ISIMA
INSTITUT D'INFORMATIQUE
Université Clermont Auvergne

Rapport de projet
Data Mining

Yasmine SAIDI

10 mars 2021

Année 2020-2021

Enseignants : Barbara MARTIN
Natacha VANDERERUCH

Table des matières

Table des figures

Introduction

Le Data Mining est une technique parmi les composants de la Big Data. Il permet d'analyser les données volumineuses. Cette année on étudie cette matière, nous avons vu les notions basiques ainsi que les principaux algorithmes supervisé et non-supervisé. Pendant le TP, nous avons découvert le language R et la construction des modèles et leurs évaluaions avec le language python et l'application web Jupyter.

Dans ce projet on va mettre en place les notions que nous avons aquis pendant les séances de cours et de TP. Dans un premier temps je vais analyser des données avec R et dans le deuxième chapitre je vais créer un modèle de classification. Et enfin, je vais aborder le sujet d'analyse de topic modeling.

1 Visualisation de données avec R

1.1 Compréhension des données

Nous avons disposé d'un fichier csv 'Cholesterol_R'. Le jeu de données est composé de 18 enregistrements. Pour chaque enregistrement, nous disposons de 5 informations : ID, Before, After4weeks, After8weeks, Margarine.

Champs	Signification
ID	Identifiant de la personne.
Before	Le taux de cholesterol avant le commencement de régime.
After4weeks	Le taux de cholesterol après 4 semaines du commencement de régime.
After8weeks	Le taux de cholesterol après 8 semaines du commencement de régime.
Margarine	La marque de margarine utilisée pendant de régime.

TABLE 1 – Visualisation avec R - Dictionnaire de données

1.2 Préparation des données

J'ai choisi de visualiser la moyenne de la variation de niveau de cholesterol pour chaque marque de margarine et la variation de taux de cholesterol chez chaque personne. Donc il fallait faire un peut de calcul.

1. AJOUT DE COLONNES : J'ai ajouté deux colonnes qui représente la variation de niveau de cholesterol pour les deux cas après 4 semaines et après 8 semaines.

```
  ï..ID Before After4weeks After8weeks Margarine VariationAfter4Weeks
1     1   6.42        5.83        5.75         B                 0.59
2     2   6.76        6.20        6.13         A                 0.56
3     3   6.56        5.83        5.71         B                 0.73
4     4   4.80        4.27        4.15         A                 0.53
5     5   8.43        7.71        7.67         B                 0.72
6     6   7.49        7.12        7.05         A                 0.37
  VariationAfter8Weeks
1                 0.67
2                 0.63
3                 0.85
4                 0.65
5                 0.76
6                 0.44
```

FIGURE 1 – Visualisation avec R - Ajout de deux colonnes

2. CALCUL DES MOYENNES : J'ai calculer la moyenne de la variation de tous les gens après avoir filtrer les données selon la marque de la margarine.

```
> meanA
[1] 0.4855556 0.5466667
> meanB
[1] 0.6466667 0.7111111
```

FIGURE 2 – Visualisation avec R - Calcul des moyennes

Du coup pour la margarine A, le niveau de cholesterol est diminué d'un taux 0.48 après 4 semaines et de 0.54 après 8 semaines. Pour la margarine B, il y'a une dimuniation de 0.64 après 4 semaines et 0.71 après 8 semaines.

1.3 Exploration de données

J'ai choisi le diagramme à bâton pour visualiser les données. J'ai constitue le diagramme comme suit :

- sur l'axe horizontal, la durée de traitement ;
- Sur l'axe vertical, le taux moyenne de variation de niveau de cholesterol chez les patiants.

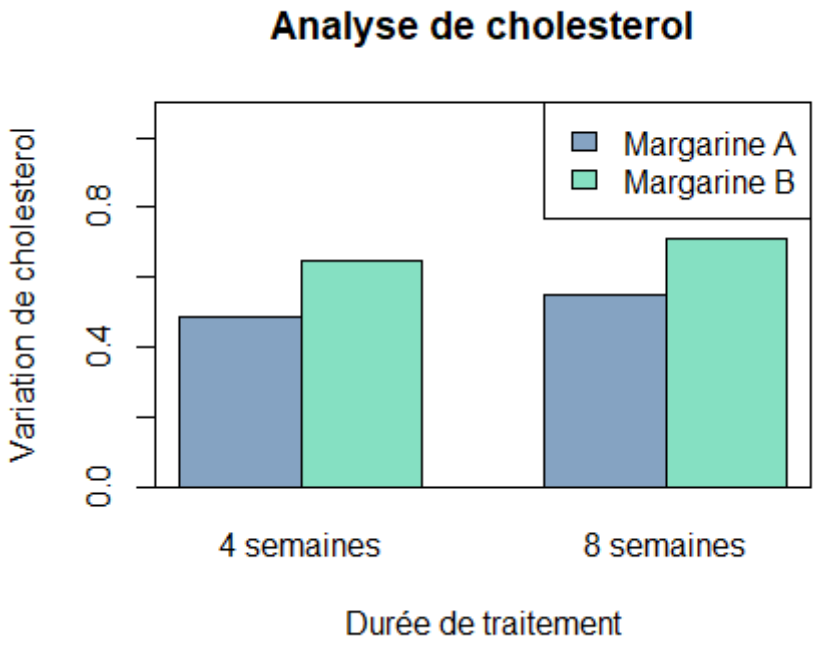

FIGURE 3 – Visualisation avec R - Diagramme à bâton

Pour visualiser le taux de cholesterol chez chaque personne dans les trois périodes j'ai choisi un diagramme linéaire.

