

# Projet Fil Rouge

**Comment peut-on expliquer et prédire l'apparition de la dermite radique après la radiothérapie ?**

Réalisé par : Farah ABDELLI



---

# TABLE DES MATIÈRES

<b>LISTE DES FIGURES</b>	<b>iii</b>
<b>LISTE DES ABRÉVIATIONS</b>	<b>iv</b>
<b>Résumé</b>	<b>v</b>
<b>INTRODUCTION</b>	<b>1</b>
<b>1 Présentation et compréhension des données</b>	<b>2</b>
1.1 Présentation des données . . . . .	3
1.2 Compréhension des données . . . . .	4
1.2.1 Typage des variables . . . . .	4
1.2.2 Modalités des variables . . . . .	5
1.2.3 Données manquantes . . . . .	6
1.3 Conclusion . . . . .	6
<b>2 Préparation des données</b>	<b>7</b>
2.1 Sécurisation des données . . . . .	8
2.1.1 Anonymisation des données . . . . .	8
2.1.2 Encodage des données . . . . .	8
2.2 Nettoyage des données . . . . .	9
2.3 Renommer les variables . . . . .	9
2.4 Imputation des valeurs manquantes . . . . .	10
2.5 Présentation de la Cohorte . . . . .	10
2.6 Conclusion . . . . .	11
<b>3 Modélisation</b>	<b>12</b>
3.1 Approche de validation . . . . .	13
3.2 Présentation de la méthode de modélisation . . . . .	13
3.3 Entraînement des modèles . . . . .	13
3.3.1 Forêt aléatoire . . . . .	13
3.3.2 Support vector machine . . . . .	14
3.4 Evaluation du modèle . . . . .	15

---

## TABLE DES MATIÈRES

---

3.4.1	Mesures de performance .....	15
3.4.2	Conclusion .....	17
<b>4</b>	<b>Analyses et critiques</b> .....	<b>18</b>
4.1	Analyse des résultats .....	19
4.1.1	Analyse technique.....	19
4.1.1.1	Modèle RF.....	19
4.1.1.2	Modèle SVM .....	19
4.1.1.3	Modèle approuvé .....	20
4.1.2	Analyse non technique.....	20
4.2	CONCLUSION .....	21
	<b>CONCLUSION GÉNÉRALE</b> .....	<b>22</b>
	<b>BIBLIOGRAPHIE</b> .....	<b>22</b>
	<b>ANNEXES</b> .....	<b>24</b>
A.1	Exemple de modalité des variables .....	24
A.2	Technique de modélisation.....	24
A.2.1	Random Forest.....	24
A.2.2	Support vector machine .....	25
A.3	Règlage des hyperparamètres .....	26
A.4	Résultats .....	26
A.4.1	Forêt aléatoire .....	26
A.4.2	Support vector machine .....	27

---

# LISTE DES FIGURES

1.1	Données fournies par les oncologues.....	3
1.2	Dictionnaire détaillé des données .....	4
1.3	Typage des variables.....	5
1.4	Modalités des variables (exemple dermite).....	5
1.5	Données manquantes .....	6
2.1	Déchiffrement d'une partition système .....	9
2.2	Cohorte .....	11
3.1	Entraînement du modèle RF avec GridSearch (code) .....	14
3.2	Entraînement du modèle RF avec les meilleurs paramètres (code).....	14
3.3	Entraînement du modèle SVM avec GridSearch (code) .....	14
3.4	Entraînement du modèle SVM avec les meilleurs paramètres SVM (code) .....	14
3.5	Matrice de confusion [2].....	15
A.1	Modalités des variables (exemple tabagisme).....	24
A.2	Modèle des forêts aléatoires .....	25
A.3	Modèle du SVM .....	25
A.4	Accuracy et F1-score du modèle RF.....	26
A.5	Matrice de confusion du modèle RF .....	26
A.6	Importance des caractéristiques (modèle RF).....	27
A.7	Accuracy et F1-score du modèle SVM .....	27
A.8	Matrice de confusion du modèle SVM .....	27
A.9	Importance des caractéristiques (modèle SVM) .....	28
A.10	Courbe ROC pour comparer les modèles.....	28



---

# LISTE DES ABRÉVIATIONS

**TP** True Positive

**TN** True Negative

**FP** False Positive

**FN** False Negative

**ML** Machine Learning

**SVM** Support Vector Machine

**RF** Random Forest

# Résumé

Le travail présenté dans ce rapport entre dans le cadre d'un projet académique dont l'objectif est d'utiliser la science des données afin de proposer un modèle capable d'expliquer et ou de prédire une/plusieurs toxicité(s) à l'aide des données fournies par les oncologues.

Le principal défi est de bien comprendre les données et d'assurer un bon pré-traitement de données avant d'attaquer la partie analyse et prédiction.

Je propose donc une solution complète qui commence par le nettoyage des données à l'aide de plusieurs techniques de préparation des données avant de les analyser. Cela fournit une base solide et fiable pour m'aider dans les prochaines phases de notre projet.

Ensuite, je propose deux modèles d'apprentissage automatique : le modèle des forêts aléatoires et le modèle SVM, deux modèles de classification supervisée très utilisés dans le domaine médical.

Ces méthodes ont été testées sur l'ensemble de données déjà nettoyé dans l'étape précédente et évaluées par des outils de mesures de performances que je les considère strictes afin d'avoir un résultat précis.

Finalement, je résume et présente des perspectives de chacune de mes réalisations.



---

# INTRODUCTION

En santé comme dans bien d'autres domaines, les progrès technologiques ont fait exploser la quantité d'informations recueillies à chaque instant. Ainsi, si dix ans ont été nécessaires pour obtenir la première séquence d'un génome humain, en 2003, il faut aujourd'hui moins d'une journée pour parvenir au même résultat. Cette accélération technologique fait croître le volume de données disponibles de manière exponentielle. Une aubaine pour la recherche en santé pour qui le big data est une source presque inépuisable de nouvelles connaissances, indispensables à l'innovation et aux progrès médicaux.

Aujourd'hui on s'intéresse à la radiothérapie mammaire qui fait partie de la prise en charge thérapeutique du cancer du sein. Ce traitement peut entraîner des effets secondaires immédiats ou retardés. Même si les nouvelles techniques utilisées limitent l'apparition des séquelles après l'irradiation, le risque de complications est tout de même présent.

Ces effets secondaires différés apparaissent plusieurs mois après la fin des séances de radiothérapie, voire plus tard. Comme pour les effets secondaires précoces, ceux-ci varient en fonction de la zone irradiée, de la dose délivrée, de l'âge du patient ainsi que de sa sensibilité aux rayons. Les machines et techniques utilisées aujourd'hui limitent fortement le risque d'apparition de ces séquelles tardives.

Dans le cadre du projet Fil Rouge, mon objectif est de répondre à cette problématique :  
Comment peut-on expliquer et prédire l'apparition de la dermite radique après la radiothérapie ?

Chapitre

**1**

---

# Présentation et compréhension des données

## Sommaire

---

<b>1.1</b>	<b>Présentation des données .....</b>	<b>3</b>
<b>1.2</b>	<b>Compréhension des données .....</b>	<b>4</b>
1.2.1	Typage des variables .....	4
1.2.2	Modalités des variables .....	5
1.2.3	Données manquantes .....	6
<b>1.3</b>	<b>Conclusion .....</b>	<b>6</b>

---



# PRÉSENTATION ET COMPRÉHENSION DES DONNÉES

Dans tous les projets de science des données, les premières questions qui se posent : quelles données nous aurons besoin ? C'est propre ? Et maintenant il faut répondre. Dans ce premier chapitre, comme l'indique son nom, nous présentons nos ensembles de données accompagnés d'une analyse des variables afin de les comprendre et de vérifier leurs qualités.

## 1.1 Présentation des données

Ci-dessous l'ensemble de données préparées par les oncologues où chaque enregistrement correspond à un patient qui a subi à une radiothérapie. Cet ensemble se compose de 2847 lignes et 1628 colonnes.

Chaque ligne contient différentes informations du patient :

- Informations personnelles (exp : id, sexe, âge, poids, taille ...), - Informations sur le diagnostic du cancer (NomDiagnostic,LateraliteDiagnostic,type,StadeClinique ...),
- Informations liées au traitement (dosesPréscrites,traitementAntalgiques,traitement\_insuline ...),
- Informations sur les différentes toxicités (dermite\_radique,hyperpigmentation,oedème,douleurs ...),et bien d'autres.

