitement entière.....	6
Figure 8: matrice des distances entre les textes.....	7
Figure 9: dendrogramme des textes séparés en 2 clusters.....	8
Figure 10: Contenu des clusters.....	8
Figure 11: dendrogramme des textes séparés en 8 clusters.....	9
Figure 12: présentation des variables.....	10
Figure 13: présentation des suites des variables.....	11
Figure 14: présentation des données.....	11
Figure 15: insertion du premier modèle d'apprentissage.....	12
Figure 16: chaîne de traitement avec le modèle KMN.....	12
Figure 17: Figure de classification KMN.....	13
Figure 18: Chaîne de traitement avec tous les modèles.....	13
Figure 19: Visualisation des performances des différents modèles d'apprentissage.....	14
Figure 20: matrice de confusion du modèle logistique régression.....	15
Figure 21: visualisation des données des individus CL classés comme D.....	16
Figure 22: visualisation des données des individus correctement classés dans D.....	16
Figure 23: présentation des données des individus CL classés comme C.....	17
Figure 24: présentation des données des individus correctement classés dans C.....	17
Figure 25: présentation des données à prédire.....	18
Figure 26: chaîne de traitement des données non étiquetés.....	18
Figure 27: étiquetage des données après la prédiction.....	19

Introduction

La classification a pour but de regrouper (partitionner, segmenter) n observations en un certain nombre de groupes ou de classes homogènes. Il existe deux principaux types de classification:

- la classification supervisée, souvent appelée simplement classification
- la classification non supervisée, parfois appelée partitionnement, segmentation ou regroupement (*Clustering* en anglais).

L'apprentissage supervisé consiste à fournir à un algorithme un ensemble de données d'entrée, également appelé ensemble de données d'apprentissage, avec des exemples étiquetés. Ces exemples étiquetés se composent de caractéristiques (ou variables indépendantes) et d'une valeur cible (ou variable dépendante), permettant ainsi à l'algorithme d'apprendre à associer les caractéristiques aux valeurs cibles correspondantes. L'algorithme utilise ces exemples étiquetés pour générer un modèle qui peut être utilisé pour prédire la valeur cible des nouvelles observations non étiquetées. Ce type d'apprentissage porte dans notre cas sur la survie des patients atteints de cirrhose qui sera la variable principale de prédiction. Elle est codée comme suit : 0 pour le décès (D), 1 pour le censuré (C) et 2 pour le censuré en raison d'une transplantation hépatique (CL).

D'autre part, l'apprentissage non supervisé consiste à fournir à un algorithme un ensemble de données d'entrée non étiquetées. L'objectif de ce type d'apprentissage est de découvrir des structures, des modèles ou des relations inhérentes aux données sans aucune information préalable sur les valeurs cibles à prédire. Les algorithmes d'apprentissage non supervisé utilisent différentes techniques, telles que le regroupement (clustering) pour regrouper les données similaires. Notre étude présente une méthode de classification de textes axée sur le thème des technologies dont l'objectif principal est de regrouper les textes en différents clusters en fonction de leurs sous-thèmes de manière précise et cohérente .

Nous pouvons observer ici que les mots les plus fréquents ne sont pas les plus utiles dans notre étude de cet apprentissage non supervisé. Pour affiner notre étude nous allons utiliser une chaîne de prétraitement de données sur Orange.

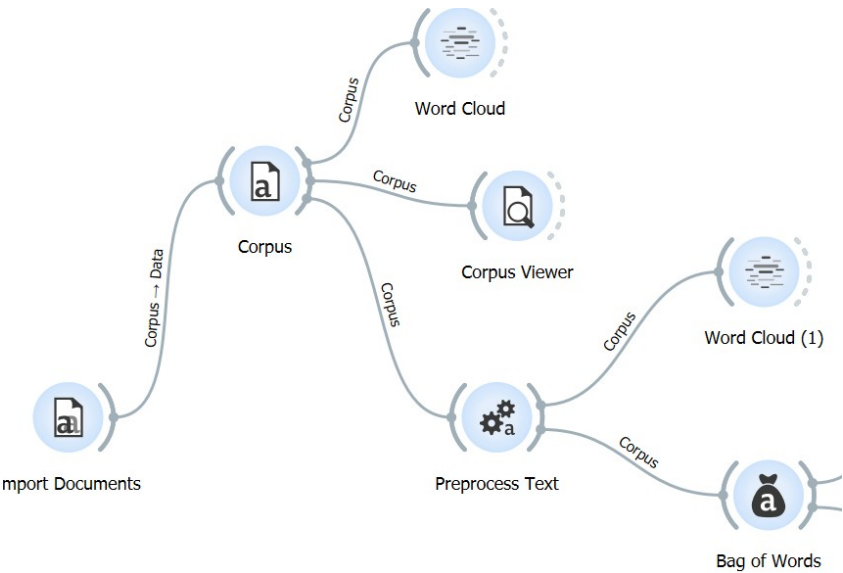


Figure 3: Figure 3: Chaîne de prétraitement

En sélectionnant l’onglet «Preprocess Text», on peut ajouter un liste de mots d’arrêt qui va nous permettre de garder seulement les mots les plus pertinents pour notre étude.

Figure 4:
Une partie
de la liste
de mots
d'arrêt

said
also
would
could
us
like
many
already
every
good
likely
want
made
says
technology
use
get
think
three
one
two
third
using
make
mr
first
using
help
user
users
used
the
that
this
on

☒ Stopwords
 English
 mot d'arret.txt
 ...

☐ Lexicon
 (none)
 ...

☒ Numbers
 Includes Numbers

☐ Regexp
 \.|\:|;|!|\?|\(|\)|\||\+|\||\"|'|"|'|\||\...|\-|_|-|\\$|&|*|>|<|\||\||

☐ Document frequency
 Relative: 0,10 0,90
 Absolute: 1 10

☐ Most frequent tokens
 100

Figure 5: Interface de « Preprocess Text »

Ainsi grâce à ce prétraitement nous pouvons observer dans le nuage de mots, des mots essentiels à notre étude ; des mots qui ont pour thème la technologie.