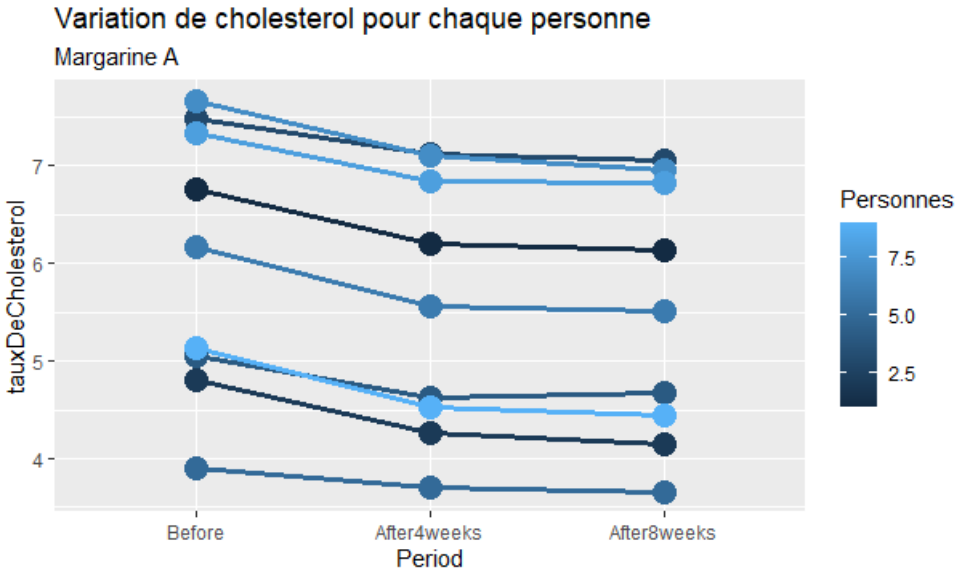

FIGURE 4 – Visualisation avec R - Diagramme linéaire - Margarine A

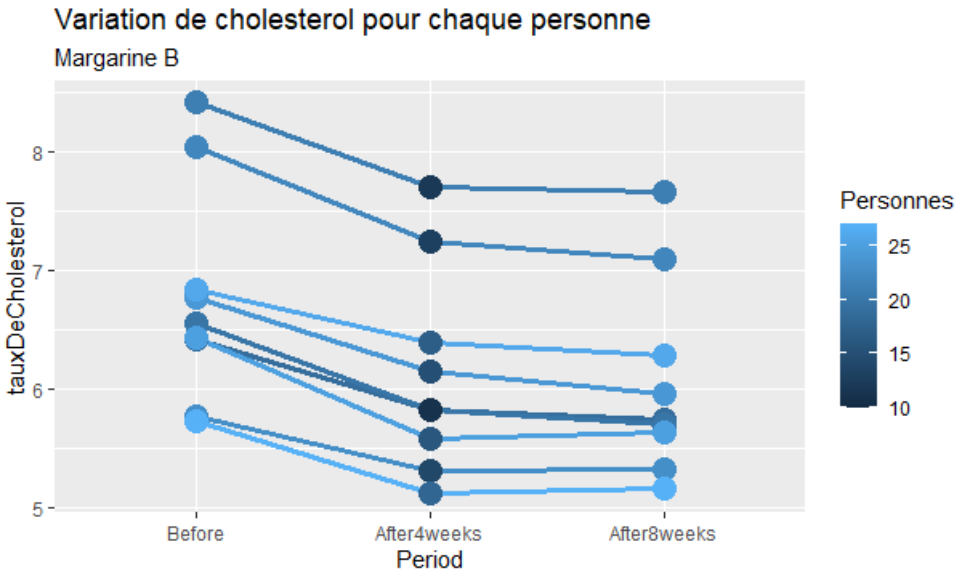

FIGURE 5 – Visualisation avec R - Diagramme linéaire - Margarine B

1.4 Analyse des résultats

D'après les figure précedantes, il apparaît que les deux margarine permettent de baisser le taux de cholesterol mais la maragarine B permet de diminuer davantage le niveau de cholesterol chez les gens.

2 Classification avec Python

Le but de cette partie est de faire une classification supervisé sur les données de patients. On veut prédire si une personne a une maladie cardiaque selon plusieurs données : l'âge, la hauteur, le poids, ...

Dans ce projet je vais suivre la méthode CRISP-DM qui signifie « Cross-industry standard process for data mining ». Chaque sous-section va être une étape de cette méthode. Donc nous aurons quatre sections dans cette partie qui sont : la compréhension métier, la compréhension des données, la préparation des données, la modélisation et enfin l'évaluation.

2.1 Compréhension du métier

Cette première phase consiste à bien comprendre les éléments métiers et la problématique qu'on vise à résoudre. je vais commencer par déterminer les objectifs stratégiques et opérationneles. Puis, je vais traduire l'objectif stratégique en concepts de Data mining.

2.1.1 Détermination des objectifs stratégiques et opérationnelles

Selon un article publié sur le site de la fondation pour la recherche médicale https ://www.frm.org : « On recense chaque année en France près de 70 000 décès liés à l'insuffisance cardiaque, et plus de 150 000 hospitalisations. ». Ces chiffres nous donnent une idée sur la propagation de ces maladies. Donc une étude de ce sujet sera très importante.

2.1.2 Détermination des objectifs du Data Mining

Puisque l'objectif stratégique est clairement défini, il convient maintenant de le traduire en concepts de Data Mining. On va opter à une technique prédictive puisque le but est de prédire si une personne peut avoir une maladie cardiaque ou non à partir des informations présentes. Parmi les techniques prédictives, on trouve les arbres de décision et les réseaux de neuronnes.

2.2 La compréhension des données

La phase de compréhension des données de CRISP-DM implique l'étude des données disponibles pour le Data mining. On doit tout d'abord decrire les données. La deuxième étape est l'exploration des données et comme étape finale vérifier la qualité des données.

2.2.1 Description des données

Pour chaque enregistrement, on a :

Code	Signification	Type
id	L'identifiant de la personne	nombre entier
age	L'age de la personne en jours	nombre entier
gender	Le sex de la personne : 1 : Si c'est une femme 2 : Si c'est un homme	variable qualitative : nombre entier
height	La longuer de la personne en cm	nombre entier
Weight	Le poids de la personne en kg	nombre réel
ap_hi	la pression artérielle systolique	nombre entier
ap_lo	la pression sanguine diastolique	nombre entier
cholesterol	le taux de cholestérol 1 : si le taux est normal 2 : si le taux est supérieur au taux normal 3 : si le taux est très supérieur au taux normal	variable qualitative nombre entier
gluc	le taux de glucose 1 : si le taux est normal 2 : si le taux est supérieur au taux normal 3 : si le taux est très supérieur au taux normal	variable qualitative nombre entier
smoke	Est que la personne fume ? 1 : si oui 0 : sinon	variable booléenne
alco	Est que c'est une personne alcoolique ? 1 : si oui 0 : sinon	variable booléenne
active	Est que c'est la personne pratique une activité sportive ? 1 : si oui 0 : sinon	variable booléenne
cardio	Est que c'est la personne a une maladie cardiaque ? 1 : si oui 0 : sinon	variable booléenne

TABLE 2 – Dictionnaire de données

La commande describe nous permet d'afficher des informations sur les attributs numérique tels que la valeur moyenne, la valeur maximale, la valeur minimale, ...

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	49972.419900	19468.865814	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129
std	28851.302323	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484
min	0.000000	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.000000

FIGURE 6 – Description des attributs

Cette matrice nous donne plusieurs informations utile. Le plus jeune personne 10798 jours donc 29 et la plus maigre personne a 10kg seulement!! Ce qui apparaît illogical.