Ce fichier est divisé en deux sous parties :

- a) La première résume la totalité des éléments rencontrés pour chaque colonne (colonnes avec écrit TOUTE\_BASE)
- b) La seconde est la saisie des données de façon temporelle.

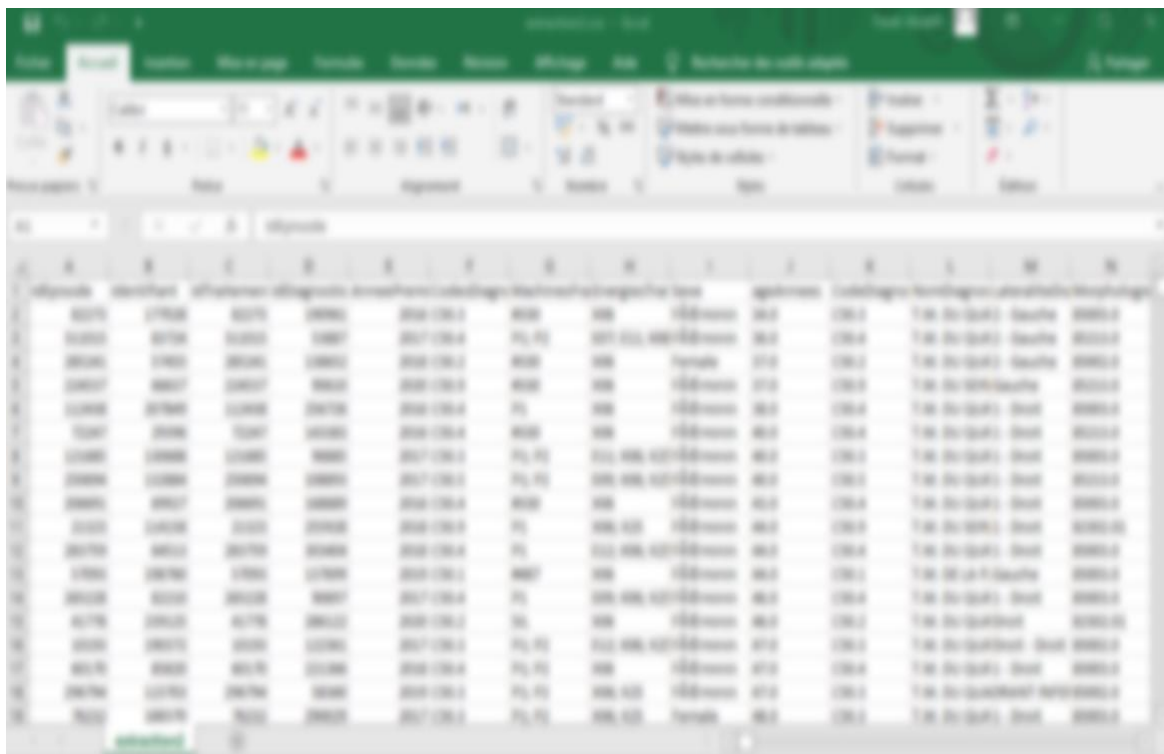
The image shows a screenshot of an Excel spreadsheet. The top part of the image shows the Excel ribbon with the 'Formules' (Formulas) tab selected. Below the ribbon, the spreadsheet grid is visible, showing a large number of columns and rows. The columns are labeled with various patient information, including identifiers, dates, and medical data. The rows contain individual patient records. The data is organized in a structured manner, with headers for different categories of information.

FIGURE 1.1 – Données fournies par les oncologues

# PRÉSENTATION ET COMPRÉHENSION DES DONNÉES

Ce jeu de données est accompagné par un dictionnaire détaillé décrivant la plupart des termes que nous pourrions rencontrer. Au début, c'était un peu compliqué de saisir la définition de chaque variable ainsi que son rôle étant donné qu'elles sont nombreuses et de même il y a un manque de connaissances dans ce métier. Mais ce fichier nous a aidé à bien comprendre chaque entrée ainsi que ses valeurs possibles.

Dictionnaire .XLSX									
File Edit View Insert Format Data Tools Help Last edit was on January 30									
100% \$ % .00 123 Default (Ca... 11 B I S A									
A1 fx Label_formulaire									
	A	B	C	D	E	F	G	H	I
1									
			Description	MedDRA v12.0 code	SNOMED CT (ConceptID)	Type_item	0	1	2
2	Label_formulaire								
3	Délai depuis la dernière radiothérapie	x	x		Données				
4	Performance status	x	x		Table	Asymptoma	Symptomatique	Symptomatique, alité moins de 50 % de la journée.	Sy
5	Poids actuel (kg)	x	x		Données				
6	Bouffées de chaleur	ctcae 5.0	10020407		Table	non	symptômes léger	symptômes modérés, interférant avec les activités instrumentales de la vie quotidienne	syn
7	Arthralgies	ctcae 5.0	10003239		Table	non	douleur légère	douleur modérée, interférant avec les activités instrumentales de la vie quotidienne	do
8	Douleur pic site traité (0-10)	x	x		Données	0	1		2
9	Traitement antalgique	x	x		Table	non	pallier 1	pallier 2	pa
10	Douleur fond site traité (0-10)	x	x		Données	0	1		2
11	Chimiothérapie/Biothérapie en cours	x	x		Données	non	Anastrozole 1 mg	Letrozole 2.5 mg/jr	Ex
12	Hormonothérapie en cours (sein)	x	x		Table				
13	Tabagisme : nb cigarette/jour (Tabagis)	x	x		Données				
14	Amplitude articulaire (ipsilatéral)	ctcae 5.0	10048706		Table	pas de dimi	perte ≤ 25 % de l	diminution > 25-> 50 %, interférant avec les activités instrumentales de la vie quotidienne	dir
15	Lymphoedème membre (ipsilatéral)	x	10050266		Table	Non	Oui		
16	Port d'une contention pour un lympho	HYPOG-m	x		Table	non	utilisation occasi	utilisation 1 à 3 fois par mois	uti
17	P 15cm proximal (ipsilat)	HYPOG	x		Données				
18	P 15cm proximal (controlat)	HYPOG	x		Données				
19	P 10cm distal (ipsilat)	HYPOG	x		Données				
20	P 10cm distal (controlat)	HYPOG	x		Données				
21	Toux	ctcae 5.0	10011224		Table	non	symptômes léger	symptômes modérés, nécessitant un traitement médical, interférant avec les activités quotidiennes	sy
22	Dyspnée	ctcae 5.0	10013963		Table	non	essoufflement lo	essoufflement lors d'un effort minime interférant avec les activités instrumentales de la vie quotidien	es
23	Fibrose pulmonaire	ctcae 4.0	10037383		Table	non	Hypoxémie légèr	Hypoxémie modérée ; signes d'hypertension pulmonaire ; fibrose pulmonaire à la radiographie 25-5	Hy
24	Hypoxie	ctcae 5.0	10021143		Table	non	Saturation en oxygène diminuée avec l'exercice (ex : oxymètre <88 %) ; nécessite une oxygénothérap	Sa	
25	Pneumopathie organisée cryptogéniqu	x	x		Table	Non	Oui	Suspectée, des investigations supplémentaires sont nécessaires	
26	Pneumopati (pneumonite)	ctcae 5.0	10035742		Table	Non	Asymptomatique	Symptomatique ; nécessitant un traitement médical ; interférant avec les activités instrumentales de	Sy
27	Péricardite	ctcae 5.0	10034484		Table	non	Asymptomatique	Péricardite symptomatique (ex : douleur thoracique)	pé
28	Infarctus du myocarde	ctcae 4.0	10028596		Table	Non	Asymptomatique et enzymes cardiaques subnormales et absence de signe d'ischémie à l'ECG		Sy
29	Insuffisance cardiaque	ctcae 4.0	10019279		Table	Non	Asymptomatique	Symptomatique lors d'un effort léger ou modéré	Sé
30	Statut ménopausique	x	x		Table	Ménopausé	Péri ménopause	Non ménopausée (règles régulières)	
31	Sécheresse vaginale	ctcae 5.0	10046904		Table	Non	Sécheresse vagin	Sécheresse vaginale modérée interférant avec les fonctions sexuelles ou provoquant une gêne fréqui	Sé
32	Reprise d'une activité sexuelle	x	x		Table	Non	Oui		
33	Dermite radique	ctcae 5.0	10061103		Table	Non	Faible érythème	Érythème modéré à vif ; desquamation suintante en plaques, affectant principalement les plis et repi	De

FIGURE 1.2 – Dictionnaire détaillé des données

## 1.2 Compréhension des données

Avant de commencer quoi que ce soit, une étape de compréhension des données est primordiale afin de vérifier leurs qualités et de préparer le terrain favorable pour le travail.

