Geoffroy DAUMER

Rapport de mi-alternance au CHU de Caen

Formation en développeur IA

Par l'Isen / Simplon / Microsoft

Tuteurs entreprise : Alain MANRIQUE

Damien LEGALLOIS

"En adressant ce document à l'enseignante, je certifie que ce travail est le mien et que j'ai pris connaissance des règles relatives au référencement, au plagiat, ainsi qu'à l'usage d'une intelligence artificielle d'aide à la rédaction de type ChatGPT."

À remettre le 07/12/2023

Résumé

L'objectifs de mon alternance est de développer des modèles d'IA qui extraient des informations sur des examens médicaux, dans le but de réaliser des études rétrospectives de ces examens.

Il y a deux types d'examens à étudier : les scintigraphies myocardiques de perfusion et les coronarographies.

La base de données sur laquelle j'ai entrainé mes algorithmes a été générée avec des expressions régulières (recherche de motifs dans du texte). Les outils que j'ai utilisés sont principalement des modèles d'IA de la librairie scikit-learn, et des programmes de reconnaissance d'expressions régulières.

Les modèles d'IA obtiennent dans la majorité d'excellente performances, sauf pour les cibles continues et pour certaines cibles multiclasses, où il est préférable de rester sur de la reconnaissance d'expressions régulières.

Table des matières

1. Introduction

Mon alternance s'est déroulée au département de médecine nucléaire au CHU de Caen. C'est là que se déroulent certains examens développés à base de traceurs radioactifs, on y dépiste le cancer, les problèmes neurologiques, les ischémies cardiaques…

Les données de santé dont dispose l'hôpital sont de plus en plus nombreuses, il y a un grand intérêt à les formater afin de les analyser, pour faire progresser la connaissance en médecine.

Le but de cette alternance est de contribuer à un projet de recherche qui porte sur la comparaison des performances diagnostiques de la scintigraphie myocardique de perfusion et de la coronarographie, ce sont des examens qui donnent des informations sur l'état de santé des artères du cœur. Je suis le seul développeur dans le projet, en collaboration avec deux médecins-chercheurs en cardiologie et médecine nucléaire : Alain MANRIQUE et Damien LEGALLOIS.

La scintigraphie myocardique de perfusion permet d'obtenir des informations sur le fonctionnement du cœur.

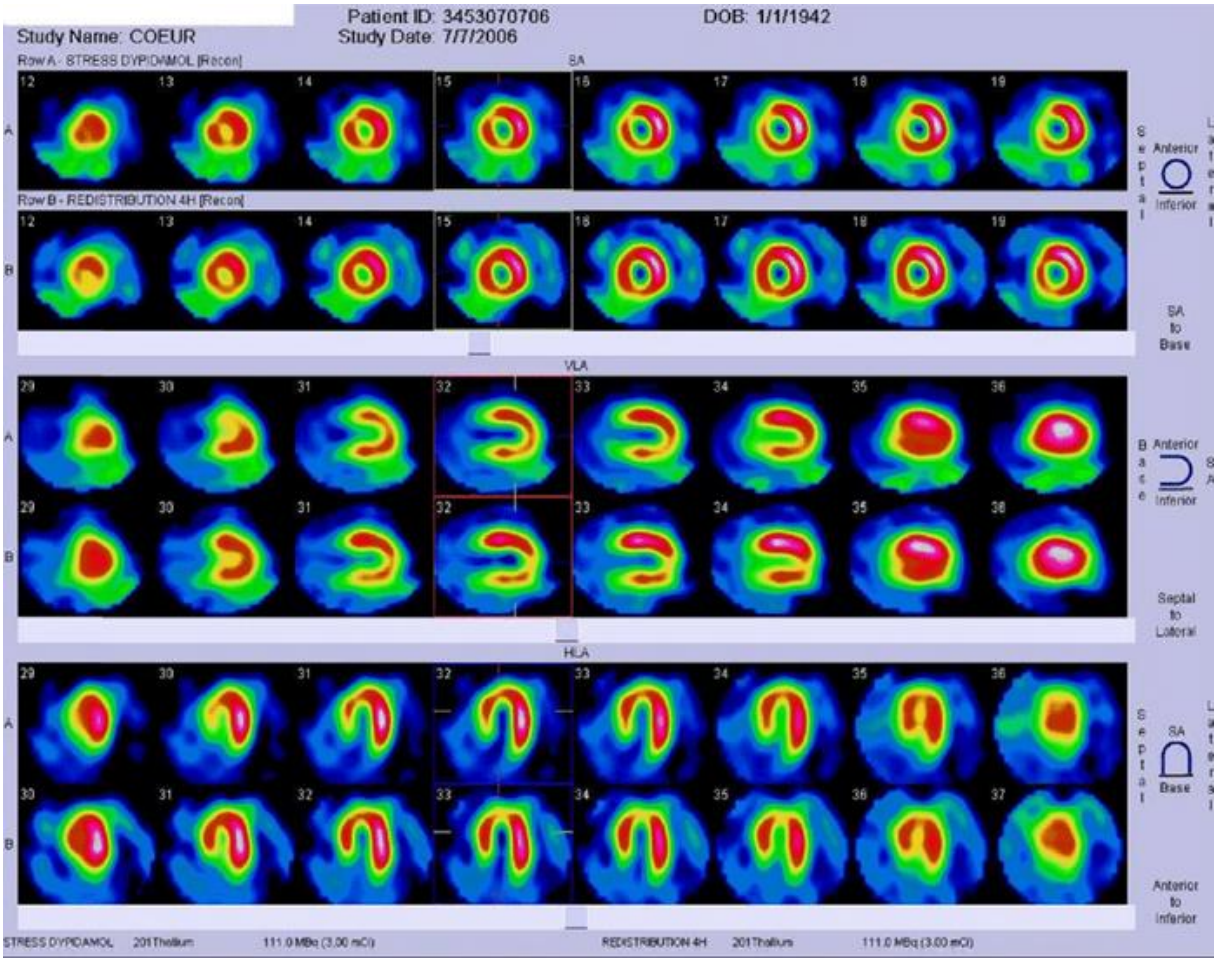


Figure 1 : Images de scintigraphies myocardiques

Lorsqu'un patient présente des symptômes de maladie coronarienne à l'issue de cet examen, il est envoyé en coronarographie pour faire un diagnostic anatomique des artères du cœur, en vue d'une potentielle opération.

La coronarographie est l'examen de référence (« gold standard ») pour dire si le patient est malade ou non.

Cet examen consiste à passer une sonde par la veine jusqu'au cœur afin de disperser un produit et d'obtenir des images des artères coronaires.