Figure 6: Sac de mots traités

Le nettoyage et le prétraitement des données sont maintenant terminés. Nous pouvons donc commencer à analyser les résultats obtenus

1.3 – Clustering

Voici la chaîne de traitement complète qui va nous permettre d'étudier nos textes.

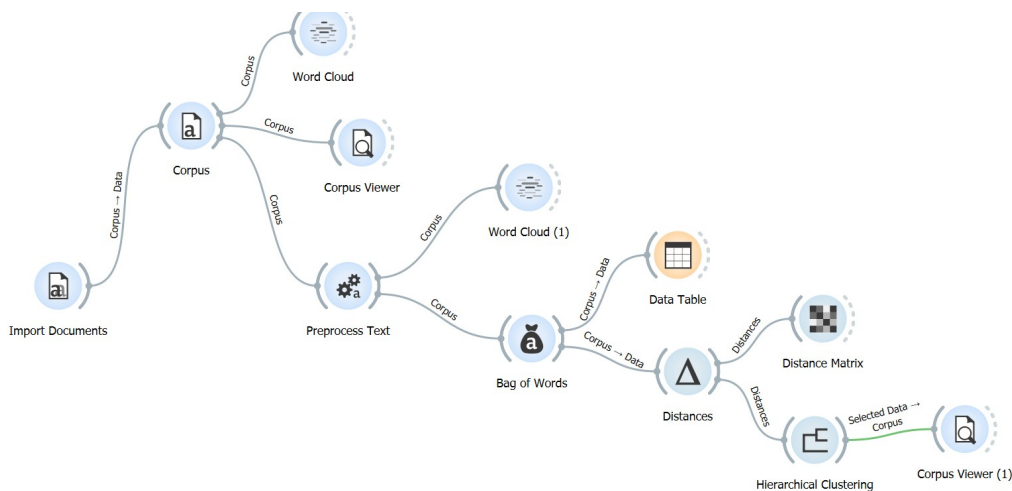


Figure 7: Chaîne de traitement entière

En sorti de notre prétraitement de données, on retrouve l'onglet « Bag of Words » . Cette action consiste à classer les textes en fonction de leur nombre d'occurrences d'un mot dans un texte.

Pour visualiser ces mots nous allons dans l'onglet « Data Table » . Ainsi nous pourrons voir pour chaque texte les mots avec le plus d'occurrences.

« Distances » nous permet de ranger les textes en fonction de leur similarité.

Dans « Distance Matrix », nous pouvons observer ce graphique avec différents nombres décimaux. Cela représente la similarité entre les textes. Plus le nombre décimal se rapproche de 1 plus les deux textes observés sont similaires. Plus le nombre décimal se rapproche de 0 moins ils le sont. 1 indique que les deux textes sont identiques et 0 indique que les deux textes ne contiennent aucun mot en commun.

	technologie_1	technologie_10	technologie_100	technologie_11	technologie_12	technologie_13	technologie_14	technologie_15	technologie_16	technologie_17	technologie_18	technologie_19	t
technologie_1		0,646	0,960	0,967	0,988	0,821	0,722	0,982	0,954	0,980	0,961	0,922	
technologie_10	0,646		0,951	0,929	0,977	0,838	0,799	0,933	0,956	0,988	0,984	0,989	
technologie_100	0,960	0,951		0,906	0,912	0,977	0,952	0,933	0,967	0,670	0,961	0,942	
technologie_11	0,967	0,929	0,906		0,968	0,946	0,955	0,956	0,933	0,938	0,931	0,971	
technologie_12	0,988	0,977	0,912	0,968		0,970	0,948	0,970	0,966	0,946	0,932	0,970	
technologie_13	0,821	0,838	0,977	0,946	0,970		0,898	0,955	0,965	0,981	0,970	0,993	
technologie_14	0,722	0,799	0,952	0,955	0,948	0,898		0,973	0,947	0,965	0,965	0,942	
technologie_15	0,982	0,933	0,933	0,956	0,970	0,955	0,973		0,963	0,969	0,935	0,957	
technologie_16	0,954	0,956	0,967	0,933	0,966	0,965	0,947	0,963		0,946	0,972	0,955	
technologie_17	0,980	0,988	0,670	0,938	0,946	0,981	0,965	0,969	0,946		0,977	0,975	
technologie_18	0,961	0,984	0,961	0,931	0,932	0,970	0,965	0,935	0,972	0,977		0,945	
technologie_19	0,922	0,989	0,942	0,971	0,970	0,993	0,942	0,957	0,955	0,975	0,945		
technologie_2	0,985	0,937	0,980	0,834	0,981	0,957	0,953	0,978	0,953	0,967	0,980	0,976	
technologie_20	0,934	0,888	0,968	0,932	0,947	0,811	0,941	0,877	0,965	0,971	0,901	0,978	
technologie_21	0,806	0,854	0,984	0,953	0,991	0,915	0,460	0,966	0,994	0,990	0,982	0,984	
technologie_22	0,944	0,951	0,922	0,969	0,916	0,987	0,966	0,934	0,971	0,983	0,944	0,975	
technologie_23	0,916	0,942	0,855	0,945	0,948	0,975	0,930	0,912	0,977	0,970	0,949	0,810	
technologie_24	0,963	0,943	0,938	0,941	0,969	0,971	0,971	0,971	0,967	0,953	0,903	0,986	
technologie_25	0,993	0,991	0,972	0,981	0,963	0,973	0,984	0,892	0,978	0,977	0,774	0,992	
technologie_26	0,994	0,941	0,932	0,932	0,984	0,963	0,985	0,983	0,981	0,987	0,973	0,987	
technologie_27	0,982	0,856	0,929	0,934	0,956	0,955	0,948	0,966	0,940	0,966	0,982	0,981	
technologie_28	0,946	0,870	0,961	0,949	0,899	0,901	0,950	0,843	0,964	0,981	0,974	0,980	
technologie_29	0,978	0,930	0,972	0,967	0,656	0,956	0,961	0,927	0,960	0,973	0,974	0,983	
technologie_3	0,971	0,901	0,968	0,944	0,867	0,926	0,971	0,857	0,964	0,996	0,983	0,992	