### 1.2.1 Typage des variables

Comme indiqué ci-dessous, la plupart des variables sont qualitatives (catégorielle) avec un pourcentage de 84,7% ce qui nous a empêché de réaliser une étude de liaison (corrélation) entre toutes les variables dans l'étape suivante.

## PRÉSENTATION ET COMPRÉHENSION DES DONNÉES

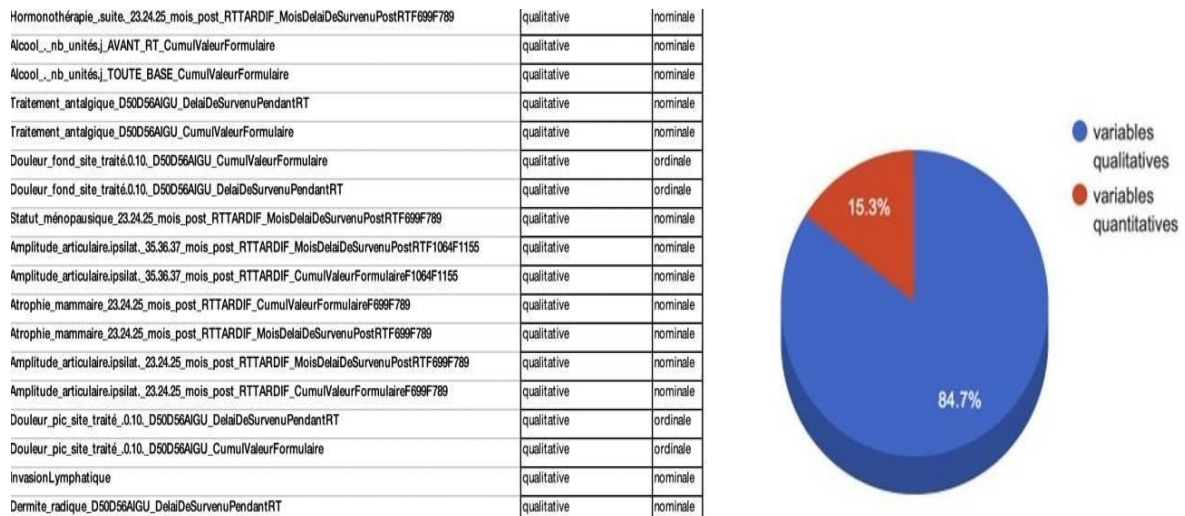


FIGURE 1.3 – Typage des variables

### 1.2.2 Modalités des variables

Ensuite, nous avons déterminé les modalités des différentes variables catégorielles afin de connaître les valeurs que peuvent prendre celles-ci. Pour chaque valeur on associe son effectif ainsi que sa fréquence et un petit graph qui résume les résultats comme l'indique l'exemple ci-dessous :

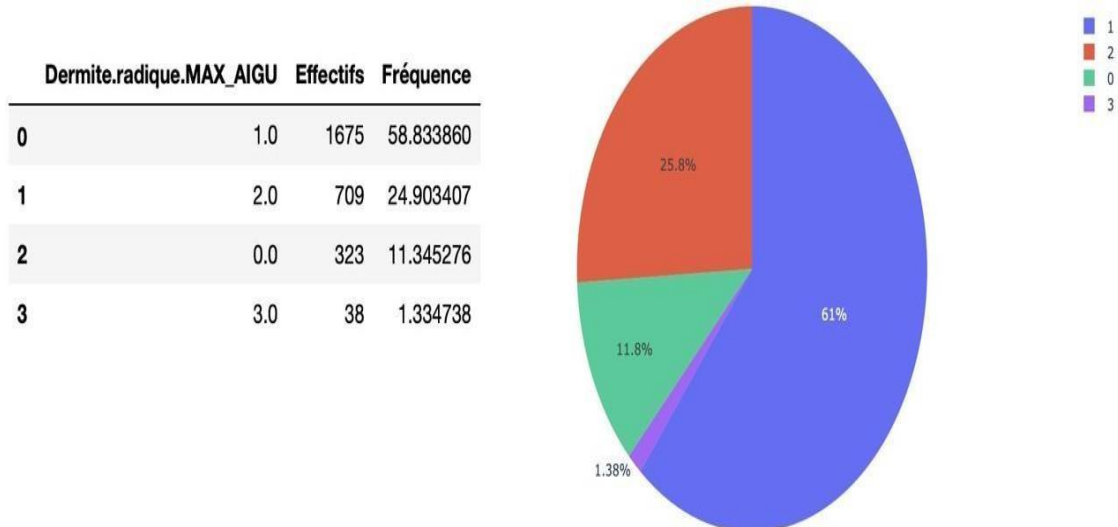


FIGURE 1.4 – Modalités des variables (exemple dermite)

## PRÉSENTATION ET COMPRÉHENSION DES DONNÉES

### 1.2.3 Données manquantes

Nous avons trouvé 1266 variables soit 77% des variables ayant plus de 85% de données manquantes ce qui est un pourcentage très élevé.

	Missing Values	% of Total Values
P100mm_distal_controlat_59.60.61.62_mois_post_RTTARDIF_MoisDelaiDeSurvenuPostRTF1794F1915	2847.0	100.0
P150mm_proximal_controlat_1_mois_post_RTTARDIF_MoisDelaiDeSurvenuPostRTF23F37	2847.0	100.0
Reprise_activité_sexuelle_41.42.43_mois_post_RTTARDIF_MoisDelaiDeSurvenuPostRTF1246F1338	2847.0	100.0
P150mm_proximal_controlat_41.42.43_mois_post_RTTARDIF_CumulValeurFormulaireF1246F1338	2847.0	100.0
P150mm_proximal_controlat_29.30.31_mois_post_RTTARDIF_MoisDelaiDeSurvenuPostRTF881F973	2847.0	100.0
...	...	...
Tabagisme_._cigarettes.j_D43D49AIGU_DelaiDeSurvenuPendantRT	2462.0	86.5
Tabagisme_._cigarettes.j_D43D49AIGU_CumulValeurFormulaire	2462.0	86.5
Traitement_prescrit_D1D7AIGU_DelaiDeSurvenuPendantRT	2445.0	85.9
Traitement_prescrit_D1D7AIGU_CumulValeurFormulaire	2445.0	85.9
Traitement.prescrit.MAX_AIGU	2431.0	85.4

1266 rows x 2 columns

FIGURE 1.5 – Données manquantes

## 1.3 Conclusion

Dans ce chapitre, nous avons présenté l'ensemble des données avec quelques descriptions, une analyse des variables a été faite afin de comprendre chaque attribut. Mais tout ce que nous économisons n'est pas suffisant pour réaliser une prédiction parfaitement précise, car toutes les données ne sont pas pertinentes. La plupart des variables sont catégorielles, il y a plusieurs données manquantes et des textes bruts qui contiennent beaucoup d'aléatoire qui affectent à l'estimation des modèles : les accents, les minuscules, les signes de ponctuation, les caractères spéciaux... Tout cela nous amène à une prochaine étape, très importante dans projets de science des données qui est la préparation de données.

Chapitre

**2**

---

# Préparation des données

## Sommaire

---

<b>2.1</b>	<b>Sécurisation des données .....</b>	<b>8</b>
2.1.1	Anonymisation des données.....	8
2.1.2	Encodage des données.....	8
<b>2.2</b>	<b>Nettoyage des données .....</b>	<b>9</b>
<b>2.3</b>	<b>Renommer les variables.....</b>	<b>9</b>
<b>2.4</b>	<b>Imputation des valeurs manquantes .....</b>	<b>10</b>
<b>2.5</b>	<b>Présentation de la Cohorte.....</b>	<b>10</b>
<b>2.6</b>	<b>Conclusion .....</b>	<b>11</b>

---

## PRÉPARATION DES DONNÉES

Comme mentionné dans le chapitre précédent, notre problème repose sur les données médicales fournies par les oncologues. Ainsi, nos entrées sont sous le format brut et ayant plusieurs données manquantes donc elles nécessitent un traitement spécifique. Nous terminons ce chapitre par l'approche de validation qui est une étape nécessaire avant de commencer l'entraînement.