2.2.2 Exploration des données

- **La matrice de corrélation :** La matrice de corrélation permet de trouver les features fortement correlé pour les supprimer en fin d'augmenter la robustesse et la stabilité du modèle et aussi simplifier les résultats obtenu par le modèle.

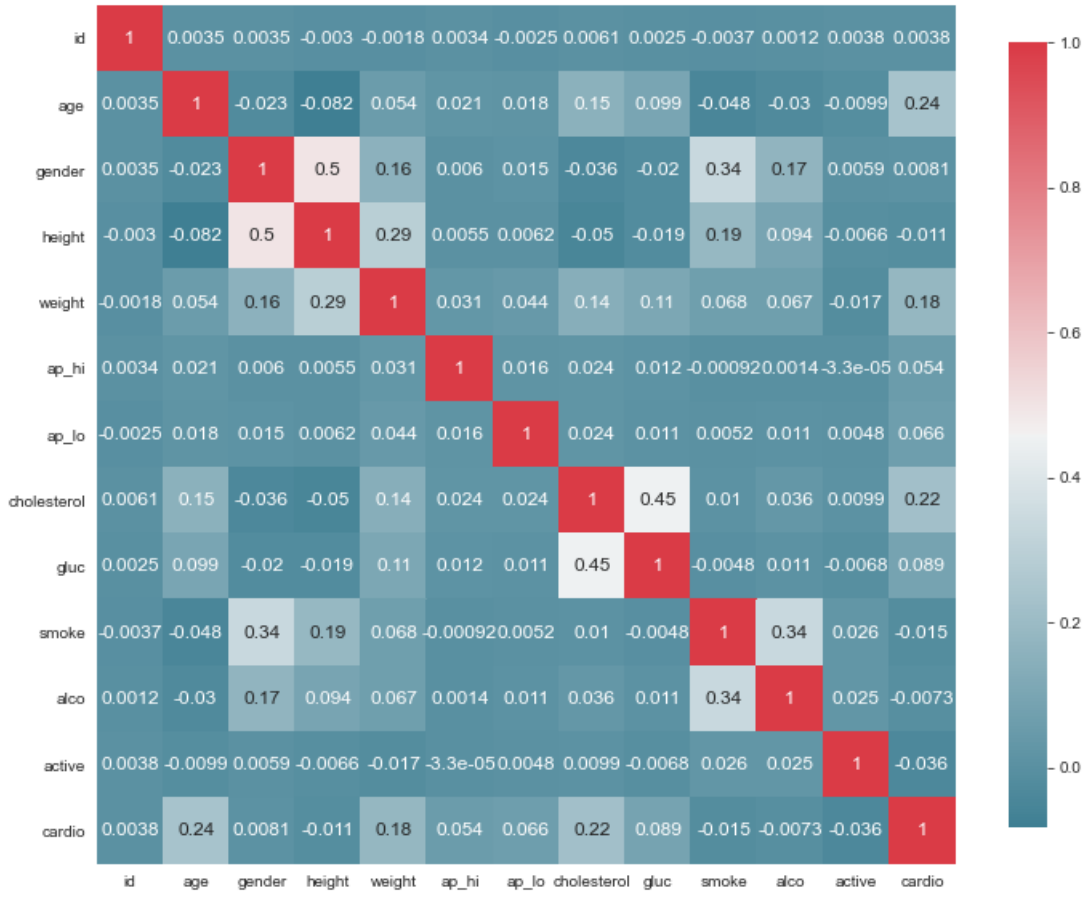

FIGURE 7 – La matrice de corrélation

D'après cette matrice, les deux variables height and gender sont très corrélées. Du coup on peut supprimer l'un des deux. Mais la variable gender est plus significative que la variable heigth. Donc on peut dispencer de cette variable pour créer notre modèle.

- **Digramme circulaire représentant le sexe des malades cardiaques :**

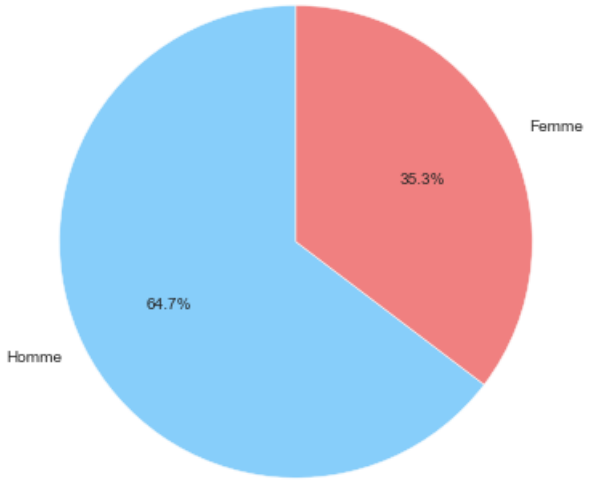

FIGURE 8 – le sexe des malades cardiaques

Ce diagrame montre que les femmes sont plus touchées aux maladies cardiaques que les hommes.

- **Diagramme en barres de taux de cholesterol et glucose des malades cardiaques :**

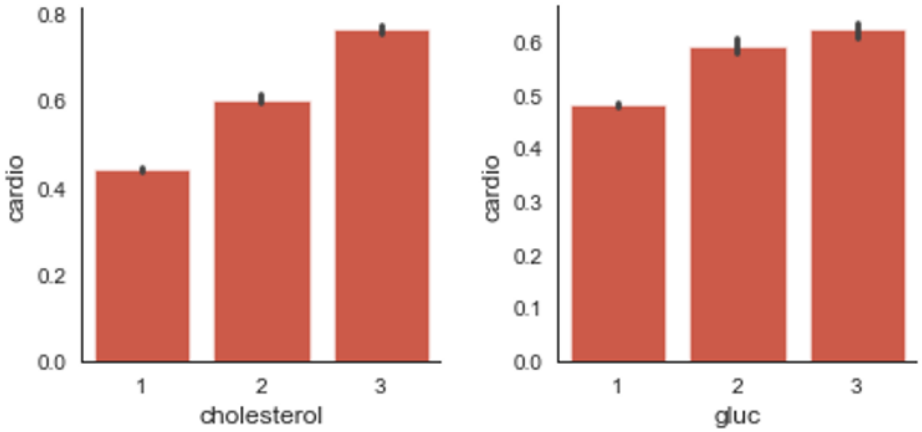

FIGURE 9 – le taux de cholesterol et glucose des malades cardiaques

Cette figure montre que le cholesterol une bonne variable de prédiction puisque la probabilité d'avoir des maladies cardiaque varie largement selon le niveau de cholesterol.