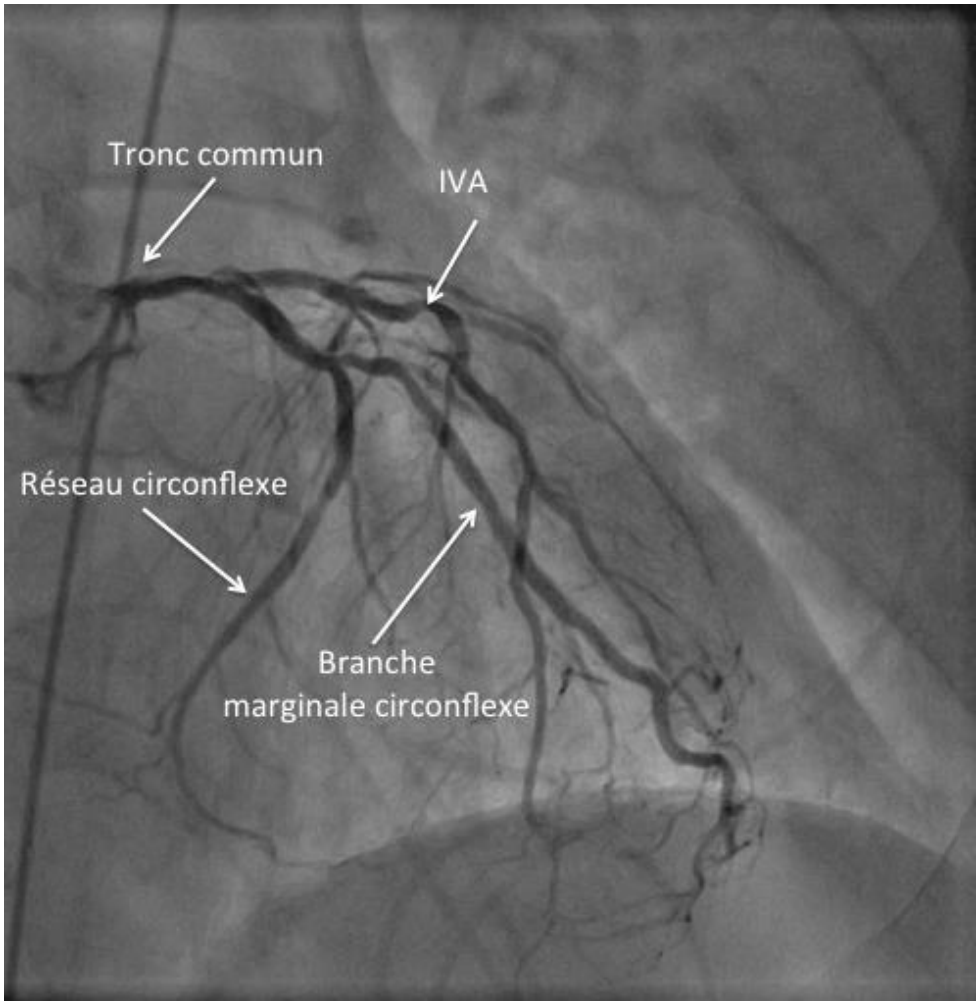


Figure 2 : Exemple d'image obtenue suite à une coronarographie

On souhaite regarder la corrélation entre les informations fonctionnelles que révèlent la scintigraphie, ainsi que leurs interprétations par les médecins, avec les réelles pathologies anatomiques qui apparaissent à la coronarographie, qui est le « gold standard » : l'examen de référence pour dire si oui ou non le patient a vraiment un problème.

La finalité du projet est d'analyser le profil des patients qui sont envoyés en coronarographie, suite à une scintigraphie, pour comprendre dans quels cas la prédiction est vraie et dans quels cas elle est fausse.

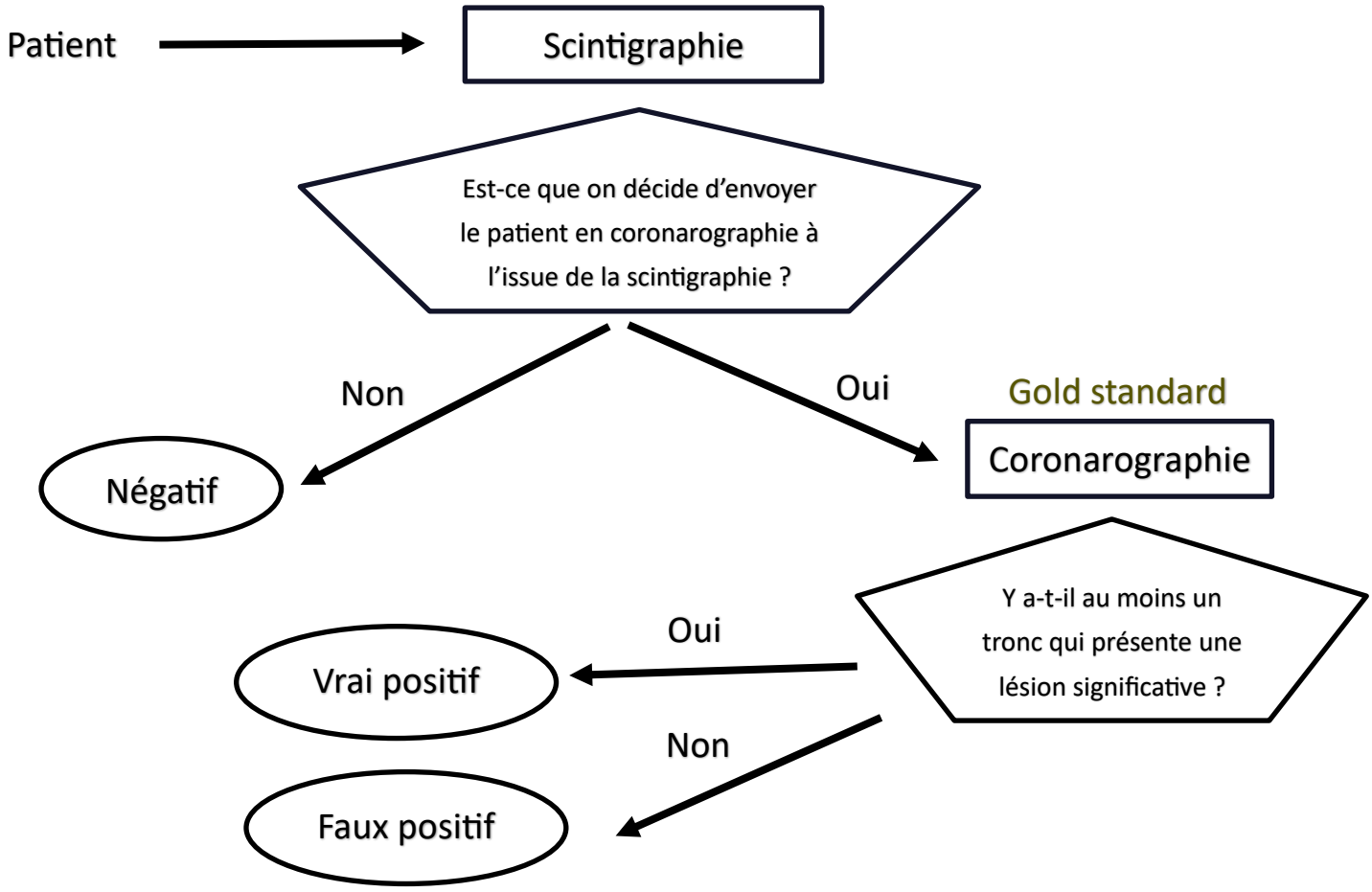


Figure 3 : Schéma explicatif de la comparaison des performances diagnostiques de la scintigraphie myocardique de perfusion

		Réalité	
		Négatif	Positif
Prédiction	Négatif	Vrai négatif (inconnu)	Faux négatif (inconnu)
	Positif	Faux positif (connu)	Vrai positif (connu)

Figure 4 : Matrice de confusion

Nous ne savons pas si les patients qui sont prédits négatifs, sont vraiment négatifs, car ils n'ont pas été envoyés en coronarographie.

La méthode de travail en équipe sera expliquée avec la partie 2 « Gestion de projet ».

La documentation sur les technologies à utiliser dans la partie 3 : « Veille technique ».

Une partie « Méthodologie » explique la réalisation du projet dans l'aspect technique.

Dans la partie « Résultats », je présenterais les métriques de performance obtenues des modèles sur les différentes informations à relever.

Une partie « Discussion » où je commenterais les résultats, et pour finir la « Conclusion ».

2. Gestion de projet

2.1 Méthodologie de gestion de projet

- Au départ nous avons fait des réunions. Mes tuteurs, Alain et Damien, m'ont expliqué le projet.
- Ensuite, je me suis renseigné sur les réglementations pour les projets de recherches, ainsi que la protection des données.
- Je me suis documenté sur les métriques de performances utilisées et sur les technologies utilisées pour ce type de projet.
- Le plus souvent, j'ai fait du code, appris à l'organiser, entraîné des modèles, fais des pipelines de prédiction, testé des IA.
- En septembre, j'ai fait une présentation de mon travail à mes tuteurs, que j'ai dû retravailler par la suite, cela m'a permis d'apprendre à vulgariser le machine learning, et d'approfondir le fonctionnement de certains modèles.

Au niveau de la communication, on a l'occasion de se parler en présentiel les jeudis.