### 2.1 Sécurisation des données

Les données de santé sont des données à caractère personnel particulières car considérées comme sensibles. Elles font à ce titre l'objet d'une protection particulière par les textes (règlement européen sur la protection des données personnelles). Pour cela, il faut appliquer des techniques d'anonymisation de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.

#### 2.1.1 Anonymisation des données

Afin d'assurer la protection des données personnelles de tous les patients, nous avons utilisé quelques techniques d'anonymisation comme indiqué ci-dessous :

- La substitution pour les variables comme Id des patients : remplacer une valeur par une autre.
- Ajouter (+/-) 5cm à la taille des patients.
- La randomisation pour les variables comme AgeDebutTraitement : remplacer un âge par une tranche d'âge .
- Changer les valeurs de la colonne Sexe Féminin par 1 et Masculin par 0.

#### 2.1.2 Encodage des données

Séparément, nous avons recours à un outil de cryptage "VeraCrypt" qui permet de chiffrer et déchiffrer nos données en toute sécurité. Ce dernier permet de créer des conteneurs chiffrés, de chiffrer une partition d'un disque ou d'un périphérique externe et également de chiffrer une partition système Windows et même l'intégralité du disque principal.

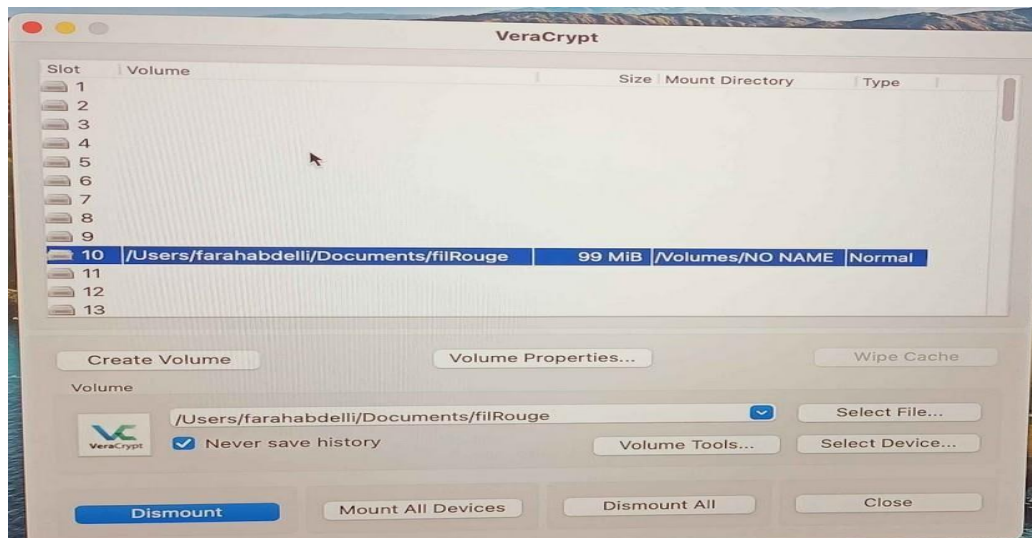


FIGURE 2.1 – Déchiffrement d'une partition système

## 2.2 Nettoyage des données

Passons à la partie de nettoyage qui consiste alors à identifier et corriger les données qui sont inexactes, altérées ou non pertinentes. Il s'agit d'une étape essentielle dans le traitement des données pour améliorer la cohérence, la fiabilité ainsi que les valeurs des informations à exploiter.

Les étapes réalisées pour faire le nettoyage sont :

- Suppression des ID manquants
- Suppression des lignes dupliquées
- Suppression des caractères spéciaux
- Suppression des colonnes ayant un pourcentage supérieur à 85% de valeurs manquantes

## 2.3 Renommer les variables

Nous avons remarqué que les noms des variables contiennent des caractères spéciaux et des signes de ponctuation. Du coup, nous les avons renommées en des en-têtes plus simple et facile à les utiliser dans les prochaines étapes. Par exemple : "Dermite\_radique'\_D1D7AIGU\_CumulValeurFormulaire" devient "Dermite\_radique\_s1" qui veut dire dermite radique semaine 1

### 2.4 Imputation des valeurs manquantes

L'imputation est le processus utilisé pour attribuer des valeurs de remplacement aux valeurs manquantes, invalides ou incohérentes qui ont échoué aux vérifications. Afin d'aborder correctement l'imputation des données manquantes il faut en distinguer les causes, surtout si elles ne sont pas le simple fruit du hasard. Une typologie a été développée par Little & Rubin les répartissant en 3 catégories [5] :

- Une donnée est MCAR, c'est-à-dire manquante de façon complètement aléatoire
- Une donnée est MNAR, c'est-à-dire manquante de façon non aléatoire
- Une donnée est MAR, c'est-à-dire manquante de façon aléatoire

Ensuite, nous avons utilisé plusieurs techniques d'imputation selon la catégorie pour remplir toutes ces valeurs par exemple :

- L'imputation par valeur moyenne consiste à remplacer la valeur manquante ou incohérente par la valeur moyenne calculée à partir des unités répondantes ayant le même ensemble de caractéristiques prédéterminées.
- L'imputation par voisin le plus proche Dans ce cas, il faut élaborer une sorte de critère pour déterminer l'unité répondante qui ressemble le plus à l'unité ayant la valeur manquante, conformément aux caractéristiques prédéterminées. L'unité la plus proche de la valeur manquante est alors utilisée comme donneur.
- Suppression de la ligne si elle contient plus de 70% de cases vides

### 2.5 Présentation de la Cohorte

Le choix des variables est basé sur une étude de corrélation et nous nous sommes concentrés sur 3 catégories d'informations :

- Les informations personnelles/relatives au patient : id, âge, taille, poids, stade clinique, tabagisme ...
- Les informations liées au traitement/machine : taille de lésion, dose délivrée, énergie de la machine, MachinesFractionsChamps, anneepremierefraction ...
- Les informations relatives aux effets de la radiothérapie : dermite, douleurfond, douleurpic, performancestatus ..

Le choix des variables est basé sur une étude de corrélation entre les différentes variables et nous avons maintenant 30 variables d'où une réduction du pourcentage des variables qualitatives à 46%.

Pour les variables catégorielles temporelles comme la dermite radique, la douleur, le tabagisme ..., nous avons ignoré la notion du temps puisque nous n'avons pas les dates exactes et ensuite réalisé des agrégations afin de les transformer en une seule variable.

Par exemple : Dans le fichier transmis par les oncologues, il y a 23 colonnes de dermite radique

Solution : Faire une agrégation par le maximum et les transférer en une seule colonne dermite radique ayant 5 classes de 0 -> 4.



Identifiant
IdDiagnostic
MachinesFractionsChamps
NomDiagnostic
LateraliteDiagnostic
StadeCliniqueDiagnostic
TailleLesionInvasive
GradeHPDiagnostic
NombreGanglionsExaminees
NombreGanglionsNegatifs
NombreGanglionsPositifs
AnneePremiereFraction
ageAnnees
Poids_moyen
Taille_moyenne
IdMachine
Energie
DoseDelivree
grade dermite aigu
tabagisme aigu
grade douleurfond aigu
grade douleurpic aigu
performance status aigu
traitement antalgique aigu
grade dermite tardif
tabagisme tardif
grade douleurfond tardif
grade douleurpic tardif
performance status tardif
traitement antalgique tardif

FIGURE 2.2 – Cohorte

## 2.6 Conclusion

Les outils de préparation des données nous ont permis de nettoyer les données avant de les analyser. Cela fournit une base solide et fiable pour nous aider dans les prochaines phases de notre projet, en particulier la modélisation en permettant au projet de s'exécuter de manière plus rapide et plus transparente.