— **Histogramme d'age des gens qui ont des maladies cardiaques :**

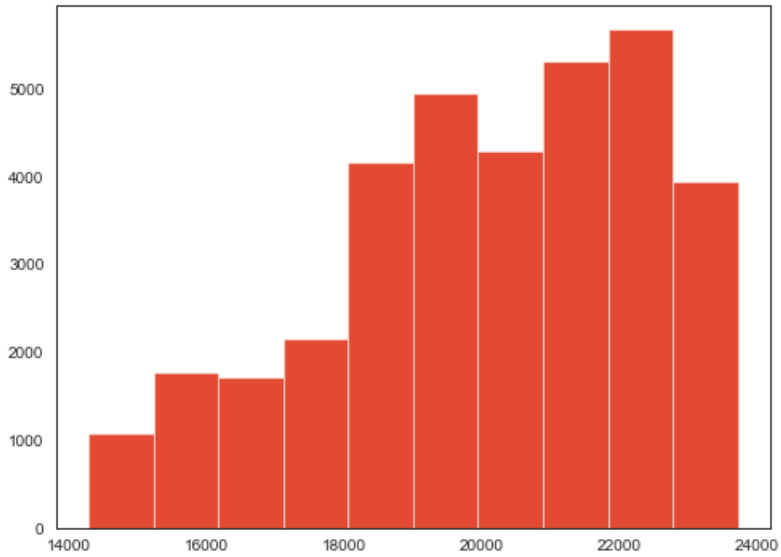

FIGURE 10 – l'âge des malades cardiaques

les gens agées sont plus touchées par les maladies cardiaques

— **Equilibrage de jeu de données :** Pour construire un modèle robuste, les classes de la variable cible doit être équilibrées. Du coup j'ai explorer cette information pour s'assurer que les données ne sont pas biaisées.

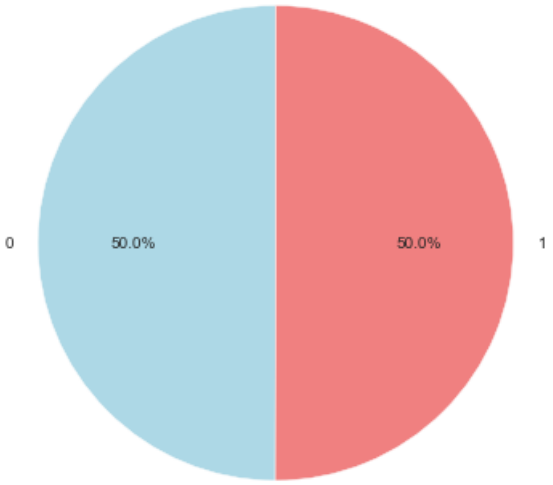

FIGURE 11 – Equilibrage du jeu de données

Heureusement les données sont parfaitement équilibrées.

2.3 Préparation des données

La préparation des données est l'un des aspects les plus importante et les plus coûteux en temps du Data Mining.

Nettoyage des données

- **Valeurs manquantes :**
  En utilisant la commande isnull, on peut savoir le nombre des valeurs manquantes dans chaque caractéristiques.
  D'après la figure suivante, on peut déduire qu'il y a aucune valeur manquante dans le jeu de données.

```
id             0
age            0
gender         0
height         0
weight         0
ap_hi          0
ap_lo          0
cholesterol    0
gluc           0
smoke          0
alco           0
active         0
cardio         0
dtype: int64
```

FIGURE 12 – Valeurs manquantes

- **Valeurs erronnées :**
  Comme déjà vu dans la section 2.2.1, j'ai remarqué une valeur anormale de poids

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
57858	82567	18804	2	165	10.0	180	1100	2	2	0	0	1	1

FIGURE 13 – Valeur anormale - poids

une personne de 51ans avec seulement 10kg est une chose normale. Puisque l'âge moyen de la population est 53ans donc on peut remplacer la valeur erronnée par la valeur moyenne des poids 74kg.
En analysant la pression astérielle systolique(ap_hi), j'ai remarqué qu'il y'a des valeurs anormales (1000, 11„ 14, 12, -150,160020). Donc j'ai décidé de visualiser le nuage des points cette caractéristique.

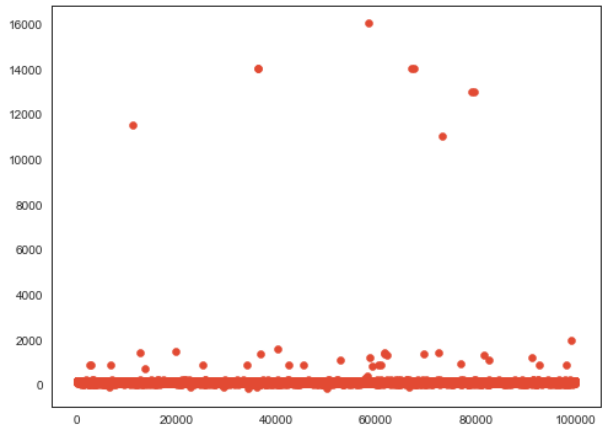

FIGURE 14 – Valeurs anormales - ap_hi

D'après la figure, il y'a 6 valeurs anormales. Puisque sont peu nombreux j'ai décidé de les supprimer.
De même pour la pression sanguine diastolique.

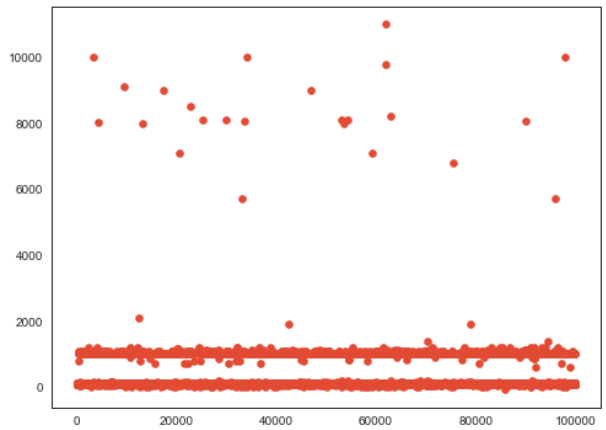

FIGURE 15 – Valeurs anormales - ap_lo

- **Nettoyage des données :**
  J'ai remplacer la valeur erronnée du poids par la mayenne des poids.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
57858	82567	18804	2	165	74.206607	180	1100	2	2	0	0	1	1

FIGURE 16 – Nettoyage - remplacement de poids anormale par la moyenne

J'ai supprimé les enregistrements qui contienent des valeurs anormales de ap_hi et ap_lo.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	
count	69966.000000	69966.000000	69966.000000	69966.000000	69966.000000	69966.000000	69966.000000	69966.000000	69966.000000	69966.000000	6
mean	49974.100163	19469.001143	1.349527	164.357974	74.207008	127.108867	93.783738	1.366907	1.226424	0.088143	
std	28852.056580	2467.426418	0.476824	8.209866	14.394833	28.278689	108.322281	0.680298	0.572244	0.283505	
min	0.000000	10798.000000	1.000000	55.000000	11.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000	
25%	25007.250000	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000	
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000	
75%	74891.750000	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000	
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	2000.000000	1900.000000	3.000000	3.000000	1.000000	

FIGURE 17 – Nettoyage - supression des lignes avec valeurs anormales

Sélection des données

Après avoir effectuer le nettoyage des données on doit maintenant choisir les données pertinentes pour nos objectifs de Data Mining.