Figure 8: matrice des distances entre les textes

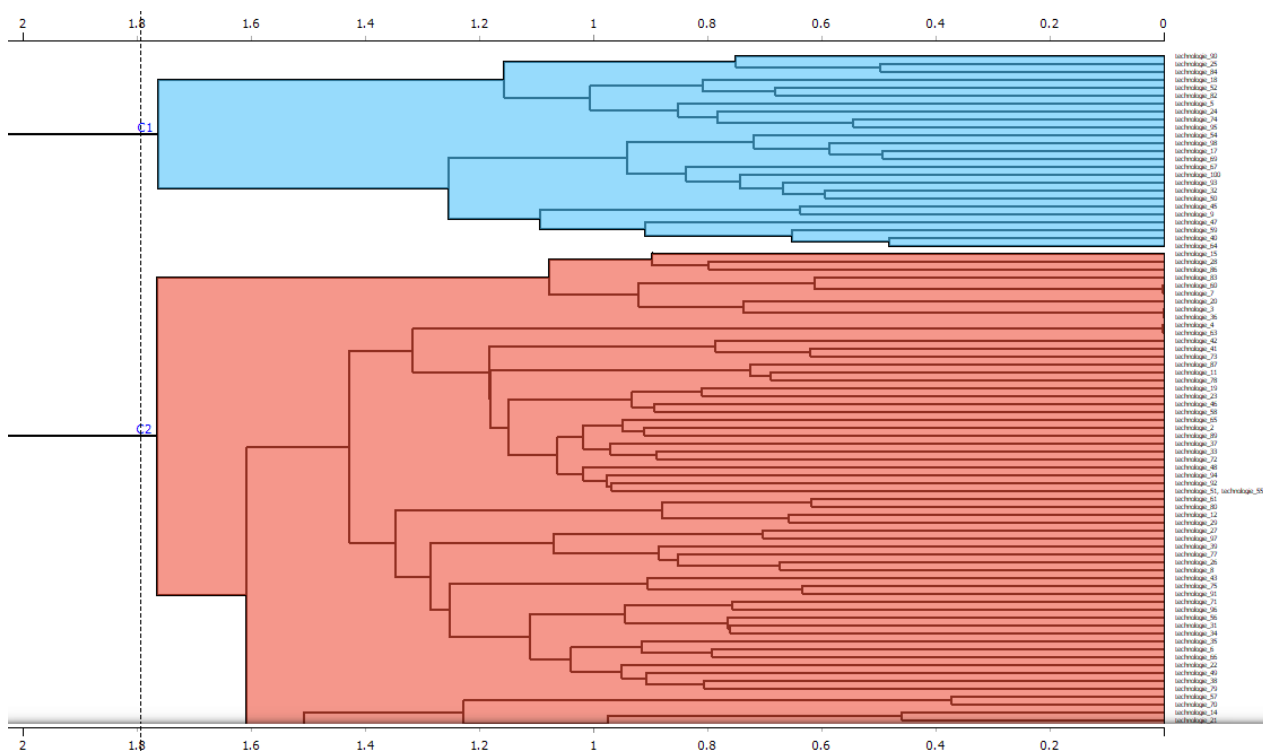


Figure 9: dendrogramme des textes séparés en 2 clusters

Le «Hierarchical Clustering» nous permet de diviser les 100 textes en différents clusters. Ici nous avons divisé les textes en 2 clusters. Les textes réunis dans un cluster sont liés par un même thème. Plus on resserre les clusters plus le thème sera précis et inversement plus on élargit le cluster moins le thème sera précis.

« Corpus Viewer (1) » nous permet de voir les textes qui sont dans les clusters sélectionnés. C'est donc à nous trouver le thème du cluster parmi les textes qui sont sélectionnés à l'intérieur.

Dans ces 2 importants clusters le thème de C1 est les appareils technologiques. et celui de C2 est Internet

RegExp Filter:	
1	technologie_1
2	technologie_10
3	technologie_100
4	technologie_11
5	technologie_12
6	technologie_13
7	technologie_14
8	technologie_15
9	technologie_16
10	technologie_17
11	technologie_18
12	technologie_19
13	technologie_2
14	technologie_20
15	technologie_21
16	technologie_22
17	technologie_23
18	technologie_24
19	technologie_25
20	technologie_26
21	technologie_27

name (1): technologie_1

path: C:/Users/rmaze/OneDrive/Documents/MIASHS/L2/Sciences des Données/Projet/technologie/technologie_1.txt

content (1): The Arizona Attorney General (AG) Mark Brnovich has filed a consumer fraud lawsuit against Google, alleging that the company used "deceptive" practices to track the location of users even after they turned off location tracking.

Brnovich shared information about the lawsuit on his Twitter account. He accused Google of using "deceptive and unfair practices to obtain users' location data". This data is exploited for advertising, which accounts for more than 80 per cent of Google's revenue.

Google collects detailed information about its users, including their physical locations, to target users for advertising. Often, this is done without the users' consent or knowledge.

Brnovich wrote on Twitter: "Google collects detailed information about its users, including their physical locations, to target users for advertising. Often, this is done without the users' consent or knowledge."

He told the Washington Post that Google has been trying to find "misleading ways" to obtain information from users who try to opt out of data collection. He added that Google may be the "most innovative company in the world" but it is not above the law.

His office's investigation of Google was initiated after reading a 2018 Associated Press report, which detailed how Google users are "lulled into a false sense of security" by privacy options, including the option to disable location history. The investigation found that many Google services on Android devices (and also iPhones) store location data even if the user selects privacy settings which purport to prevent Google from doing so.

The findings were confirmed by computer scientists at Princeton University.

The lawsuit specifically alleges that Google maintained location tracking for certain features, including weather and search engine queries, after the user disabled app-specific location tracking. Only when the user turns off broader system-level tracking did Google stop tracking location.