2.2 Cahier des charges

2.2.1 Règlementations relatives à la protection des données, et aux travaux de recherche.

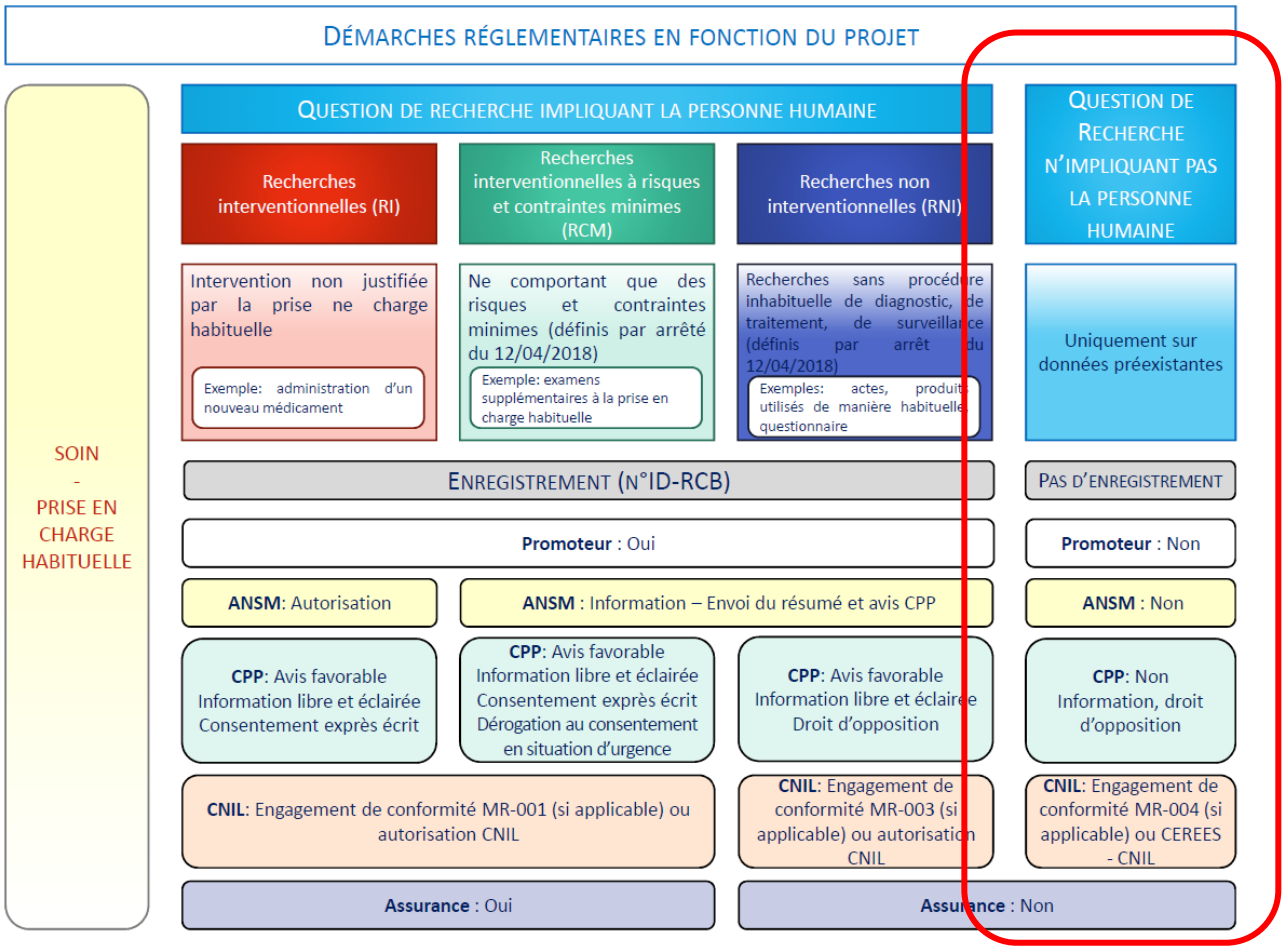


Figure 5 : Schéma des réglementation pour les travaux de recherche impliquant les données des patients

Pour réaliser un travail de recherche en santé, il faut obtenir une autorisation d'un comité d'éthique. Dans le cas de recherches portant uniquement sur des données préexistantes, c'est au CLERS (Comité Local d'Ethique pour la Recherche en Santé) de délivrer l'autorisation.

Nous pouvons voir à partir du schéma que l'utilisation des données rétrospectives de patients nécessite une conformité à la MR-004.

Les règles de la MR-004 :

- Ne collecter que des données strictement nécessaires au regard des objectifs de la recherche.

- Pour transférer des données hors UE (exemple, sur les « clouds »), le transfert doit être strictement nécessaire, et les données doivent être indirectement identifiantes.

- « Chaque projet conforme à la MR-004 » doit être enregistré dans un répertoire public tenu par la plateforme des données de santé

- Les données des patients peuvent être conservées jusqu'à 2 ans après la dernière publication des résultats

- Seuls les professionnels et leurs collaborateurs intervenant dans la recherche, dans un lieu de recherche peuvent conserver le lien entre l'identité codée des personnes se prêtant à la recherche utilisée pour associer les données de santé à caractère personnel et leurs noms et prénoms.

- Une information générale, et individuelle, concernant la réutilisation des données pour les activités de recherche, doit-être assurée auprès des personnes concernées. Cette note comprend les détails relatifs à la réutilisation des données, comme le droit de rétractation.

- Dans le cas où l'information des patients représente un effort disproportionné, on peut demander une dérogation (données pré-2019 dans le cas du CHU).

- Lors de la publication de la recherche, il faut anonymiser un maximum les données (pas d'initiales, dates de naissance, lieux…)

2.2.2 Structuration des données au sein de l'établissement

Un « Entrepôt des Données de Santé » est en construction au CHU (EDS), par la startup CODOC. Il permettra de requêter les données facilement et de manière anonymes.

Un comité scientifique et éthique de cet entrepôt de données est mis en place, auquel il faudra soumettre une demande pour exploiter certaines données, en plus de la demande au CLERS.

Cependant, il n'est pas encore accessible, donc mon code ne sera pas basé ni adapté dessus.

2.2.3 Exigences de performances des modèles

L'objectif est d'obtenir pour chaque information à extraire des comptes rendus une aire sous la courbe ROC supérieure à 90% (exemple maladie cardiovasculaire ?).

La courbe ROC donne le taux de vrais positifs en fonction du taux de faux positifs, à différents seuils de discrimination. Elle est souvent utilisée pour montrer les progrès d'un classificateur binaire lorsque le seuil de discrimination varie.

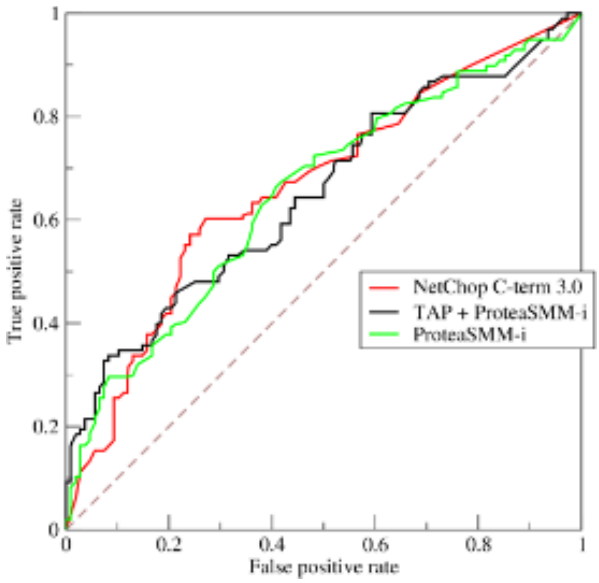


Figure 6 : Exemple de courbe ROC, représentant les progrès de détermination de 3 caractéristiques

Dans la pratique, on veut le point où on a la meilleure sensibilité (ordonnée haute), ainsi que la meilleure spécificité (abscisse basse).