J'ai éliminé une seule variable qui est la longueur puisqu'elle a une grande dépendance avec la variable genre.

	age	gender	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	62.0	110	80	1	1	0	0	1	0
1	20228	1	85.0	140	90	3	1	0	0	1	1
2	18857	1	64.0	130	70	3	1	0	0	0	1
3	17623	2	82.0	150	100	1	1	0	0	1	1
4	17474	1	56.0	100	60	1	1	0	0	0	0

FIGURE 18 – Sélection des données pour la modélisation

J'ai pris 70% des enregistrements pour l'apprentissage et 30% pour la validation.

2.4 Modélisation

La quatrième étape est la modélisation qui est au cœur de tout projet d'apprentissage automatique.

Cette étape est responsable des résultats qui devraient satisfaire ou aider à atteindre les objectifs du projet. Bien que ce soit la partie glamour du projet, c'est aussi la plus courte dans le temps.

2.4.1 Sélection des techniques de modélisation

Il y' a plusieurs types de modélisation. Le choix du modèle le plus adéquat sera généralement basé sur les critères suivants :

— Les types de données disponibles pour l'exploration : Les champs sont numériques.

— Les objectifs de data mining : Notre problème est un problème de classification supervisé. Nous avons plusieurs méthodes de classification supérvisé : k plus proches voisins, Les forêt aléatoires, support vecot machine, Régression logistique, ... Nous avons vu quelques méthodes dans le cours et d'autres dans le TP.

On va entraîner tous ces modèles et on va choisir celle qui donne la meuilleuire accuracy.

2.4.2 Explication des techniques de modélisation

— Random Forests : C'est un algorithme qui effectue l'apprentissage sur plusieurs arbres de décision entrainées sur des sous ensembles du jeu de données différents.
— Support Vector Machines : Ce sont une famille d'algorithmes permettant de résoudre des problèmes de classification, de régression et de détection d'anomalie. Ils ont pour but de séparer les données en classes à l'aide d'une frontière de telle façon que la distance entre les différents groupes de données et la frontière qui les sépare soit maximale.
— Gradient Boosting : est une technique qui permet de résoudre des problèmes de classification et de régression. Elle donne un poids à chaque individu et au fur à mesure les poids sont corrigés.
— KNN : L'algorithme de plus proche voisins permet de construire des ensemble d'invidus à partir des données d'apprentissage et pour chaque nouveau individu on cherche la plus proche classe pour lui a associe cette classe.
— Gaussian naive bayes : C'est un classifieur linéaire basé sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses.
— Régression logistique : C'est un modèle de régression binomiale permet d'associer à un vecteur de variables aléatoires une variable aléatoire binomiale.

2.4.3 Construire le modèle

En utilisant la fonction fit, j'ai pu entrainer tous les modèles.

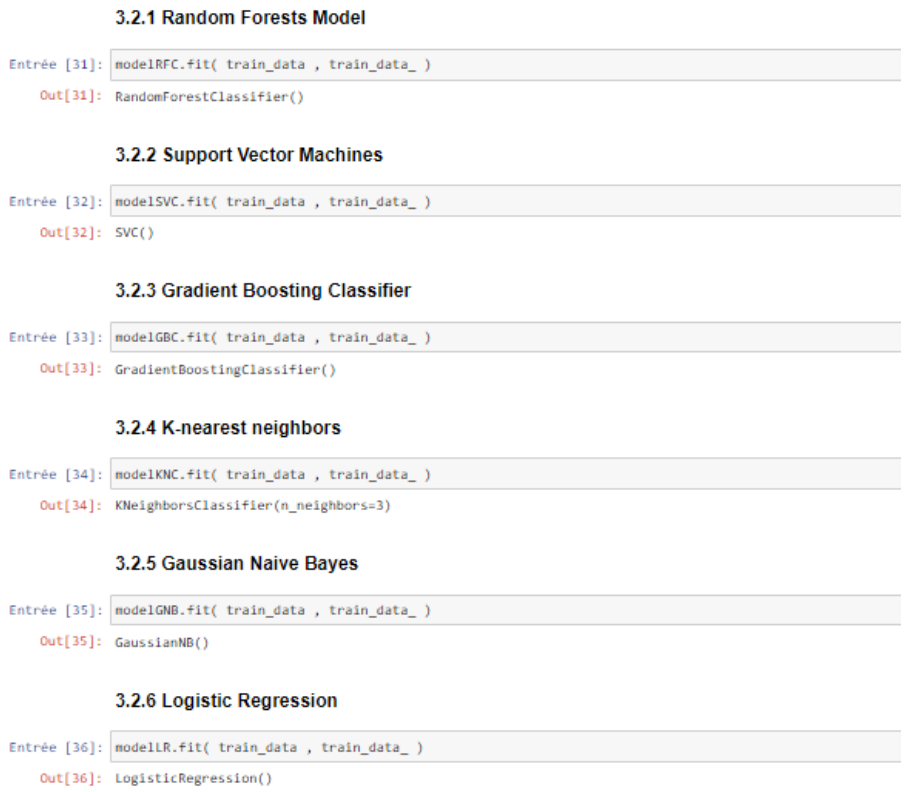

FIGURE 19 – la construction des modèles

2.5 Evaluation

2.5.1 Comparaison des modèles

J'ai utilisé 70% pour faire l'apprentissage du modèle et le 30% des données on va l'utiliser pour valider le modèle. On a visualiser l'accuricy de tous les modèle pour qu'on puisse faire une comparaison et choisir le meuilleur modèle.