"We brought forward this action to put a stop to Google's deceptive collection of user data, obtain monetary relief and require Google to disgorge gross receipts arising from its Arizona activities," Brnovich added in another tweet.

Figure 10: Contenu des clusters

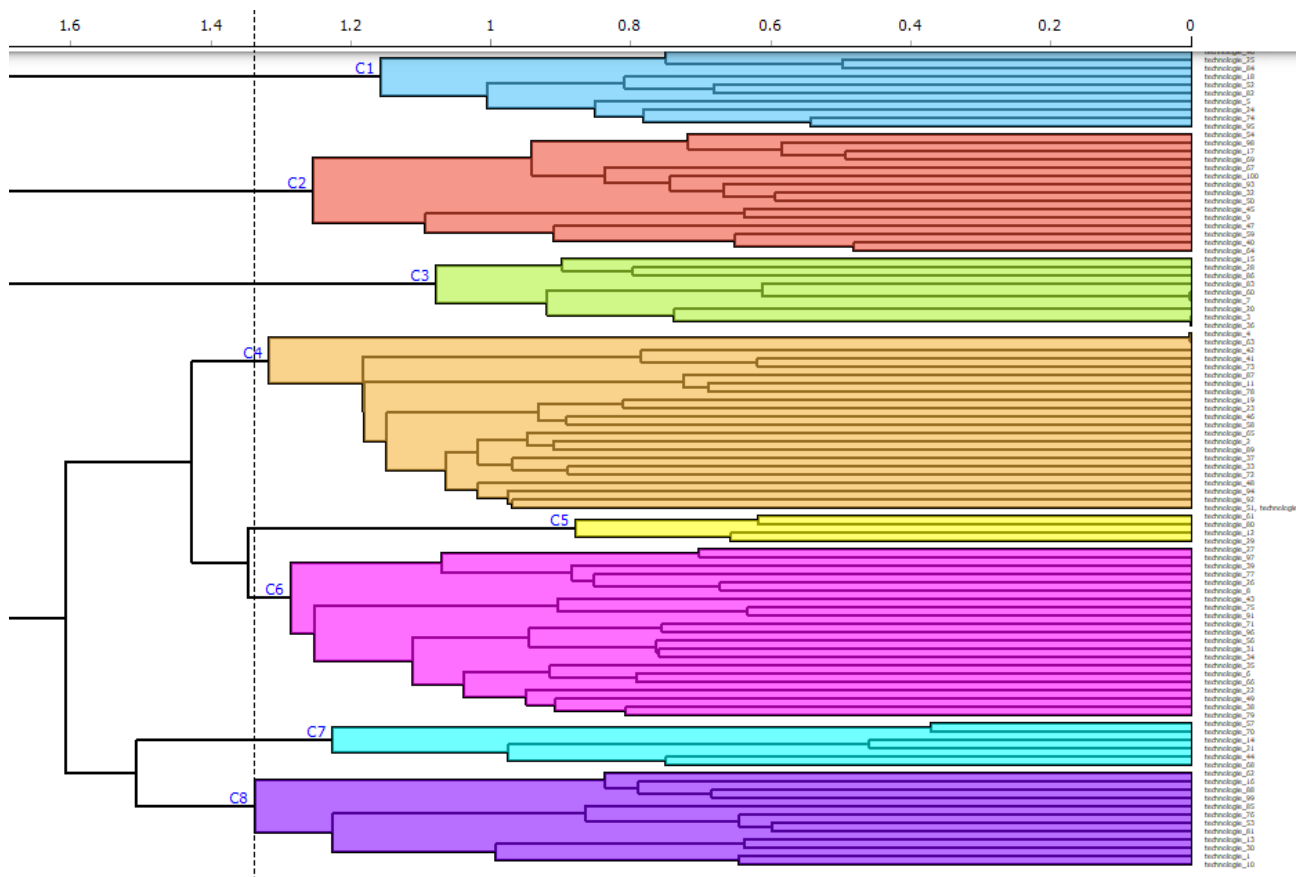


Figure 11: dendrogramme des textes séparés en 8 clusters

Nous avons voulu savoir le contenu de ces textes plus précisément. Au niveau 1,35 cette coupe permet d'obtenir 8 clusters. Chaque clusters ont un sous-thème. Grâce à l'utilisation de «Corpus Viewer (1)» nous avons pu trouver ces sous-thème.

- C1:Jeux Vidéos
- C2 :Téléphone portable
- C3:Microsoft
- C4:Appareils technologiques portables
- C5:Linux push
- C6:Cyber criminalité
- C7:Apple
- C8:Google

2- Apprentissage Supervisé

2.1-Introduction:

La cirrhose du foie est une maladie chronique qui peut entraîner de graves complications et une mortalité élevée. Il est essentiel de prédire la survie des patients atteints de cirrhose afin de pouvoir prendre des décisions cliniques éclairées et de fournir les meilleurs soins possibles. Dans cette analyse, nous allons classer les données sur Orange selon une variable indiquant l'état de survie des patients atteints de cirrhose du foie.

2.2- Collecte de jeux de données:

Le jeu de donnée sur lequel porte l'étude a été récupéré du site web [Kaggle.com](https://www.kaggle.com)

Pour classer ces données, nous utiliserons Orange, un logiciel d'analyse de données qui offre de nombreuses fonctionnalités pour le prétraitement, la visualisation et la modélisation des données. Dans ce cadre nous allons traiter un apprentissage supervisé qui porte sur la prédiction de la survie des patients atteints de cirrhose: les états de survie seront codés comme suit : 0 pour le décès (D), 1 pour le censuré (C) et 2 pour le censuré en raison d'une transplantation hépatique (CL).

2.3-Nettoyage et prétraitement des données:

Dans le cadre de notre prétraitement des données, nous avons choisi de supprimer toutes les lignes qui contenaient des valeurs manquantes (NA) dans la colonne "Drogue".