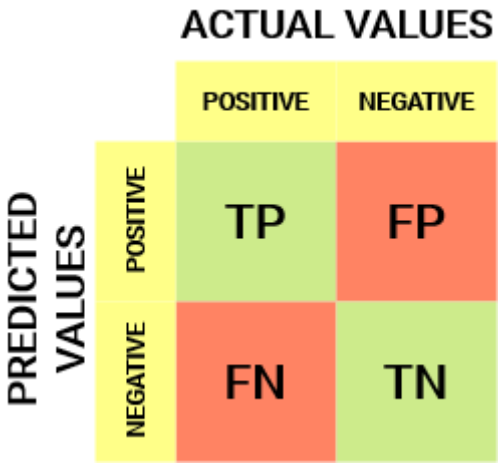
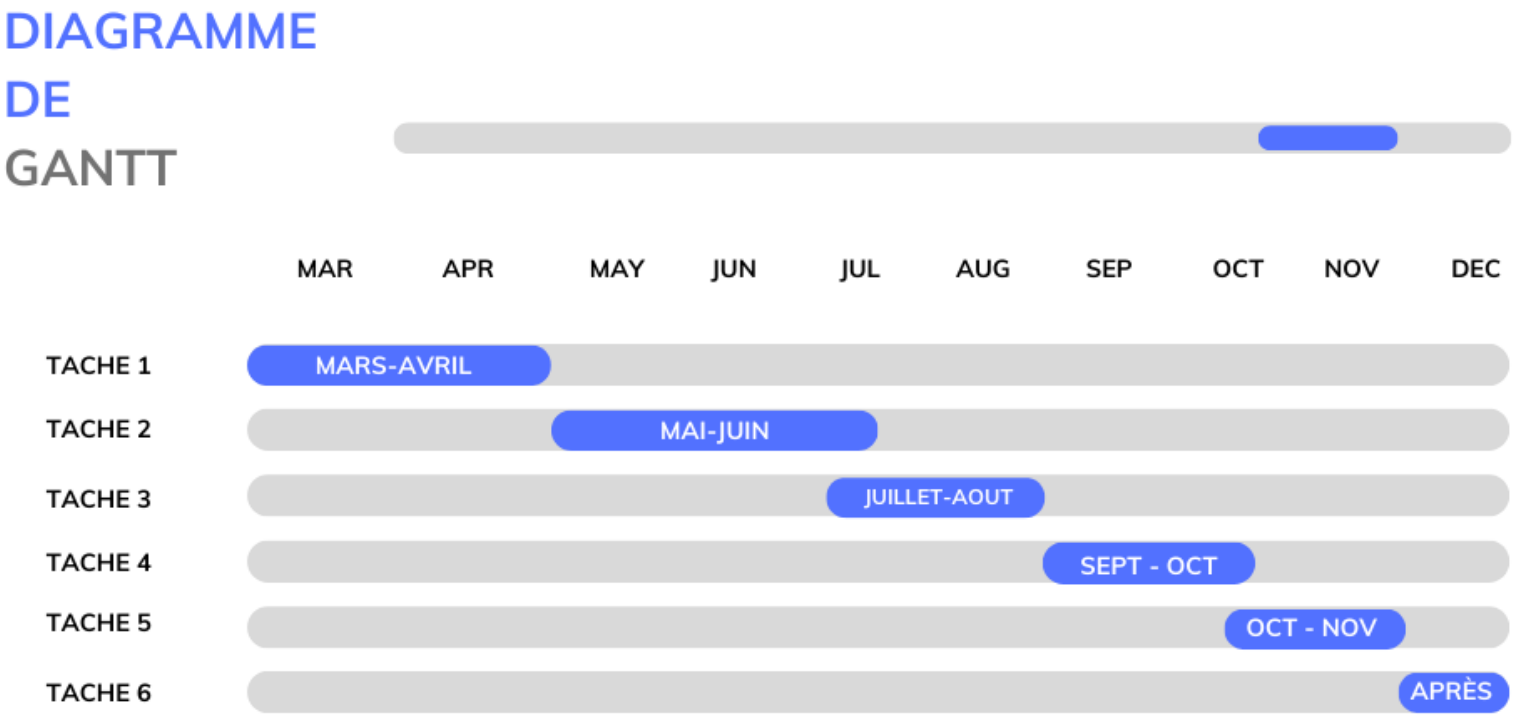


Figure 7 : Matrice de confusion

2.3 Diagramme de GANTT



Tache 1 : Recherche sur les réglementations pour les projets de recherches, ainsi que pour la protection des données.

Tache 2 : Veille technique, documentation sur les métriques de performance utilisées.

Tache 3 : Développement d'un code permettant de comparer différents modèles de la librairie scikit-learn.

Tache 4 : Organisation du code, développement des fonctions finales de prédiction.

Tache 5 : Mise en production des modèles avec une application qui extrait les informations sur les comptes rendus de scintigraphies.

Tache 6 : Mise au propre du code, amélioration des modèles.

Figure 8 : Diagramme de Gantt du projet

3. Veille technique

Dans cette partie, je vous présente les technologies qui permettent d'extraire des données sur du texte.

- Les expressions régulières :

  Ce n'est pas de l'IA, mais plutôt un programme qui reconnait des motifs spécifiques de texte.

  Mon jeu d'entraînement pour mes modèles d'IA fut généré avec cette technologie.

  On peut générer aussi bien des valeurs continues que des valeurs catégoriques, et donc alimenter divers types de modèles avec les données générées :

Exemple pour les valeurs continues :

On peut chercher le groupe de chiffres qui se situe après « vts », à condition que :

- Avant plus loin il y ai « acquisition de stress »
- Il y ai entre 0 et 14 caractères non numériques qui précèdent le groupe de chiffres
- Après 0 à 10 caractère suivant le groupe de chiffres il y ai « ml »

Si on le trouve on met sa valeur, sinon on ne met rien.

Exemple pour les valeurs catégoriques :

Chercher dans le texte si il y a :

- « cmd »
- Ou « cardiopathie dilatee »
- Ou « cardiomyopathie dilatee »

Si on trouve un des motifs on met « cmd » dans la colonne « cmd » de notre tableau,

Sinon on met « no_cmd ».

Recherche d'ischémie myocardique chez patiente de 58 ans. Douleur constrictive à type "d'étau" au repos rare de durée brève de localisation plutot digestive. Dyspnée NYHA 2. ATCD CV : AOMI FDR CV : Tabagisme actif à 40 PA, diabète de type 2 diagnostiqué il y a 3 ans non insulinodépendant, hérédité cardiovasculaire IMC 31 kg/m2. ECG : RRS 80/min, qRs fins, pas de trouble significatif de la repolarisation ETT : FEVG normale Traitement : METFORMINE KARDEGIC ATORVASTATINE L'acquisition d'effort a été synchronisée à l'électrocardiogramme, et pratiquée à la suite après l'injection de 3.7 MBq/kg de 99mTc-Sestamibi. Epreuve de stress : [] Effort sur bicyclette explorant % de la FMT (W max) [] Test mixte associant Persantine (0.56 mg/kg) et épreuve d'effort sur bicyclette [X] Persantine [] Regadenoson [] Dobutamine, réalisée par [] Pr AGOSTINI [] Pr MANRIQUE [] Dr MARNEFFE [X] Dr TAGER, cliniquement négative et électriquement négative. Acquisition de stress : Perfusion : homogène Cinétique segmentaire : normale FEVG : 83 % Volumes ventriculaires : VTD = 52ml et VTS = 9 ml Tomoscintigraphie myocardique de perfusion normale. FEVG normale. Traitement médical à poursuivre.

Figure 9 Exemple de texte

```python
vts_stress = re.findall(
    r"(?<=acquisition de stress).*?(?<=vts)\D{0,14}(\d+).{0,10}?(?=ml)",
    text,
    re.IGNORECASE,
)
```

Figure 10 Exemple d'expression régulière en python

vts_stress = 9 ml

AX
vtd_stress
54
82
167
104

CMD

```r
pattern <- "cmd|cardiopathiedilate|cardiomyopathiedilate"
spect$cmd <- ifelse(spect$search %like% pattern == TRUE,"cmd","no_cmd")
spect$cmd <- ordered(spect$cmd, levels = c("no_cmd", "cmd"))
```

Figure 11 Exemple d'expression régulière en R

cmd = no_cmd

M
cmd
no_cmd
no_cmd
no_cmd
no_cmd
no_cmd

Schéma explicatif de l'utilisation des expressions régulières (1) pour la construction de jeux de données

Avantages :

- Ne requiert pas d'annoter manuellement les comptes rendus, comme il le faudrait pour entraîner des algorithmes d'IA, gain de temps significatif.
- Autant d'erreurs manuelles que d'erreurs machine (1).
- Des pipelines de NLP basés sur des règles peuvent être mises en place, dans le cas où on aurait différente manières de décrire une pathologie.