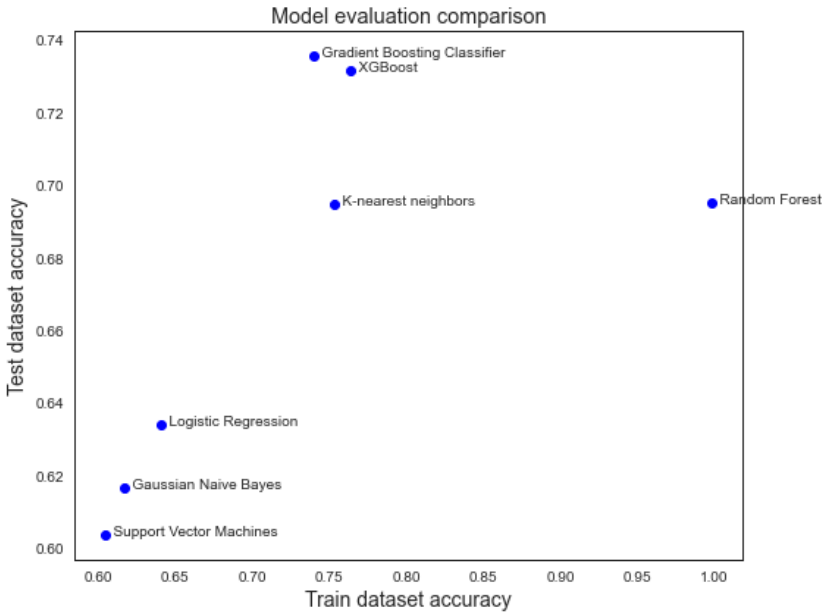

FIGURE 20 – Comparaison des modèles

Le modèle qui donne une valeur d'accuracy la plus haute est Gradiant boosting Classifier.

2.5.2 Amélioration

J'ai essayé plusieurs paramètres pour les deux algorithmes KNN et random forest et aussi un autre algorithme qui est XGBoost. Et dans tous les cas c'était l'algorithme Gradiant boosting Classifier qu'avait toujours le meilleurs score.

2.5.3 Importance des variables

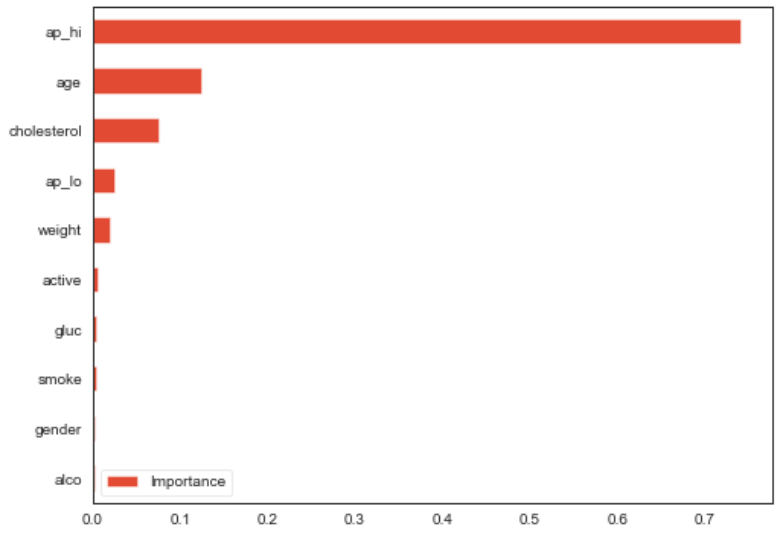

FIGURE 21 – Importance des variables

C'est la pression artérielle systolique qui joue un grand rôle pour prédire la variable cible selon le modèle de Gradiant boosting Classifier.

2.6 Déploiement

C'est la dernière étape de CRISP-DM. J'ai enregistré les résultats de la validation (identifiant de la personne et la valeur de la variable cible).

3 Topic Modeling

Dans cette partie, on va faire le topic modeling qui consiste à savoir le sujet ou le thème abstrait dans un document.

3.1 Prétraitement des données

On va suivre les huit étapes vu dans les séances de TP depuis la Tokenisation jusqu'à Chunking. Je vais utiliser seulement la colonne texte puisque c'est elle qui donne plus d'informations.

3.1.1 Tokenisation

La tokenisation est le process de découper un texte en tokens(mots). C'est la fonction word_tokenize() qui permet de faire cette tâche.

```
4250624

Out[30]: ['If',
          'I',
          'smelled',
          'the',
          'scent',
          'of',
          'hand',
          'sanitizers',
          'today',
          'on',
          'someone',
          'in',
          'the',
          'past',
          ',',
          'I',
          'would',
          'think',
          'they',
```

FIGURE 22 – Tokenisation

Le document contient 4250624 mots.

3.1.2 Fréquence des mots

FreqDist nous permet d'avoir la fréquence de chaque mots et la fonction most_common donne les plus fréquents mots dans le document.

```
Out[33]: [('#', 266985),
          (':', 208423),
          ('https', 177118),
          ('the', 87362),
          ('@', 85902),
          (',', 80194),
          ('COVID19', 77745),
          ('.', 76884),
          ('to', 71257),
          ('of', 57334)]
```

FIGURE 23 – Fréquence des mots

Il parait que parmi les tokens les fréquents on trouve : « : », « the », « to » et « of ». D'où vient l'importance de l'étape de stop words.

3.1.3 Stemming : Normalisation des mots

Dans cette étape en associe chaque mot à sa racine et on supprime les terminologies. Il existe plusieurs méthodes pour le stemming. J'ai choisi la fonction stem.

```
Out[38]: ['If',
          'I',
          'smell',
          'the',
          'scent',
          'of',
          'hand',
          'sanit',
          'today',
          'on',
          'someon',
          'in',
          'the',
          'past',
          ',',
          'I',
          'would',
          'think',
          'they',
          'were',
```

FIGURE 24 – Stemming

Par exemple, on peut voir que le mot « smelled » est remplacé par « smell ».

3.1.4 Lemmatisation : Base des mots

La lemmatisation est par définition une action consistant à l'analyse lexicale d'un texte avec pour but de regrouper les mots d'une même famille. La lemmatisation prend en compte : le genre (masculin ou féminin), le nombre (singulier ou pluriel), la possession (moi, toi, eux…) et le mode (indicatif, impératif…). La fonction lemmatize permet de faire ce travail.