2.4- Modèles d'apprentissages:

Nous allons importer le jeu de donnée «Prédiction de la survie des patients atteints de cirrhose» disponible sur le site kaggle grâce à l'outil File.

il contient 304 individus avec:

*2 variable numérique contenant l'identifiant et le nombre du jour de chaque patients atteints

*9 variables quantitatives caractérisant chaque individu

*1 variable nominale (classification) indiquant l'état de survie des patients atteints de cirrhose du foie: Les états de survie sont les suivants : 0 = D (décès), 1 = C (censuré), 2 = CL (censuré en raison d'une transplantation hépatique).

	Name	Type	Role	Values
1	ID	N numeric	feature	
2	N_Days	N numeric	feature	
3	Status	C categorical	target	C, CL, D
4	Drug	C categorical	feature	D-penicillamine, Placebo
5	Age	N numeric	feature	
6	Sex	C categorical	feature	F, M
7	Ascites	C categorical	feature	N, Y
8	Hepatomegaly	C categorical	feature	N, Y
9	Spiders	C categorical	feature	N, Y
10	Edema	C categorical	feature	N, S, Y
11	Bilirubin	N numeric	feature	
12	Cholesterol	N numeric	feature	

Figure 12: présentation des variables













	Name	Type	Role	Values
9	Spiders	 categorical	feature	N, Y
10	Edema	 categorical	feature	N, S, Y
11	Bilirubin	 numeric	feature	
12	Cholesterol	 numeric	feature	
13	Albumin	 numeric	feature	
14	Copper	 numeric	feature	
15	Alk_Phos	 numeric	feature	
16	SGOT	 numeric	feature	
17	Tryglicerides	 numeric	feature	
18	Platelets	 numeric	feature	
19	Prothrombin	 numeric	feature	
20	Stage	 numeric	feature	

Figure 13: présentation des suites des variables

Une fois les données chargées dans Orange, nous pourrons ajouter Data Table en sortie de File et parcourir les données de façon à bien les comprendre :

Status	Drug	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin
C	D-penicillamine	F	N	N	N	N	1.0	239	3.77
C	Placebo	M	N	N	N	N	1.8	460	3.35
D	Placebo	M	N	Y	N	N	2.3	178	3.00
C	Placebo	F	N	N	N	N	0.9	400	3.60
C	D-penicillamine	F	N	N	N	N	0.9	248	3.97
D	Placebo	F	Y	Y	Y	Y	2.5	188	3.67
D	D-penicillamine	F	N	N	N	N	1.1	303	3.64
CL	Placebo	F	N	Y	N	N	1.1	464	4.20
D	D-penicillamine	F	Y	Y	N	N	2.1	?	3.90
C	Placebo	F	N	N	N	N	0.6	212	4.03
D	D-penicillamine	F	N	N	N	N	0.4	127	3.50
C	Placebo	F	N	N	N	N	0.5	120	3.61
D	D-penicillamine	F	N	Y	Y	N	1.9	486	3.54
CL	D-penicillamine	F	N	N	N	N	5.5	528	4.18
D	Placebo	F	N	Y	Y	N	2.0	267	3.67

Figure 14: présentation des données

L'objectif est d'entraîner des modèles à reconnaître, à partir des variables quantitatives, l'étiquette contenue dans la variable classification. Autrement dit, nous cherchons à créer des modèles classant l'état de survie des patients atteints de cirrhose du foie. Une fois entraînés, nous allons utiliser les modèles sur un jeu de données ne contenant pas la variable classification.

Pour commencer, nous allons ajouter l'algorithme des k plus proches voisins en sortie de File :

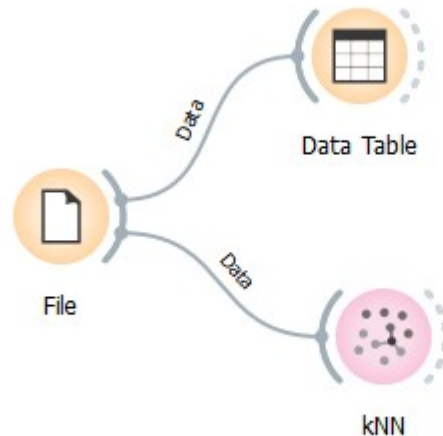


Figure 15: insertion du premier modèle d'apprentissage

En sortie de kNN, nous allons ajouter l'outil Test and Score sans oublier de prendre en entrée les données issues de File ceci permet de générer une validation croisée du modèle et en sortie de Test and Score nous ajoutons une matrice de confusion

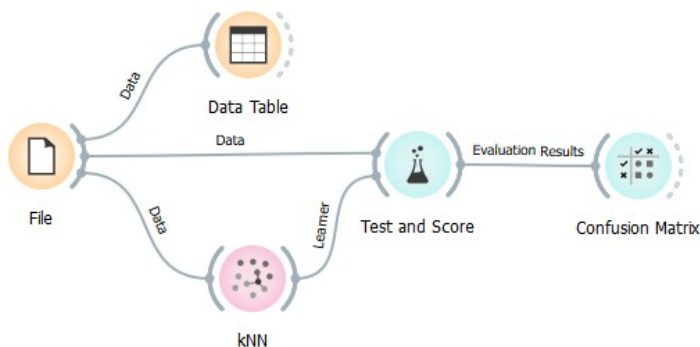


Figure 16: chaîne de traitement avec le modèle KMN

et on obtient cette matrice à l'issue de la matrice du confusion

		Predicted			
		C	CL	D	Σ
Actual	C	147	0	19	166
	CL	15	0	3	18
	D	57	0	62	119
	Σ	219	0	84	303

Figure 17: Figure de classification KMN

On peut observer que 147 C , 62 D et 0 CL ont été correctement classés. Par contre, 19 C ont été classés comme D, 15 CL ont été classés C , 3 CL ont été classés D par contre 0 CL ont été classés dans CL et 57 D ont été classés comme C et 0 D ont été classés dans CL. Le modèle est donc plutôt bon mais loin d'être infaillible puisque il prédit jamais CL.