Inconvénient :

- Les outils d'IA ont de meilleurs résultats sur de larges jeux de données (syntaxe variée).
- Limité pour obtenir des informations complexes, exemple : caractéristiques des tumeurs, et peut s'arrêter à seulement un type de spécimen ou d'organes.

"An accessible, efficient, and accurate natural language processing method for extracting diagnostic data from pathology reports (1)"

Il y a deux principaux types de modèles d'IA pour extraire les données sur du texte :

- Les modèles de classification
- Les modèles de reconnaissance d'entité nommées

Les modèles de classification sont utilisés pour classifier du texte lorsque la cible peut avoir un nombre fini de valeurs, exemple : le patient a-t-il de la dyslipidémie ? oui ou non.

Les modèles de reconnaissance d'entités nommées (NER) sont utiles pour relever des informations qui peuvent prendre une infinité de valeurs dans le texte, comme une entité mesurable (exemple en % ou en ml).

- Les modèles de classification :

  Ces modèles trouvent des relations entre le texte et la cible après avoir transformé ce dernier en vecteur de nombres, exemple la librairie scikit-learn.

  Avantage :
  - Nécessite peu de ressources de calcul
  - Simple d'utilisation

  Inconvénient :
  - Pas les meilleures performances

- L'extraction d'entités nommées (NER) :

  Elle permet de reconnaitre des mots dans un contexte, exemple « 57 ml » pourra être prédit comme un volume télé diastolique.
  Les modèles de NER sont produits à partir des modèles de langues.

  Avantage :
  - Meilleures performances que les expressions régulières

  Inconvénient :
  - Nécessite beaucoup de ressources
  - Les données d'entraînement doivent être annotées avec la position de l'entité à extraire dans chaque document.

- Les modèles de langues :

  La technologie à la pointe du traitement du langage naturel. Ces modèles sont des réseaux de neurones entrainés sur des grandes quantités de données textuelles. Ils nécessitent d'importantes ressources de calcul pour être « réglés sur des jeux de données spécifiques ». A savoir que les modèles de langues peuvent être utilisés de différentes façons, comme la traduction, classification, régression, NER…

- Les réseaux de neurones :

  Cette technologie, lorsqu'elle est utilisée hors du contexte des grands modèles de langues (LLM), a d'excellent résultats.

  Des chercheurs ont employé le word embedding et un réseau de neurones convolutionnel pour reconnaitre la classification internationnale des maladies (ICD-10), et ont surclassé les méthodes actuelles, avec une préparation de données minimum.

  "An NLP tool for data extraction from electronic health records: COVID-19 mortalities and comorbidities" – 2022 (2)

  Avantage :
  - Excellents résultats

  Inconvénient :
  - Nécessite beaucoup de ressources

Table 2
Comparison of manual and automated data extraction methods for grade and location of dysplasia in colorectal surveillance biopsies among patients with inflammatory bowel disease.

	Data extraction method		Statistics
	Manual	Automated	
Dysplasia grade			
Negative for dysplasia	249 (81.4%)	249 (81.4%)	
Indefinite for dysplasia	0 (0.3%)	1 (0.3%)	Concordance: 99.0%
Low-grade dysplasia	50 (16.3%)	49 (16.0%)	Cohen's κ: 0.97
High-grade dysplasia	5 (1.6%)	4 (1.3%)	*P* value: <.001
Adenocarcinoma	2 (0.7%)	3 (1.0%)	
Dysplasia location			
Rectum/sigmoid	79 (25.8%)	75 (24.5%)	
Descending colon/SF	67 (21.9%)	64 (20.9%)	Concordance: 97.1%
Transverse colon	35 (11.4%)	34 (11.1%)	Cohen's κ: 0.96
Ascending colon/HF	63 (20.6%)	64 (20.9%)	*P* value: <.001
Cecum/Ileocecal valve	26 (8.5%)	27 (8.8%)	
Other	36 (11.8%)	42 (13.7%)	
Time invested			
Person*hours	6.5	0.17	

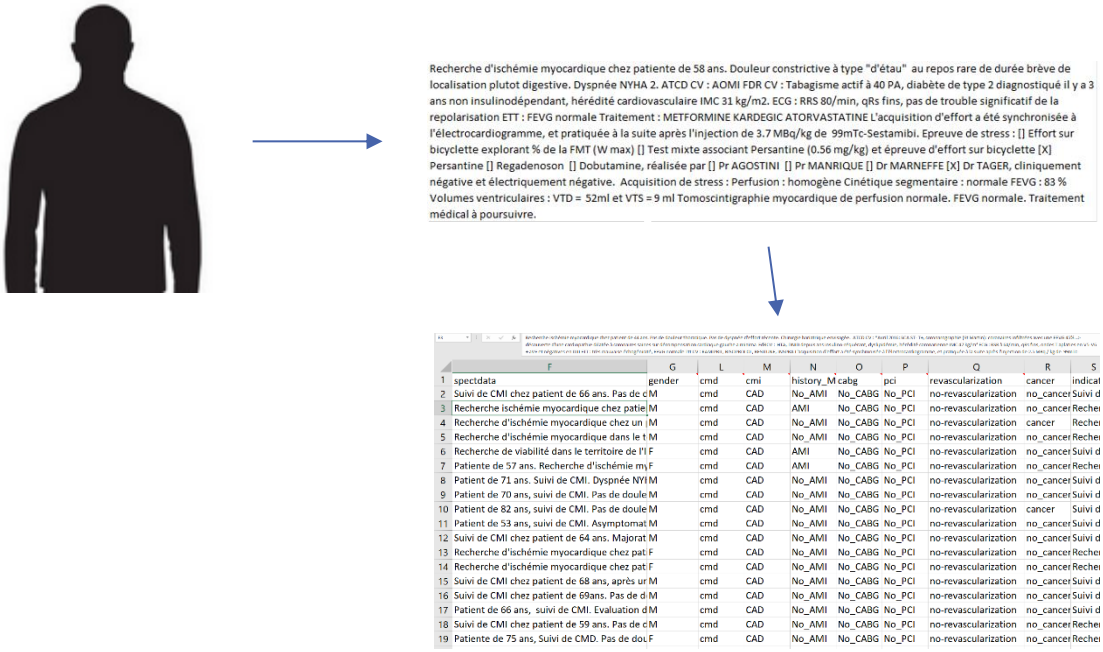
Figure 12 Une partie des résultats de l'article (1) qui utilise les expressions régulière pour extraire les informations médicales sur du texte

Métrique d'évaluation : coefficient kappa (établis la concordance entre deux observateurs en tenant compte de la concordance due au hasard), valeur P pour établir un seuil de significativité (1).

4. Méthodologie

4.1 Résumé de la tâche à réaliser

Pour étudier les patients, il faut transformer le compte rendu de leurs examens en jeux de données, afin de les rendre analysables avec des statistiques.