3.1.5 Stop words

Dans cette étape on supprime les mots inutiles comme par exemple : « the », « a », « : ». Mais avant j'ai mis tous les mots en minuscule.

```
Out[75]: ['smell',
          'scent',
          'hand',
          'sanit',
          'today',
          'someon',
          'past',
          'would',
          'think',
          'intox',
          'that…',
          '//t.co/qzvybrogb0',
          'hey',
          'yanke',
          'yankeespr',
          'mlb',
          'would',
          "n't",
          'made',
          'sen',
```

FIGURE 25 – Stop words

3.1.6 Part of speech tagging (POS)

Dans cette partie on attribue à chaque mot sa nature (un nom, un verbe, un adjectif,...).

```
[('smell', 'NN')]
[('scent', 'NN')]
[('hand', 'NN')]
[('sanit', 'NN')]
[('today', 'NN')]
[('someon', 'NN')]
[('past', 'NN')]
[('would', 'MD')]
[('think', 'NN')]
[('intox', 'NN')]
[('that…', 'NN')]
[('//t.co/qzvybrogb0', 'NN')]
[('hey', 'NN')]
[('yanke', 'NN')]
[('yankeespr', 'NN')]
[('mlb', 'NN')]
[('would', 'MD')]
[("n't", 'RB')]
[('made', 'VBN')]
[('sen', 'NN')]
```

FIGURE 26 – Part of speech tagging (POS)

3.1.7 Reconnaissance des différentes entités

Cette étape consiste à organiser les mots dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.

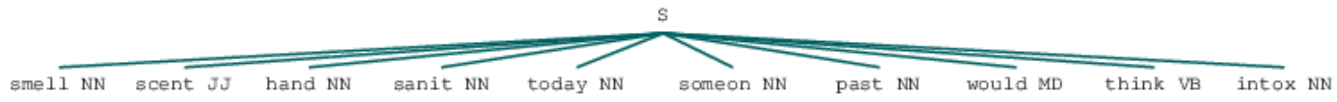

FIGURE 27 – Named entity recognition

3.1.8 Chunking

Le Chunking est un processus d'extraction de phrases à partir d'un texte non structuré, ce qui signifie analyser une phrase pour identifier les constituants (groupes de noms, verbes, groupes de verbes, etc.).

```
(S
  (NP smell/NN)
  (NP scent/JJ hand/NN)
  (NP sanit/NN)
  (NP today/NN)
  (NP someon/NN)
  (NP past/NN)
  would/MD
  think/VB
  (NP intox/NN))
```

FIGURE 28 – Chunking

3.2 Visualisation

Le nuage de mots est une méthode de visualisation qui permet d'avoir les mots les plus fréquents dans un document. Plus le mot est fréquent plus il est de grand taille.

FIGURE 29 – Nuage de mots

Cette représentation nous donne une idée sur les mots les plus utilisé dans les tweets sur covid. On trouve : covid19, covid, pandem, mask, ...

3.3 Topic modeling avec LDA

LDA est une méthode non supervisée de Natural Language Processing. Il permet de présenter un documents sous forme de thèmes pour le systhétiser.

3.3.1 La fréquence de mots

La première étape consiste à construire un dictionnaire où la clé est la fréquence du mot et la valeur est le mot.

```
smell 155
scent 13
hand 1189
sanit 479
today 5227
someon 842
past 672
would 2491
think 2422
intox 2
that… 279
```

FIGURE 30 – La fréquence des mots

3.3.2 Gensim filters

J'ai gardé seulement les 100000 mots plus fréquents qui apparait dans moins de 15 documents et plus que un seul document.

3.3.3 Gensim doc2bow

Dans cette phase, on attribue à chaque mot dans un document sa fréquence.

```
Word 8 ("would") appears 1 time.
Word 9 ("made") appears 1 time.
Word 10 ("player") appears 1 time.
Word 11 ("respect") appears 1 time.
Word 12 ("yanke") appears 1 time.
```

FIGURE 31 – Gensim doc2bow

Ce modèle a génèré 9 ensembles de document. Pour chaque ensemble j'ai affiché la probabilité de TF-IDF.

3.3.4 Running LDA using Bag of Words

J'ai choisi d'extraire 6 topics et pour chaque topic d'afficher les 10 plus fréquents mots-clés. Un nuage de mots a donné le résultat suivant :