On peut dire qu'il est difficile de prédire le sujet "Prédiction de la survie des patients atteints de cirrhose" par l'algorithme KMN (K-means clustering) car il est traditionnellement utilisé pour des tâches de clustering non supervisées, où l'on n'a pas de variable cible prédéfinie à prédire. Dans notre cas les données fournissent une variable cible décrivant l'état de survie des patients atteints de cirrhose du foie, avec 3 états possibles : décès (D), censuré (C) ou censuré en raison d'une transplantation hépatique (CL).

Pour prédire cet état de survie à partir des autres variables d'entrée, on pourrait utiliser des techniques de classification supervisée plutôt que de clustering. Donc nous allons ajouter les modèles disponibles dans le menu Model tel que Logistic regression, Naive Bayes, SVM, Tree, Random Forest et Neural Network puis comparer kNN avec les autres modèles de classification

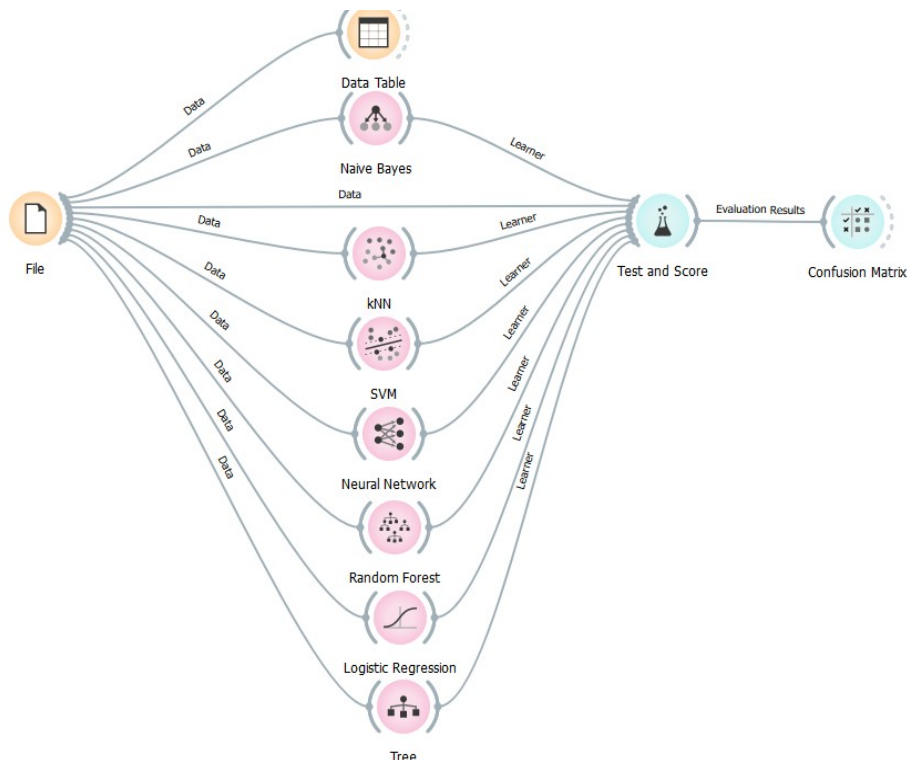


Figure 18: Chaîne de traitement avec tous les modèles

Ces résultats indiquent la performance des différents modèles d'apprentissage:

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.885	0.795	0.770	0.748	0.795	0.606
Logistic Regression	0.933	0.845	0.836	0.833	0.845	0.707
kNN	0.742	0.690	0.658	0.658	0.690	0.397
Tree	0.820	0.759	0.757	0.755	0.759	0.552
Naive Bayes	0.875	0.690	0.730	0.802	0.690	0.513
Random Forest	0.846	0.779	0.755	0.734	0.779	0.575
Neural Network	0.901	0.822	0.811	0.807	0.822	0.663

Figure 19: Visualisation des performances des différents modèles d'apprentissage

L'AUC (Area Under the Curve), également appelée précision, mesure la capacité du modèle à classer correctement l'état de survie des patients atteints de cirrhose du foie. Plus l'AUC est proche de 1, meilleure est la performance du modèle. On peut donc voir que la régression logistique et le réseau de neurones obtiennent les meilleurs résultats avec des AUC de 0,933 et 0,901 respectivement, ce qui suggère qu'ils sont capables de bien classer les données

Le CA (Classification Accuracy), ou taux de classification correcte, mesure la précision globale du modèle en termes de pourcentage de prédictions correctes. Dans ce cas la régression logistique obtient le meilleur CA par rapport aux autres modèles avec 0,845.

Le F1-score(F1) est une mesure de précision globale qui combine la précision et le rappel du modèle. Il est important de noter que le F1-score est une moyenne harmonique des deux métriques et est plus équilibré lorsque les classes sont déséquilibrées. Les meilleurs modèles en termes de score F1 sont donc la régression logistique 0,836, le réseau de neurones 0,811 et le SVM 0,770

La précision (PREC) mesure le taux de prédictions correctes parmi toutes les prédictions positives du modèle. On remarque que la régression logistique a la meilleure précision avec 0,833, suivie du réseau de neurones 0,807 et du SVM 0,748.

Le rappel (RECALL) mesure le taux de prédictions correctes parmi toutes les instances réellement positives. La régression logistique a également le meilleur rappel avec 0,845.

En résumé, la régression logistique semble être le modèle le plus performant en termes de capacité de classification et de précision globale

l'analyse des données sur la prédiction de la survie des patients atteints de cirrhose à l'aide d'Orange a montré que régression logistique était le modèle le plus performant avec une précision de prédiction de 83%. Ce modèle a obtenu une précision élevée dans la prédiction des différents états de survie, y compris les cas de censure en raison d'une transplantation hépatique. Ces résultats suggèrent que la régression logistique peut être utilisée comme outil précieux dans la prise de décision clinique pour prédire la survie des patients atteints de cirrhose.

		Predicted			
		C	CL	D	Σ
Actual	C	152	2	12	166
	CL	10	4	4	18
	D	16	3	100	119
	Σ	178	9	116	303

Figure 20: matrice de confusion du modèle logistique régression

On remarque que même si logistique régression est le modèle le plus performant il a du mal à bien prédire CL, comme on peut le voir sur la matrice il y'a que 4CL qui ont été classé correctement.