	F	G	L	M	N	O	P	Q	R	S
1	spectdata	gender	cmd	cmi	history_M	cabg	pci	revascularization	cancer	indicat
2	Suivi de CMI chez patient de 66 ans. Pas de c	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
3	Recherche ischémie myocardique chez patie	M	cmd	CAD	AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
4	Recherche d'ischémie myocardique chez un	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	cancer	Reche
5	Recherche d'ischémie myocardique dans le t	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
6	Recherche de viabilité dans le territoire de l'I	F	cmd	CAD	AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
7	Patiente de 57 ans. Recherche d'ischémie my	F	cmd	CAD	AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
8	Patient de 71 ans. Suivi de CMI. Dyspnée NYI	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
9	Patient de 70 ans, suivi de CMI. Pas de doule	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
10	Patient de 82 ans, suivi de CMI. Pas de doule	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	cancer	Suivi d
11	Patient de 53 ans, suivi de CMI. Asymptomat	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
12	Suivi de CMI chez patient de 64 ans. Majorat	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
13	Recherche d'ischémie myocardique chez pat	F	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
14	Recherche d'ischémie myocardique chez pat	F	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
15	Suivi de CMI chez patient de 68 ans, après ur	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
16	Suivi de CMI chez patient de 69ans. Pas de d	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
17	Patient de 66 ans, suivi de CMI. Evaluation d	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
18	Suivi de CMI chez patient de 59 ans. Pas de c	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
19	Patiente de 75 ans, Suivi de CMD. Pas de dou	F	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche

Figure 13 Schéma explicatif du processus de transformation des données

4.2 Contexte de travail

On pourrait, pour extraire les données du texte, utiliser directement les expressions régulières.

On les a utilisées, mais on voudrait se servir du jeu de données généré pour entraîner des IA qui reconnaitraient par elles même les informations dans le texte.

La question est la suivante :

Est-ce qu'une IA entraînée sur un jeu de données issu d'expressions régulières sera plus intelligent que les expressions régulières elles-mêmes ?

Non

On est donc face à un dilemme :

- Abandonner l'IA
- Faire de l'IA pour s'entraîner
- Annoter des données à la main et faire de l'IA
- Appliquer des modèles de langues intelligent directement au problème, sans ajuster le modèle avec des jeux d'entraînement, seulement vérifier les résultats.

4.3 Jeux d'entraînement

Pour entraîner des algorithmes d'IA, il faut un jeu d'entraînement. Mes tuteurs ont réalisé un programme de fouille de données avec le langage R, qui utilise les expressions régulières. Il est basé sur les scintigraphies, les données de coronarographies ont, elles, été générées à la main pour l'année 2019, le tout formant un jeu d'entraînement d'une année sur les deux examens.

Voici à quoi ressemblent les jeux d'entraînement pour les modèles :

F3 | Recherche ischémie myocardique chez patient de 44 ans. Pas de douleur thoracique. Pas de dyspnée d'effort récente. Chirurgie bariatrique envisagée. ATCD CV : *Avril 2016: SCA ST- T+, coronarographie (St Martin): coronaires infiltrées avec une FEVG 45% => découverte d'une cardiopathie dilatée à coronaires saines sur décompensation cardiaque gauche a minima FdRCV : HTA, DNID depuis ans insulino-réquérant, dyslipidémie, hérédité coronarienne IMC 47 kg/m² ECG : RSR à 64/min, qRs fins, ondes T aplaties en V5-V6 + aVF et négatives en DIII ETT : très mauvaise échogénicité, FEVG normale Ttt CV : RAMIPRIL, BISOPROLOL, RESITUNE, INSPRA L'acquisition d'effort a été synchronisée à l'électrocardiogramme, et pratiquée à la suite après l'injection de 2.5 MBq / kg de 99mTc-

	F	G	L	M	N	O	P	Q	R	S
1	spectdata	gender	cmd	cmi	history_M	cabg	pci	revascularization	cancer	indicat
2	Suivi de CMI chez patient de 66 ans. Pas de d	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
3	Recherche ischémie myocardique chez patie	M	cmd	CAD	AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
4	Recherche d'ischémie myocardique chez un	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	cancer	Reche
5	Recherche d'ischémie myocardique dans le t	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
6	Recherche de viabilité dans le territoire de l'I	F	cmd	CAD	AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
7	Patiente de 57 ans. Recherche d'ischémie my	F	cmd	CAD	AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
8	Patient de 71 ans. Suivi de CMI. Dyspnée NYI	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
9	Patient de 70 ans, suivi de CMI. Pas de doule	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
10	Patient de 82 ans, suivi de CMI. Pas de doule	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	cancer	Suivi d
11	Patient de 53 ans, suivi de CMI. Asymptomat	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
12	Suivi de CMI chez patient de 64 ans. Majorat	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
13	Recherche d'ischémie myocardique chez pat	F	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
14	Recherche d'ischémie myocardique chez pat	F	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
15	Suivi de CMI chez patient de 68 ans, après ur	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
16	Suivi de CMI chez patient de 69ans. Pas de d	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
17	Patient de 66 ans, suivi de CMI. Evaluation d	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Suivi d
18	Suivi de CMI chez patient de 59 ans. Pas de d	M	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche
19	Patiente de 75 ans, Suivi de CMD. Pas de dou	F	cmd	CAD	No_AMI	No_CABG	No_PCI	no-revascularization	no_cancer	Reche

Figure 14 : Fraction du jeu de données d'entrainement pour l'IA, partie scintigraphie

BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC
Âge	Date coro	uivi de CM	Dyspnée	DTT	DTA	Diabète	ATCD PAC	TCD STENT	N.I	N.IDM	CFR	FEVG	Volumes	ion myocai
62	2019-03-28 00:00:00	OUI					MIG-IVA,	MIG-MID-I'	1 (A)		? (ANT)	N	N	
63	2019-02-22 00:00:00	OUI	OUI		OUI			CD	2 (IL)			P	N	
55	2019-03-28 00:00:00		OUI						1-2 (ASM, ILM)			N	N	
66	2019-03-12 00:00:00		OUI		OUI	OUI				2-3 (INF)		N	N	
67	2019-02-27 00:00:00		OUI			OUI			2-3 (L)			N	N	
79	2019-03-13 00:00:00				OUI				2-3 (I)			?	?	
47	2019-06-05 00:00:00					OUI						N	N	
84														
90	2019-04-08 00:00:00	OUI			OUI		Saph-Biss,	Saph-IVA	2 (ILM, ISE	1-2 (A, IA)		N	N	
66														

Figure 15 : Fraction des données d'entrainement, partie coronarographie

4.4 Bilan de la veille technique appliquée à mon cas particulier

- Les expressions régulières sont très pratiques et obtiennent de bonnes performances
- Les pc du CHU ne possèdent pas de grandes ressources de calcul
- Limitations de logiciels
- Grande vigilance quand à la sécurité des données