Chapitre

**3**

---

# Modélisation

## Sommaire

---

<b>3.1</b>	<b>Approche de validation.....</b>	<b>13</b>
<b>3.2</b>	<b>Présentation de la méthode de modélisation.....</b>	<b>13</b>
<b>3.3</b>	<b>Entrainement des modèles.....</b>	<b>13</b>
3.3.1	Forêt aléatoire.....	13
3.3.2	Support vector machine.....	14
<b>3.4</b>	<b>Evaluation du modèle .....</b>	<b>15</b>
3.4.1	Mesures de performance .....	15
3.4.2	Conclusion .....	17

---

A ce stade, diverses techniques de modélisation sont sélectionnées et appliquées, et leurs paramètres sont calibrés aux meilleures valeurs. Généralement, il existe plusieurs techniques pour le même type de problème d'exploration de données. Alors maintenant, nous nous tournons vers Le cœur du projet est de modéliser nos données, ce qui nous permettra d'atteindre nos principaux objectifs.

### 3.1 Approche de validation

L'approche de validation est Train-Test Split. L'ensemble de données doit être divisé en 2 parties. La première (la plus grande) de 80% est utilisée pour l'entraînement et le deuxième ensemble de 20% est utilisé pour tester les performances du modèle sélectionné. Les différents modèles peuvent être les mêmes d'un point de vue général, mais leurs hyper paramètres diffèrent. Donc, ils ne sont techniquement pas les mêmes et se comportent différemment.

### 3.2 Présentation de la méthode de modélisation

Nous avons un échange à long terme de recherche sur les algorithmes de classification. Ce que nous recherchons vraiment, ce ne sont que des algorithmes puissants. Les résultats étaient variables et c'était à nous de décider, d'où venait le choix de la forêt aléatoire et SVM (voir annexe A.2 pour plus de détails). Nous avons choisi ces deux modèles car en premier nous traitons un problème de classification (prédiction d'une classe) et ces deux modèles sont adaptés avec la classification des données en plus ils sont les plus utilisés dans le domaine médical.

### 3.3 Entraînement des modèles

#### 3.3.1 Forêt aléatoire

Cet exemple montre comment un classifieur est optimisé par validation croisée, qui est effectuée à l'aide de l'objet GridSearchCV sur un ensemble de développement qui ne comprend que 80% des données étiquetées disponibles. (voir annexe pour plus de détails sur la validation croisée avec GridSearch)

Les performances des hyper-paramètres sélectionnés et du modèle formé sont ensuite mesurées sur un ensemble d'évaluation dédié qui n'a pas été utilisé lors de l'étape de sélection du modèle.

Le résultat du GridSearch était : {'max\_depth' : 13, 'min\_samples\_leaf' : 2, 'n\_estimators' : 75}

```

#Créez un dictionnaire appelé rf_params et remplissez quelques paramètres
#pour les n_estimators, max_depth et min_samples_leaf
rf_params = {'n_estimators':np.arange(25,150,25),'max_depth':np.arange(1,15,2),
             'min_samples_leaf':np.arange(2,15,3)}

#n_jobs:Nombre de tâches à exécuter en parallèle. -1 signifie utiliser
#tous les processeurs.
#cv:Détermine la stratégie de fractionnement de la validation croisée.
#ici j'ai choisi 5 plis
gs_rf = GridSearchCV(clf,rf_params,cv=5,n_jobs=-1)
gs_rf.fit(x_train_scaled,y_train)
b = gs_rf.best_params_

```

FIGURE 3.1 – Entraînement du modèle RF avec GridSearch (code)

```

#Entraîner le modèle RF avec les meilleurs paramètres
RF = RandomForestClassifier(n_estimators=b['n_estimators'],max_depth=b['max_depth'],
                           min_samples_leaf=b['min_samples_leaf'],random_state=0)
model = RF.fit(x_train_scaled,y_train)
y_pred3 = model.predict(x_test_scaled)
print(classification_report(y_test, y_pred3))

```

FIGURE 3.2 – Entraînement du modèle RF avec les meilleurs paramètres (code)

### 3.3.2 Support vector machine

Pareil pour le modèle SVM, nous avons entraîné le modèle avec la technique d'optimisation GridSearchCV pour obtenir automatiquement les meilleurs paramètres.

Le résultat du GridSearch était : {'kernel' : 'rbf', 'C' : 100}

```

from sklearn.model_selection import GridSearchCV

#Créez un dictionnaire appelé param_grid et remplissez quelques paramètres
#pour les noyaux, C et gamma
param_grid = {'C': [0.1,1, 10, 100], 'gamma': [1,0.1,0.01,0.001],
              'kernel': ['rbf', 'poly', 'sigmoid']}
grid = GridSearchCV(SVC(),param_grid,refit=True,verbose=2)
#Créer un objet GridSearchCV et l'adapter aux données d'entraînement
grid.fit(x_train_scaled,y_train)
print(grid.best_estimator_)

```

FIGURE 3.3 – Entraînement du modèle SVM avec GridSearch (code)

```

#Entraîner le modèle SVM avec les meilleurs paramètres
grid_predictions = grid.predict(x_test_scaled)
print(confusion_matrix(y_test,grid_predictions))
print(classification_report(y_test,grid_predictions)) #Output

```

FIGURE 3.4 – Entraînement du modèle SVM avec les meilleurs paramètres SVM (code)

### 3.4 Evaluation du modèle

#### 3.4.1 Mesures de performance

L'évaluation d'un modèle consiste à mesurer les écarts entre les prédictions du modèle et les résultats attendus. Nous utilisons des indicateurs de précision pour mesurer la qualité du classificateur. Cela exprime le nombre de classes prédites correctes par rapport au nombre total de prédictions de classes effectuées. Nous choisissons d'utiliser des métriques strictes et nous considérons la sortie comme équitable uniquement lorsque toutes les catégories prédites sont correctes.

- **Matrice de confusion**

En apprentissage automatique supervisé, la matrice de confusion [2] est une matrice qui mesure la qualité du système de classification. L'un de ses avantages est qu'elle peut rapidement montrer si le système de classification est capable de classer correctement.

---

		Classe réelle	
		-	+
Classe prédite	-	True Negatives (vrais négatifs)	False Negatives (faux négatifs)
	+	False Positives (faux positifs)	True Positives (vrais positifs)

---

FIGURE 3.5 – Matrice de confusion [2]

#### a) La performance :

La performance ou la justesse (accuracy en anglais) est l'un des critères d'évaluation des modèles de classification. De manière informelle, elle fait référence à la proportion de prédictions correctes faites par le modèle. [4]

Equation :

$$\text{Justesse} = \frac{VP + VN}{VP + VN + FP + FN}$$

### b) Le rappel :

Le rappel représente la capacité du classificateur à trouver tous les échantillons de la classe grâce à sa prédiction.

[3]

Équation :

$$Rappel = \frac{TP}{TP + FN}$$

### c) La précision :

Par conséquent, nous nous intéresserons également à la précision qui mesure le ratio de prédictions précises pour chaque catégorie par rapport au nombre de prédictions faites pour chaque catégorie. [3]

Équation

$$Précision = \frac{TP}{TP + FP}$$

### d) F-score :

Nous pouvons calculer la « mesure F », afin d'évaluer le compromis entre rappel et précision, qui est leur moyenne harmonique.[3]

Equation :

$$F - mesure = 2 \times \frac{Précision \times Rappel}{Précision + Rappel} = \frac{2TP}{2TP + FP + FN}$$

#### • Courbe ROC

Une courbe ROC (receiver operating characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs. [7]

Le taux de vrais positifs (TVP) est l'équivalent du rappel. Il est donc défini comme suit :

$$TVP = \frac{VP}{VP + FN}$$

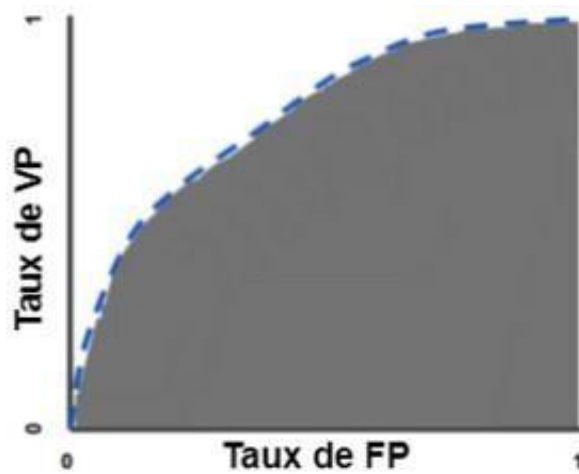
## MODÉLISATION

Le taux de faux positifs (TFP) est défini comme suit :

$$TFP = \frac{FP}{FP + VN}$$

- AUC

AUC signifie "aire sous la courbe ROC". Cette valeur mesure l'intégralité de l'aire à deux dimensions situées sous l'ensemble de la courbe ROC (par calculs d'intégrales) de (0,0) à (1,1).[7]



### 3.4.2 Conclusion

Cette phase était une transition obligatoire pour assurer un bon modèle qui nous permettrait enfin d'anticiper nos besoins de ce projet. Nous passons maintenant à l'étape suivante pour en finir une solution adaptée et efficace.