Ceci peut être expliqué par un :

-Déséquilibre des classes : Il se peut que la classe CL soit sous-représentée par rapport aux autres classes (D et C). Cela peut entraîner un biais dans l'apprentissage du modèle, où il se concentre davantage sur les classes majoritaires et n'apprend pas suffisamment sur les cas de censure en raison d'une transplantation hépatique.

2-Manque de données : Les modèles d'apprentissage automatique dépendent fortement de la qualité des données d'entraînement. Si les données utilisées pour entraîner le modèle sont bruyantes, incomplètes ou mal étiquetées, cela peut affecter la performance du modèle.

3-Caractéristiques prédictives insuffisantes : Les caractéristiques ou variables utilisées pour prédire l'état de survie CL peuvent ne pas être pertinentes ou informatives. Si les caractéristiques utilisées ne capturent pas efficacement les facteurs qui influencent la transplantation hépatique, le modèle ne pourra pas prédire avec précision cette classe.

4-caractéristiques non linéaires : il y'a des modèles qui sont linéaires par défaut, ce qui signifie qu'ils peuvent avoir du mal à modéliser des relations non linéaires entre les caractéristiques et la variable cible. Si la relation entre les caractéristiques et la survie des patients atteints de cirrhose est non linéaire, cela peut rendre difficile la prédiction

5-Paramètres de modèle inappropriés : Les modèles ont plusieurs paramètres qui peuvent être ajustés lors de l'entraînement pour optimiser la performance. Si ces paramètres ne sont pas correctement réglés, le modèle peut ne pas être en mesure de capturer les motifs dans les données ou de gérer le sur-ajustement.

Afin d'identifier les critères sur lesquels les variables sont classées selon le modèle logistique régression, nous allons ajouter un Data Table en sortie de Confusion Matrix puis sélectionner les 4 individus classés comme D alors que ce sont des CL en cliquant sur la case correspondante ce qui nous permet d'observer les données de ces individus

Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides
6.1	1712	2.83	89	3681.0	158.10	139
1.3	316	3.51	75	1162.0	147.25	137
3.5	325	3.98	444	766.0	130.20	210
5.5	528	4.18	77	2404.0	172.05	78

Figure 21: visualisation des données des individus CL classés comme D

Après nous allons les comparer avec les données des individus de la classe D qui ont été classés correctement dans D

1.3	408	4.22	67	1387.0	142.60	137
3.2	?	3.56	77	1790.0	139.50	?
8.0	468	2.81	139	2009.0	198.40	139
4.0	416	3.99	177	960.0	86.00	242
1.4	259	4.16	46	1104.0	79.05	79

Figure 22: visualisation des données des individus correctement classés dans D

Nous pouvons observer que les valeurs de l'albumine et des triglycérides sont très similaires. Par exemple, pour un taux de triglycérides de 139, le taux d'albumine est de 2,83, ce qui est presque identique à la valeur de la figure 22 (2,83). Par conséquent, nous pouvons conclure que la variable "D" est classée en fonction des taux d'albumine et de triglycérides.

		Predicted			
		C	CL	D	Σ
Actual	C	152	2	12	166
	CL	10	4	4	18
	D	16	3	100	119
	Σ	178	9	116	303

Figure 20(b) :matrice de confusion du modèle logistique régression

On s'intéresse maintenant au 10 individus classés comme C alors que ce sont des CL et on observe leur données

Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin
1.1	464	4.20	38	1644.0	151.90	102	348	10.3
3.5	348	3.20	121	938.0	120.00	146	296	10.0
3.4	450	3.37	32	1408.0	116.25	118	313	11.2
3.2	339	3.18	123	3336.0	205.00	84	304	9.9
0.5	201	3.73	44	1345.0	54.25	145	445	10.1
3.4	356	3.12	188	1911.0	92.00	130	318	11.2
1.1	432	3.57	45	1406.0	190.00	77	248	11.4
1.6	?	3.07	136	1995.0	128.00	?	372	9.6
2.1	387	3.77	63	1613.0	150.35	33	185	10.1
8.7	310	3.89	107	637.0	117.00	242	298	9.6

Figure 23: présentation des données des individus CL classés comme C

On compare ces données avec les données des individus de la classe C qui ont été classés correctement en sélectionnant la case correspondante

0.5	280	4.23	36	377.0	56.00	146	227	10.6
3.2	375	3.14	129	857.0	89.00	?	375	9.5
3.6	374	3.50	143	1428.0	188.00	44	151	10.1
1.0	317	3.56	44	1636.0	84.00	111	394	9.8
2.0	310	3.36	70	1257.0	122.00	118	143	9.8
8.6	546	3.73	84	1070.0	127.00	153	291	11.2
1.7	434	3.35	39	1713.0	171.00	100	234	10.2

Figure 24: présentation des données des individus correctement classés dans C

Les résultats montrent une similitude remarquable entre les valeurs de la bilirubine et du taux de prothrombine. De plus, pour une valeur de 0,5 de bilirubine, nous obtenons un taux de prothrombine de 10,6, ce qui est presque identique à la valeur obtenue dans la figure 24 (10,1). Par conséquent, il est possible d'affirmer que la variable C est classée en fonction de la bilirubine et du taux de prothrombine.