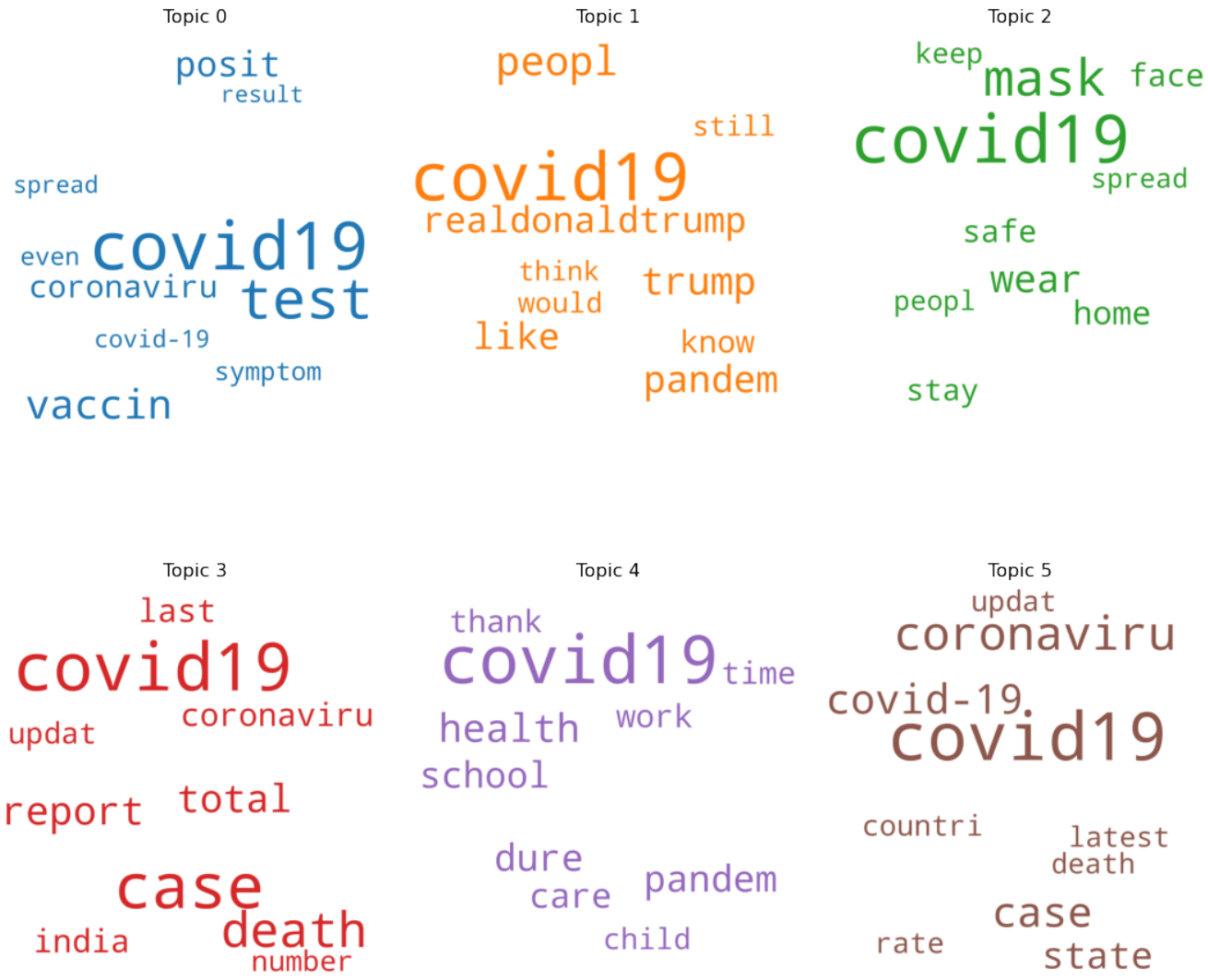

FIGURE 32 – Running LDA using Bag of Words

3.3.5 TF-IDF

Terme Frequency - Inverse Document Frequency (TF-IDF). Cette méthode consiste à rassembler un ensemble de documents partageant les mêmes mots clés. Deux documents ont plus de chance d'être dans le même ensemble s'ils possèdent des mots-clés en une grandes fréquence, et que ces mots sont rares dans d'autres documents.

```
[(0, 0.31200066919227304),
 (1, 0.34386088345342114),
 (2, 0.36804123747488343),
 (3, 0.43730975990531323),
 (4, 0.33144355605282877),
 (5, 0.3978718001045878),
 (6, 0.2666917498141616),
 (7, 0.21853705022170467),
 (8, 0.26585967117591836)]
```

FIGURE 33 – TF-IDF

Ce modèle a génèré 9 ensembles de document. Pour chaque ensemble j'ai affiché la probabilité de TF-IDF.

3.3.6 Running LDA using TF-IDF

De même pour TF-IDF, j'ai selectionné 6 topics et affiché 10 mots pour chaque topic.

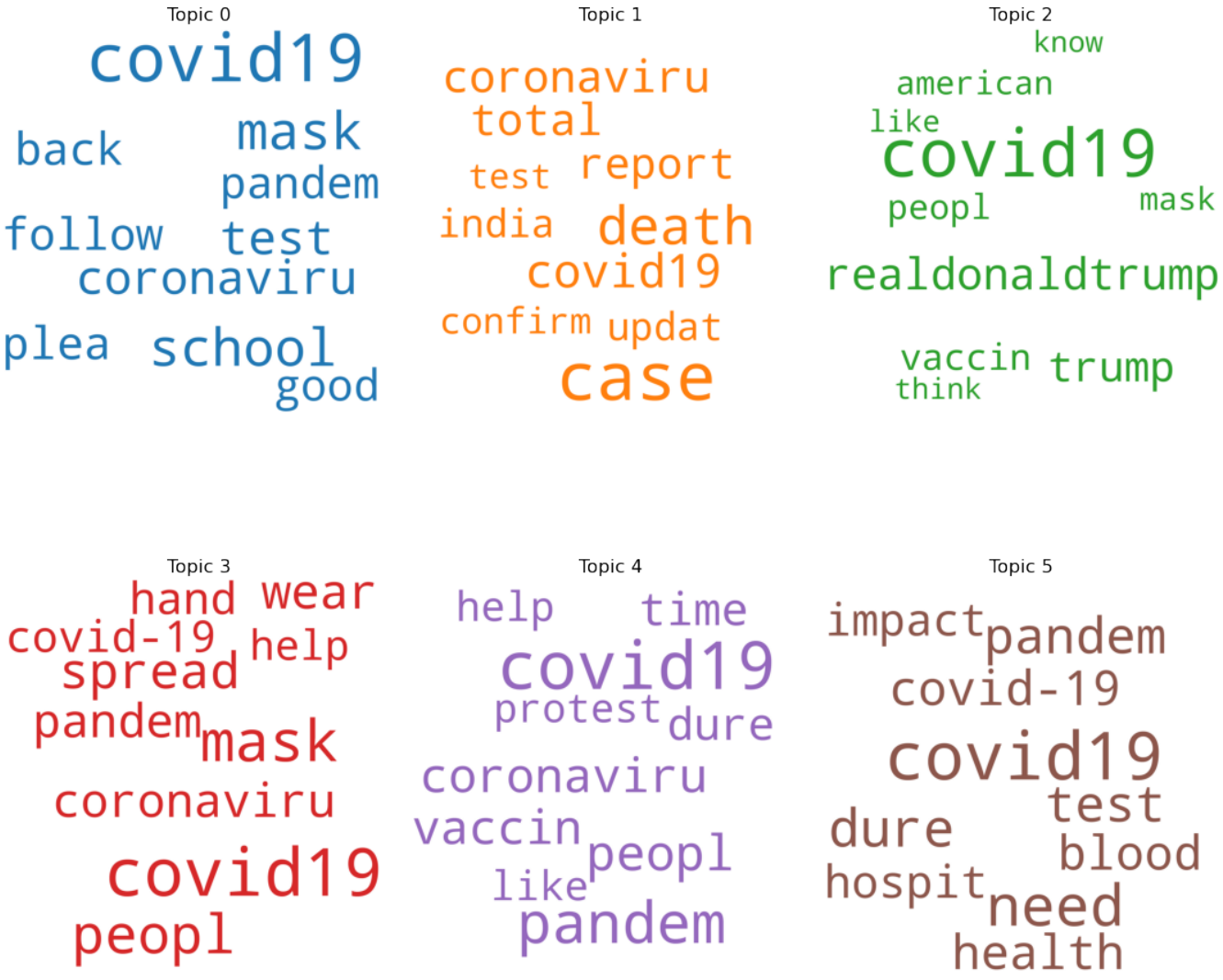

FIGURE 34 – LDA en utilisant TF-IDF

3.3.7 Evaluation du modèle

J'ai évalué le même document avec les deux types de modèle. Et j'ai remarqué que les deux algorithmes ont asssocié le document à deux topics différents.

Conclusion

A travers ce projet, j'ai pu en effet consolider mes connaissances acquises et d'enrichir mon expérience en matière de data mining. A travers les trois sujets traités, on constate que le date mining peut être appliqué dans tous les domaines et s'avère utile pour tous les décideurs.

Pour atteindre notre objectif j'ai mis en place la démarche CRISP-DM (CRoss Industry Standard Process for Data Mining). Il s'agit d'un modèle de processus de data mining qui décrit une approche communément utilisée par les experts en data mining pour résoudre les problèmes qui se posent à eux et les algorithme de NLP.