---

# Analyses et critiques

## Sommaire

---

<b>4.1</b>	<b>Analyse des résultats.....</b>	<b>19</b>
4.1.1	Analyse technique.....	19
4.1.2	Analyse non technique .....	20
<b>4.2</b>	<b>CONCLUSION .....</b>	<b>21</b>

---



La modélisation se fait en différentes itérations, il est temps de voir à quoi ressemblaient les différents résultats et quel est le résultat de notre analyse. Pour cela, nous utiliserons les outils de test dont nous avons parlé plus tôt.

### 4.1 Analyse des résultats

(Voir annexe A.4)

#### 4.1.1 Analyse technique

##### 4.1.1.1 Modèle RF

Dans les figures A.4 et A.5 dans l'annexe A.4, le modèle de classification RF était puissant et précis avec un accuracy de 92 % et un f1-score de 89% sur les données d'entraînement et un accuracy de 83 % et un f1-score de 81% sur les données de test ce qui est un bon signe.

Maintenant, en regardant la matrice de confusion, dans la ligne 0, il y a 238 patients et 174 des patients sont correctement attribués au grade 0, tandis que 60 sont incorrectement attribués au grade 1 et 4 sont incorrectement attribués au grade 2.

Dans la ligne 1, il y a 183 patients, mais seulement 166 sont correctement attribués au grade 1, tandis que 14 sont incorrectement attribués au grade 0. Dans la ligne 5, il y a qu'un seul patient qui est au grade 4 mais le modèle n'a pas pu le prédire correctement.

=> Le modèle a bien prédit les classes 0,1 et 2 avec un pourcentage  $\geq 80\%$  mais malheureusement il n'était pas performant avec les grades 3 et 4. Et cela, parce que dans l'ensemble des données il n'a pas plusieurs cas pour pouvoir bien entraîner le modèle du coup c'est tout à fait normal d'avoir des résultats pareils.

Nous avons étudié l'importance des variables afin d'expliquer l'apparition de la dermite radique et comme indiqué dans la figure A.6, le modèle considère que la dose délivrée, le tabagisme et le poids du patient sont les données les plus importantes qui participent à l'apparition de cette toxicité.

##### 4.1.1.2 Modèle SVM

Dans les figures A.7 et A.8 dans l'annexe A.4, nous avons eu un accuracy de 95 % et un f1-score de 69% sur les données d'entraînement et un accuracy de 75 % et un f1-score de 77% sur les données de test.

Maintenant, en regardant la matrice de confusion, dans la ligne 0, il y a 238 patients et 168 des patients sont correctement attribués au grade 0, tandis que 62 sont incorrectement attribués au grade 1 et 8 sont incorrectement attribués au grade 2.

Dans la ligne 1, il y a 183 patients, mais seulement 155 sont correctement attribués au grade 1, tandis que 26 sont

incorrectement attribués au grade 0. Dans la ligne 5, il y a qu'un seul patient qui est au grade 4 mais le modèle n'a pas pu le prédire correctement.

=> Le modèle a bien prédit les classes 0,1 et 2 avec un pourcentage  $\geq 75\%$  mais malheureusement il n'était pas performant avec les grades 3 et 4.

Et cela, parce que dans l'ensemble des données il n'a pas plusieurs cas pour pouvoir bien entraîner le modèle du coup c'est tout à fait normal d'avoir des résultats pareils. De même, comme indiqué dans la figure A.9, le modèle considère que le poids, la taille du patient et la dose délivrée sont les données les plus importantes qui participent à l'apparition de cette toxicité.

### 4.1.1.3 Modèle approuvé

D'après les résultats obtenus, c'est vrai que le modèle RF était plus performant avec un f1-score légèrement plus élevé que le modèle SVM mais les 2 modèles n'ont pas été précis au niveau des classes 3 et 4 et comme expliqué ci-dessus il faut avoir plus de cas pour pouvoir entraîner le modèle.

Afin de comparer entre les 2 modèles, nous avons utilisé la mesure ROC CURVE comme indiqué dans la figure A.10.

La courbe ROC montre que notre courbe RF est plus proche de la bordure gauche du graphique (modèle plus précis). Ce que nous pouvons dire, c'est que les deux algorithmes ont échangé des faux positifs contre des vrais positifs et vrais positifs aux faux positifs.

Pour le choix du modèle, nous avons opté pour le modèle RF et ce pour ces points :

- AUC = 86%
- f1-score = 81%
- Moins de faux positifs que SVM : Il est très important pour nous de ne pas avoir de faux positifs.

### 4.1.2 Analyse non technique

Nous avons réussi à mettre en place 2 modèles d'apprentissage automatique afin de prédire l'apparition de la dermite radique.

Les résultats obtenus (Voir annexe A.4) étaient performants par rapport aux 3 premiers grades de la toxicité qui sont respectivement : "pas de dermite", "Faible érythème ou desquamation sèche" et "Érythème modéré à vif ; desquamation suintante en plaques, affectant principalement les plis et replis cutanés ; œdème modéré".

Pour les grades 3 et 4 qui sont respectivement "Desquamation suintante en plaque, affectant d'autres zones que les plis et replis cutanés ; saignement induit par des traumatismes ou abrasions mineurs" et "Mise en jeu du pronostic vital ; nécrose cutanée ou ulcération de toute l'épaisseur du derme ; saignement spontané des sites affectés ; indication de greffe cutanée", les deux modèles choisis n'ont pas pu fournir de bons résultats parce qu'il y a peu de cas dans les données fournies ce qui fait que les modèles ne sont pas bien entraînés sur ces deux grades d'où des résultats pareils.

Nous n'avons pas pris en considération le dernier grade, qui est "décès", parce que dans notre jeu de données il n'y a pas de cas grade 5.

Certes les résultats étaient très proches mais bien évidemment il faut en choisir un. Pour cela, nous nous sommes basés sur le fait d'avoir un modèle plus précis et qui a moins de faux positifs (les individus présentant un test positif alors qu'ils ne sont pas infectés). D'où le choix du modèle des forêts aléatoires.

Le modèle approuvé considère que la dose délivrée, le tabagisme et le poids du patient sont les données les plus importantes qui participent à l'apparition de cette toxicité.

La conclusion que nous pouvons en tirer :

- L'apparition de la dermite radique est favorable chez un patient fumeur.
- Nous proposons de regarder les doses délivrées qui peuvent être la cause de cette apparition.

## 4.2 CONCLUSION

Cette étape est très constructive, elle nous permet de revoir l'ensemble du projet pour vérifier s'il répond aux normes que nous nous sommes fixées au départ, surtout d'analyser les résultats obtenus en précisant les points positifs et négatifs.



---

# CONCLUSION GÉNÉRALE

Ce projet a été un vrai challenge pour moi. Dans ce travail, l'intérêt s'est porté sur la prédiction de la toxicité "dermite radique" avec des techniques d'apprentissage automatique.

L'ensemble de données considéré s'appuie sur les diagnostics de plusieurs patients qui ont subi des séances de radiothérapie. Un prétraitement de données a été effectué avant une visualisation et une analyse des données.

Durant ce projet, j'ai rencontré plusieurs difficultés dès le lancement : La plupart des données étaient catégorielles qui n'aident pas ni à faire des analyses ni à réaliser des corrélations ...

L'absence de l'axe du temps pour les variables temporelles m'a empêché d'utiliser des modèles plus performants dédiés pour les séries temporelles.