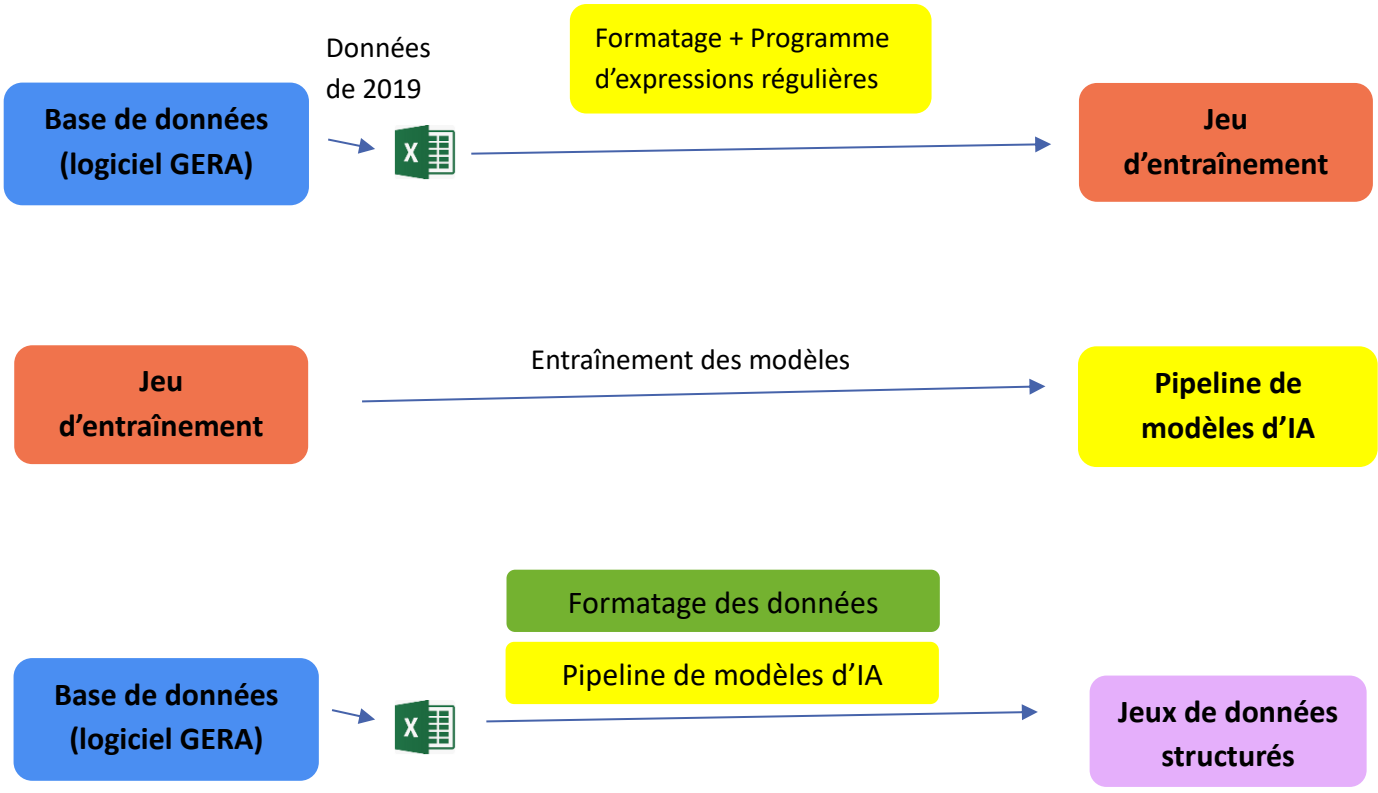
L'IA n'a pas beaucoup d'intérêt dans ce cas précis. Elle reste néanmoins la meilleure solution à long terme, permettant d'avoir des bons résultats sur des textes avec des syntaxes variées, l'inconvénient étant sa mise en place (jeux de données générés à la main).

4.5 Schématisation du cas pratique

Les comptes rendus de scintigraphies peuvent être extraits du logiciel dédié, GERA, sous forme de tableaux excel. Ces tableaux doivent être formatés (prétraitement des colonnes) pour être prédits par les modèles d'IA, afin d'obtenir un jeu de données structuré.

Les comptes rendus de coronarographie proviennent d'un autre logiciel, Cardioreport, sur lequel je n'ai pas encore eu de contact avec le responsable pour savoir s'il est possible de le requêter simplement.

Examen de scintigraphie



Examen de coronarographie

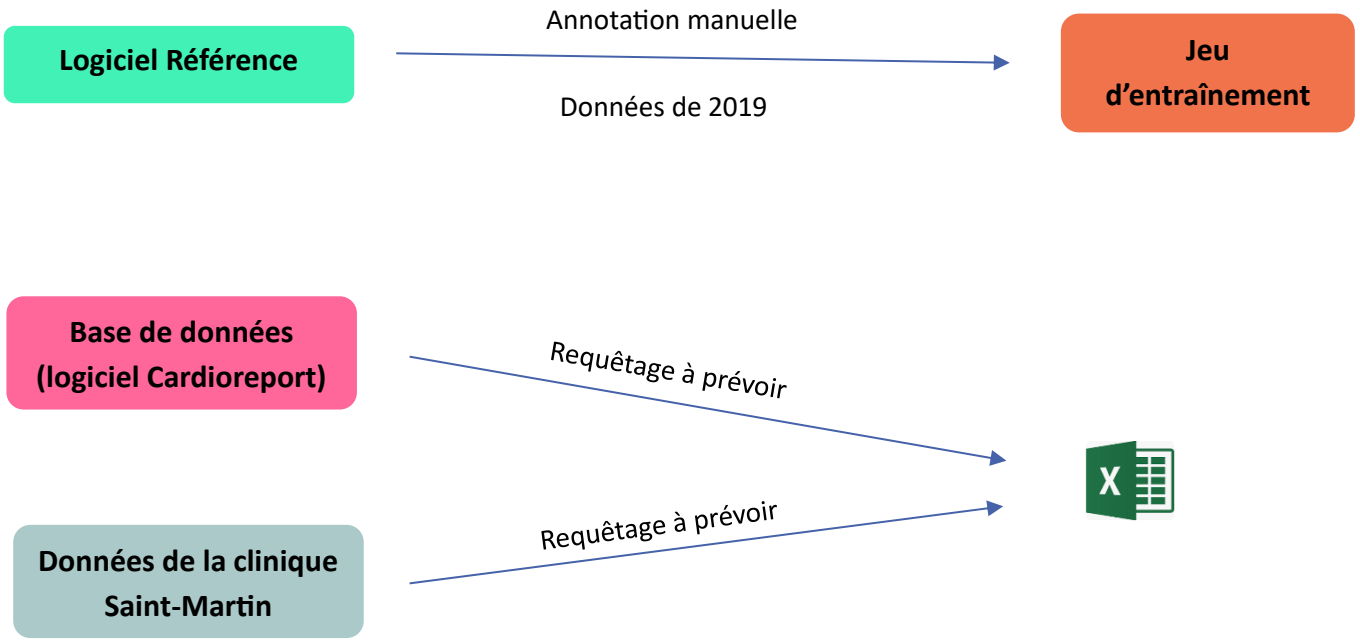


Figure 16 Schématisation du cas pratique

4.6 Procédé technique

Ayant une faible puissance de calcul sur mon ordinateur, je me suis restreint aux modèles qui demandent moins de ressources, je ne peux pas non plus utiliser des cartes graphiques sur des clouds car cela compromettrait la sécurité des données.

Les technologies utilisées :

- Les modèles scikit-learn
- Les programmes de reconnaissance d'expressions régulières

Les différentes informations à extraire sur les comptes rendus se classent en trois groupes :

- Les cibles catégoriques binaires
- Les cibles catégoriques multiclasses
- Les cibles continues

Pour les cibles catégoriques (binaires et multiclasses) :

J'ai réalisé un code qui trouve automatiquement le meilleur modèle parmi plusieurs paramètres.

Les encodages de textes en valeurs numériques testés sont :

- Tf-idf Vectorizer
- Count Vectorizer

Les modèles comparés pour ces types de cibles sont :

- AdaBoostClassifier
- SGDClassifier
- LinearSVC
- LogisticRegression
- RandomForestClassifier
- DecisionTreeClassifier
- KNeighborsClassifier
- MultinomialNB
- GradientBoostingClassifier

Les modèles utilisés pour les cibles continues sont des fonctions de reconnaissance d'expressions régulières. Les modèles de régression ne sont pas adaptés à ce type de problèmes, on veut extraire des entités, pas les calculer.

Les métriques que j'ai utilisées :

Pour les modèles de classification :

- Précision
- Rappel
- F1 score
- Aire sous la courbe ROC

Pour les fonctions régulières (cibles continues) :

- R2 score
- RMSE
- MAE

Voici la structure de mon notebook d'entraînement (modèles de classification) :

- Import des librairies
- Fonction qui :
    - Prétraite les données du jeu d'entraînement
    - Sépare le texte (« X ») et les cibles (les « y »)

- Pour chaque type de cibles (binaires / multiclasses) :
    - Pour chaque cible :
        - Séparation en jeu entraînement et de test stratifiés
        - Choix d'un transformer qui change le texte au format numérique
        - Instanciation d'une classe python qui entraîne puis teste des modèles. Enregistre le meilleur.
          Cette partie du code diffère pour les cibles binaires et multiclasses

Les cibles continues sont extraites par des expressions régulières. Des fonctions inscrivent des entités dans les colonnes dédiées de mon jeu de données lorsqu'elles reconnaissent des motifs spécifiques dans le texte.

Il s'agit de volumes, comme des fractions d'éjections de ventricules, en millilitres.

5. Résultats

Voici les tableaux qui regroupent les métriques de performances de mes modèles, pour chaque information à extraire.