2.5-Prédictions:

Maintenant après avoir comparer les modèles et identifier celui qui produit les meilleurs résultats, nous allons pouvoir faire des prédictions .Pour cela, on commence par récupérer le fichier des données a prédire qu'on a crée et le chargez dans Orange à l'aide d'un second File et visualisez son contenu à l'aide d'un Data Table :

	Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin
1	?	D-penicillamine	21464	F	Y	Y	Y	Y	14.5	261	2.60
2	?	D-penicillamine	20617	F	N	Y	Y	N	1.1	302	4.14
3	?	D-penicillamine	25594	M	N	N	N	S	1.4	176	3.48
4	?	D-penicillamine	19994	F	N	Y	Y	S	1.8	244	2.54
5	?	Placebo	13918	F	N	Y	Y	N	3.4	279	3.53
6	?	Placebo	24201	F	N	Y	N	N	0.8	248	3.98
7	?	Placebo	20284	F	N	Y	N	N	1.0	322	4.09
8	?	Placebo	19379	F	N	N	N	N	0.3	280	4.00
9	?	D-penicillamine	15526	F	N	N	Y	N	3.2	562	3.08

Figure 25: présentation des données a prédire

Comme on peut le voir, le fichier contient 9 individus pour lesquels les variables caractéristiques des patients atteints ont été spécifiées. Par contre, il ne contient pas de variable classification permettant de déterminer les états de survie si la ligne correspond a un D,CL,C. C'est cette variable que on va prédire grâce aux modèles entraînés précédemment.

En sortie de File avec lequel on a chargé les données à prédire, on ajoute un Prédictions (menu Evaluate). En plus des données,on ajoute une connexion de façon à ce que Prédictions prenne en entrée le modèle le plus performant

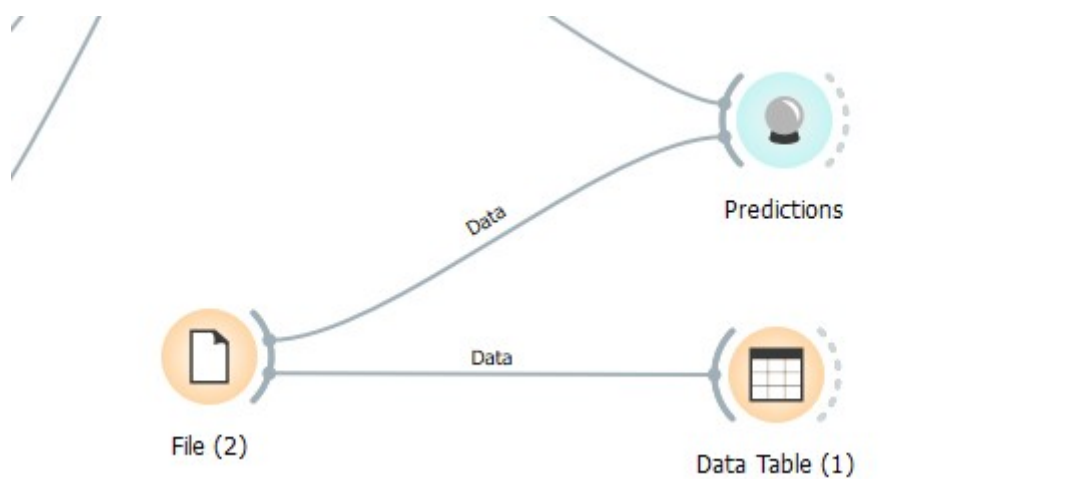


Figure 26: chaîne de traitement des données non étiquetés

	Logistic Regression	ID	N_Days	Status	Drug	Age
1	D	1	400	?	D-penicillamine	21464
2	C	2	4500	?	D-penicillamine	20617
3	D	3	1012	?	D-penicillamine	25594
4	D	4	1925	?	D-penicillamine	19994
5	D	5	1504	?	Placebo	13918
6	D	6	2503	?	Placebo	24201
7	D	7	1832	?	Placebo	20284
8	D	8	2466	?	Placebo	19379
9	D	9	2400	?	D-penicillamine	15526

Figure 27: étiquetage des données après la prédiction

Comme on peut le voir le modèle classe toutes les lignes comme D sauf la deuxième ligne il la classe comme C.

2.6-Discussion:

L'analyse de la prédiction de la survie des patients atteints de cirrhose du foie peut fournir des informations précieuses pour les professionnels de la santé. Par exemple, en identifiant les facteurs qui influencent la survie des patients, nous pourrions mieux comprendre la progression de la maladie et développer des stratégies de traitement plus efficaces. De plus, cette analyse peut également permettre d'identifier les patients à haut risque de décès et de prioriser leur traitement.

2.7-Conclusion:

La classification des données sur Orange selon l'état de survie des patients atteints de cirrhose du foie permettra d'explorer et de prédire la survie de ces patients. L'analyse de ces résultats pourrait avoir un impact significatif sur la prise en charge des patients atteints de cirrhose, en fournissant des informations essentielles pour la planification des soins et le développement de nouveaux traitements.

Conclusion

Dans le cas de l'apprentissage supervisé, les modèles sont plus faciles à interpréter et à comprendre car l'ensemble de données d'entraînement est étiqueté. Cela permet de prédire avec précision les valeurs de sortie pour de nouvelles données en utilisant les caractéristiques apprises du jeu de données d'entraînement. De plus, les résultats de l'apprentissage supervisé peuvent être plus fiables et cohérents, car le modèle est formé pour se conformer à des valeurs cibles spécifiques.

Cependant, l'apprentissage supervisé nécessite un ensemble de données étiqueté, ce qui peut être coûteux et fastidieux à obtenir dans de nombreux domaines. De plus, les modèles entraînés en utilisant l'apprentissage supervisé peuvent être plus sensibles aux erreurs ou aux biais présents dans les données d'entraînement.

D'autre part, l'apprentissage non supervisé permet d'explorer des structures ou des patterns cachés dans un ensemble de données non étiqueté. Cela peut être utile pour la segmentation de données, la détection d'anomalies ou la recommandation personnalisée. De plus, l'apprentissage non supervisé peut être utilisé pour réduire la dimensionnalité des données et faciliter la visualisation.

Cependant, l'apprentissage non supervisé a des limites en termes de capacité à prédire des valeurs spécifiques pour de nouvelles données, car il n'y a pas de données d'entraînement étiquetées. De plus, l'interprétation des résultats de l'apprentissage non supervisé peut être plus difficile, car il n'y a pas de cibles spécifiques à évaluer.

En conclusion, le choix entre l'apprentissage supervisé et non supervisé dépend des objectifs spécifiques de l'analyse des données, de la disponibilité des données étiquetées et de la nature des données à traiter.