La présence de plusieurs données manquantes m'a fait perdre beaucoup de temps parce que l'imputation n'était pas évidente et je ne maîtrise pas cette compétence. De plus, il y a le manque de connaissances du métier qui demandait beaucoup de recherches et de documentations.

Je voulais essayer plus de méthodes d'imputation, de nettoyage et d'apprentissage. Cependant, en raison de la contrainte de temps et de la complexité de la méthode elle-même, cette partie n'a pas été atteinte. Se concentrer sur la partie donnée manquantes, tester d'autres méthodes d'imputation et peut être essayer d'autres techniques d'optimisation pourrait améliorer les résultats obtenus.

Ce fut une expérience très intéressante et précieuse au cours de laquelle nous avons étudié et acquis des compétences et des connaissances techniques.

En raison du rythme de l'alternance et des cours qui étaient en parallèle, l'avancement du projet était un peu compliqué, mais enfin je peux conclure que ce projet a été bénéfique à plusieurs niveaux et je suis très heureuse de cette expérience qui m'a donné l'occasion de participer à un tel défi.



---

## Bibliographie

- [1] **MonCoachData**. Tous les modèles de Machine Learning expliqués brièvement **[en ligne]**. Disponible sur : <https://moncoachdata.com/blog/modeles-de-machine-learning-expliques/>
- [2] **WIKIPEDIA** . Matrice de confusion **[en ligne]**. Mis à jour en janvier 2022 . Disponible sur : [https://fr.wikipedia.org/wiki/Matrice\\_de\\_confusion](https://fr.wikipedia.org/wiki/Matrice_de_confusion)
- [3] **OPENCLASSROOM** . Évaluez un algorithme de classification qui retourne des valeurs binaires **[en ligne]**. Mis à jour en septembre 2021 . Disponible sur : <https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4308256-evaluez-un-algorithme-de-classification-qui-retourne-des-valeurs-binaires>
- [4] **DEVELOPPERS** . Classification : Accuracy **[en ligne]**. Mis à jour en mars 2022 . Disponible sur : <https://www.overleaf.com/project/62a331a20c038a6ff2962a0d> <https://developers.google.com/machine-learning/crash-course/classification/accuracy?hl=fr>
- [5] **Pierre Baudin** . Comprendre et identifier les données manquantes **[en ligne]**. Mis à jour en décembre 2020 . Disponible sur : <https://blog.avanci.fr/classification-des-donnees-manquantes>
- [6] **Great Learning Team** . Hyperparameter Tuning with GridSearchCV **[en ligne]**. Mis à jour en septembre 2020 . Disponible sur : <https://www.mygreatlearning.com/blog/gridsearchcv/>
- [7] **DEVELOPPERS**. Classification : ROC Curve and Area under curve **[en ligne]**. Mis à jour en février 2020 . Disponible sur : <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=fr>



# ANNEXES

## A.1 Exemple de modalité des variables

Tabagisme...cigarettes.j.MAX_AIGU	Effectifs	Fréquence
0	0.0	1788 62.802950
1	10.0	73 2.564103
2	15.0	39 1.369863
3	20.0	30 1.053741
4	5.0	25 0.878117
5	3.0	16 0.561995
6	6.0	13 0.456621
7	1.0	13 0.456621
8	2.0	13 0.456621
9	4.0	10 0.351247
10	7.0	7 0.245873

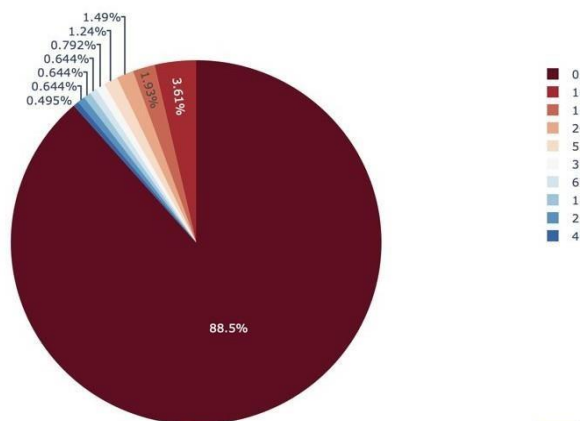


FIGURE A.1 – Modalités des variables (exemple tabagisme)

## A.2 Technique de modélisation

### A.2.1 Random Forest

Les forêts d'arbres décisionnels ou forêts aléatoires (Random Forest) sont une technique d'apprentissage d'ensemble qui s'appuie sur des arbres de décision. Le modèle de forêt aléatoire implique la création de plusieurs arbres de décision en utilisant un ensemble de données séparés à partir des données d'origine. Et en sélectionnant au hasard un sous-ensemble de variables à chaque étape de l'arbre de décision. Ensuite, le modèle sélectionne tous les modes prédits pour chaque arbre de décision. [1]

Les paramètres les plus importantes de RF sont :

- n\_estimators = nombre d'arbres dans la prévision
- max\_depth = nombre maximum de niveaux dans chaque arbre de décision
- min\_samples\_leaf = nombre minimal de points de données autorisés dans un nœud feuille

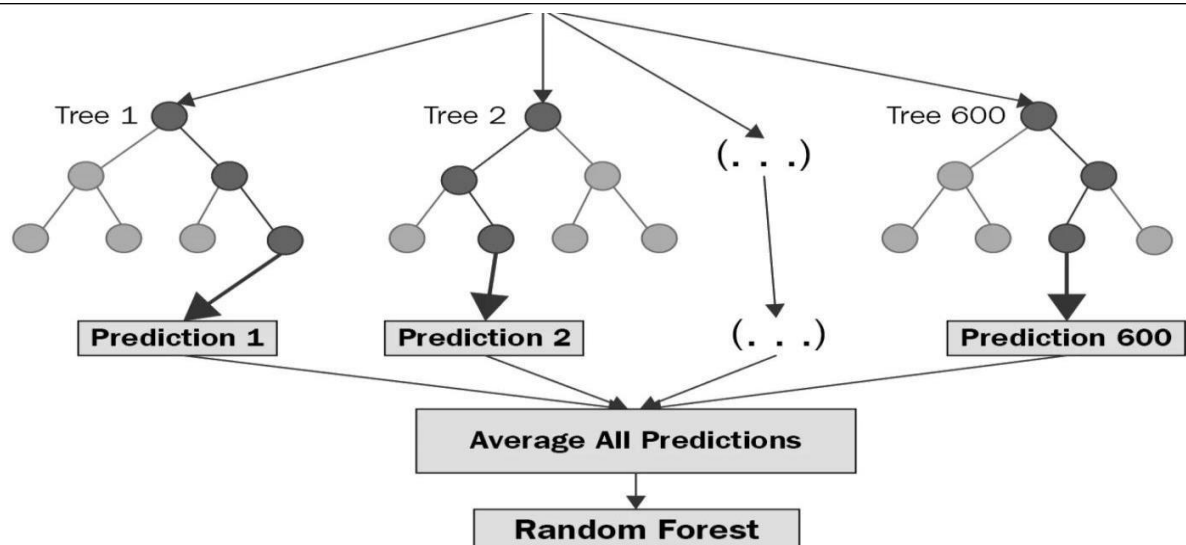


FIGURE A.2 – Modèle des forêts aléatoires

### A.2.2 Support vector machine

La machine à vecteurs de support est une technique de classification supervisée, qui peut en fait devenir assez compliquée, mais elle est très intuitive au niveau le plus élémentaire. [33] On suppose qu'il existe deux types de données. La machine à vecteurs de support trouvera un hyperplan ou une frontière entre les deux classes de données, ce qui maximisera la marge entre les deux classes (voir ci-dessous). Il existe plusieurs plans pour séparer les deux catégories, mais un plan peut maximiser la marge ou la distance entre les catégories. [1]

Les paramètres les plus importantes du SVM sont :

- kernel : le noyau du modèle
- C : C'est le paramètre de régularisation, coût, du terme d'erreur.

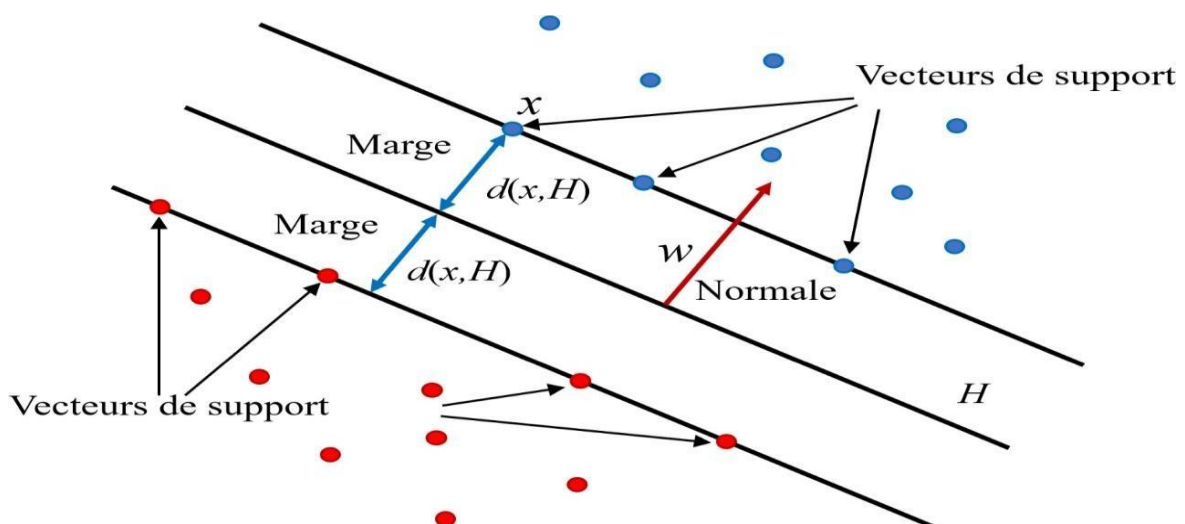


FIGURE A.3 – Modèle du SVM

### A.3 Réglage des hyperparamètres

**GridSearchCV** [6] : est le processus de réglage des hyperparamètres afin de déterminer les valeurs optimales pour un modèle donné. Les performances d'un modèle dépendent de manière significative de la valeur des hyperparamètres. Il n'y a aucun moyen de connaître à l'avance les meilleures valeurs pour les hyperparamètres, donc idéalement, nous devons essayer toutes les valeurs possibles pour connaître les valeurs optimales. Faire cela manuellement peut prendre beaucoup de temps et de ressources et nous utilisons donc GridSearchCV pour automatiser le réglage des hyperparamètres.

### A.4 Résultats

#### A.4.1 Forêt aléatoire

	accuracy	F1-score
Training data	92%	89%
Testing data	83%	81%

FIGURE A.4 – Accuracy et F1-score du modèle RF

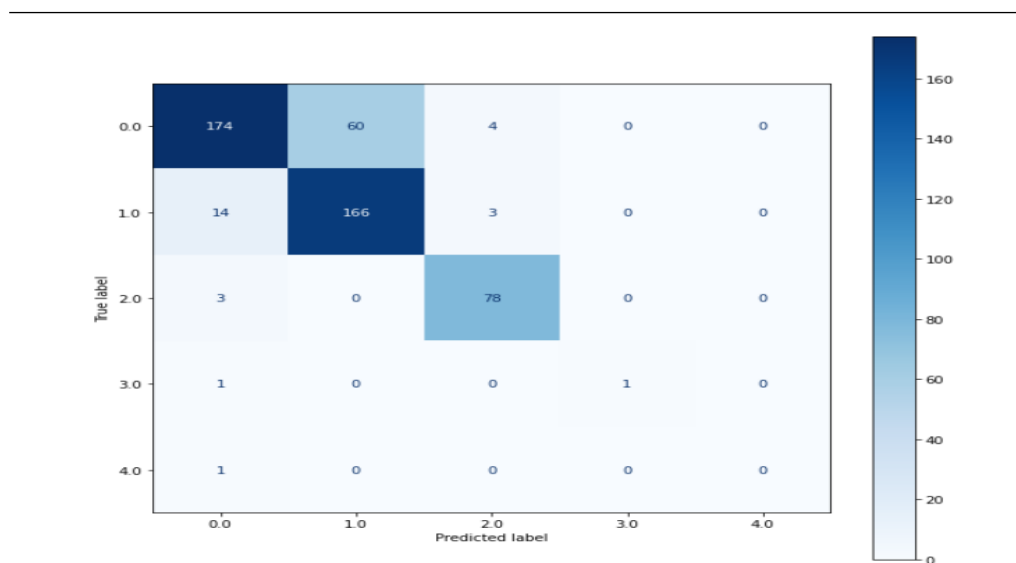


FIGURE A.5 – Matrice de confusion du modèle RF



	importance
DoseDelivree	48.224568
tabagisme aigu	35.213655
Poids.moyen.pendant.le.traitement	31.256448
TailleLesionInvasive	25.458745
performance status aigu	18.669322
Taille.moyenne	10.113489

FIGURE A.6 – Importance des caractéristiques (modèle RF)

#### A.4.2 Support vector machine

	accuracy	F1-score
Training data	95%	69%
Testing data	75%	77%

FIGURE A.7 – Accuracy et F1-score du modèle SVM

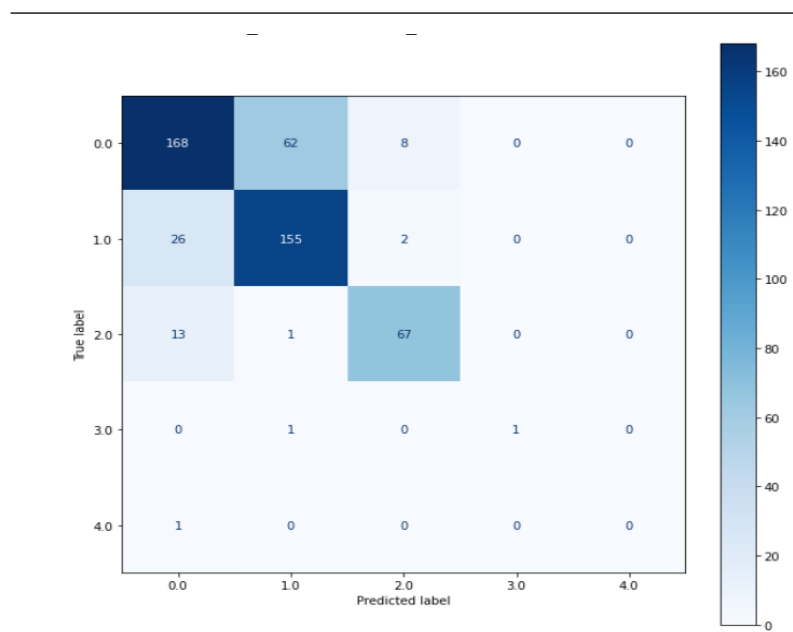


FIGURE A.8 – Matrice de confusion du modèle SVM

---

	importance
Poids.moyen.pendant.le.traitement	37.223456
Taille.moyenne	32.564981
DoseDelivree	24.245646
TailleLesionInvasive	19.332115
performance status aigu	16.477812
tabagisme aigu	9.563329

---

FIGURE A.9 – Importance des caractéristiques (modèle SVM)

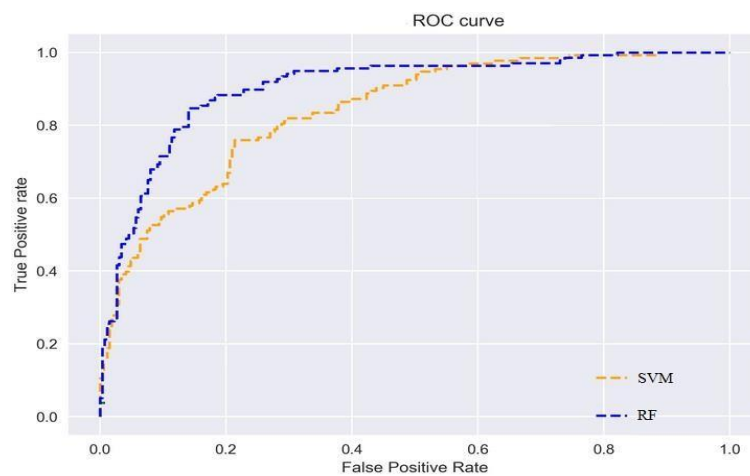


FIGURE A.10 – Courbe ROC pour comparer les modèles