Les modèles sont enregistrés sous forme de pipelines qui comprennent le transformer plus le modèle.

En production, les prédictions sont effectuées par des fonctions qui regroupent le prétraitement de texte et les pipelines.

	cmd	cmi	history_MI	cabg	pci	revascularization	cancer
traitement_de_texte		nettoyage de texte classique		nettoyage	nettoyage	nettoyage de texte	nettoyage
transformer_name		count vectorizer		tf-idf	tf-idf	count vectorizer	tf-idf
model_name	DecisionTreeC	DecisionTree	GradientBoost	DecisionT	DecisionT	DecisionTreeClassi	AdaBoost
precision	1	0,9573	0,9145	0,6897	0,9717	0,927	1
recall	1	0,9515	0,8742	0,6061	0,9493	0,931	0,8065
f1	1	0,9544	0,8939	0,6452	0,9604	0,929	0,8929
roc_auc_score	1	0,9587	0,9259	0,7967	0,9689	0,9488	0,9032

	sent_to_angio	spect_perfusion	nb_segment_isch	signif_ischemia	ischemia	nb_segment_IDM
traitement_de_texte	nettoyage de tex	nettoyage de text	nettoyage de texte	nettoyage de tex	nettoyage	nettoyage de texte c
transformer_name	count vectorizer	tfidf vectorizer	tfidf vectorizer	tfidf vectorizer	count vect	tfidf vectorizer
model_name	GradientBoostin	SGDClassifier	SGDClassifier	SGDClassifier	GradientB	DecisionTreeClassifi
precision	0,9539	0,9839	0,8512	0,942	0,9607	0,9151
recall	0,9797	0,9851	0,8649	0,9405	0,9942	0,9149
f1	0,9667	0,9845	0,8551	0,9409	0,9771	0,9127
roc_auc_score	0,984	trop compliqué à r	trop compliqué à m	trop compliqué à	0,9794	trop compliqué à me

	LBBB	AF	aap	antico	bb	ivabradine	statine	autre_hypolipemiant
traitement_de_texte	nettoyage	nettoyage	nettoyage	nettoyage	nettoyage	nettoyage de	nettoyage	nettoyage de texte clas
transformer_name	tf-idf	tf-idf	count vect	count vect	tf-idf	tf-idf	tf-idf	tf-idf
model_name	AdaBoost	GradientB	AdaBoost	AdaBoost	DecisionT	AdaBoostClas	AdaBoost	AdaBoostClassifier
precision	1	0,9398	0,9952	0,9892	0,9914	1	0,9975	1
recall	0,9565	1	0,9859	0,9583	0,9692	1	0,9778	0,988
f1	0,9778	0,9689	0,9905	0,9735	0,9802	1	0,9875	0,9939
roc_auc_score	0,9783	0,9962	0,9898	0,9784	0,9807	1	0,9874	0,994

	iec	diuretique	ica	insuline	ado	dt	effort	typique	NYHA
traitement_de_texte	nettoyage	nettoyage	nettoyage	nettoyage	nettoyage	nettoyage	nettoyage	nettoyage	nettoyage
transformer_name	count vect	tf-idf	count vect	count vect	count vect	tfidf vecto	tfidf vecto	tfidf vecto	tfidf vecto
model_name	AdaBoost	DecisionT	AdaBoost	DecisionT	AdaBoost	SGDClassi	SGDClassi	SGDClassi	SGDClassi
precision	0,9896	0,9869	0,9944	0,9833	1	0,8648	0,7797	0,8253	0,9212
recall	0,955	0,9557	0,9514	0,9077	0,9583	0,8716	0,8081	0,8378	0,9216
f1	0,972	0,9711	0,9724	0,944	0,9787	0,8557	0,7803	0,8254	0,9209
roc_auc_score	0,9716	0,9761	0,9748	0,9531	0,9792	trop comp	trop comp	trop comp	trop comp

	indication_exam	Hypertension	diabetes	tabacco	Dyslipidemia	PAR	familial_history_cad
traitement_de_texte	nettoyage de text	nettoyage de t	nettoyage	nettoyage	nettoyage de	nettoyage	nettoyage de texte cla
transformer_name	tfidf vectorizer	count vectorize	tf-idf	count vec	count vectoriz	count vec	count vectorizer
model_name	SGDClassifier	AdaBoostClass	AdaBoost	LogisticRe	AdaBoostClass	DecisionT	DecisionTreeClassifie
precision	0,9484	0,996	0,9739	0,9925	0,9866	1	0,9585
recall	0,9486	0,9939	0,9825	0,9851	0,9822	0,9245	0,972
f1	0,9478	0,9949	0,9782	0,9888	0,9844	0,9608	0,9652
roc_auc_score	trop compliqué à	0,9929	0,9854	0,9881	0,9808	0,9623	0,9774

	fevg_stress	fevg_repos	vtd_stress	vts_stress	vtd_repos	vts_repos
traitement_de_texte	classique: pas d'accents, de majuscules et de ponctuation, concaténation de					
transformer_name						
model_name	er					
precision						
recall						
f1						
roc_auc_score	ettre en place					
R2_score	1	1	1	1	1	1
RMSE	0	0	0	0	0	0
MAE	0	0	0	0	0	0

Figure 17 Performance des modèles pour l'extraction de données des scintigraphies

Dans mon tableau de scores, je répertorie pour chaque modèle :

1. Le prétraitement de texte effectué :
2. Le type d'encodage en valeur numérique
3. Le modèle
4. Les métriques de performance

On remarque que les modèles fonctionnant le mieux pour les cibles catégoriques binaires sont :

- DecisionTreeClassifier
- AdaBoostClassifier
- GradientBoostingClassifier
- LogisticRegression

Pour les cibles multiclasses c'est SGDClassifier qui obtient les meilleures performances.

6. Discussion

Les résultats sont excellents, sauf pour certaines cibles multiclasses : la douleur thoracique (dt), la douleur à l'effort (effort), la douleur typique ou atypique (typique).

Ces modèles sont sujet à erreurs car les encodages de texte que j'utilise comptent l'occurrence des mots dans le texte. Les modèles apprennent quels mots ont le plus d'impact sur la cible à chercher.

Le problème est que la syntaxe grammaticale dans le cas de ces cibles dépend beaucoup de la négation, on a souvent « pas de douleur thoracique » ou « douleur thoracique », or le nombre de pas dans le texte est assez indépendant de la douleur thoracique, exemple « […] présentant des douleurs thoraciques constrictives au repos. Pas de dyspnée. Pas d'ATCD CV » le patient présente des symptômes de douleur thoracique mais pas de symptômes sur les autres plans, l'algorithme est donc confus.

La meilleure solution serait de rester sur des expressions régulières dans ces cas précis, ou bien d'utiliser des modèles de traitement de langage de pointe comme les modèles de langue utilisables avec la librairie « Hugging face ». L'avantage énorme de ces modèles est qu'ils comprennent la relation entre les mots. Cependant il faut une carte graphique.

Je peux tenter de développer de l'hyperparamétrage de modèles dans mon code mais l'amélioration sera probablement mauvaise comparée à l'exactitude qu'obtiendraient les expressions régulières.

Limites rencontrées :

- L'hôpital est un environnement protégé niveau machines et réseau, cela diminue la flexibilité du développeur.

- Je suis sur Windows et j'aurais voulu essayer la librairie d'auto ML « auto sklearn » qui ne fonctionne qu'avec Linux.

- Les données ne doivent pas sortir de l'ordinateur, question de sécurité.

7. Conclusion

Le projet porte sur l'extraction de caractéristiques à partir de comptes rendus, des modèles de classification de la librairie « Scikit-learn » ont été utilisés pour extraire les caractéristiques de type variables catégoriques, pour les caractéristiques de type variables continues, les expressions régulières ont été utilisées.

Une application « Tkinter » de mise en production de ces modèles a été développée. Il ne reste qu'à optimiser certains modèles, probablement par un retour aux expressions régulières.

Les gros points bloquants pour moi ont été l'écriture du code pour trouver les meilleurs modèles, il me manque une méthode à suivre pour aborder et résoudre les problèmes.

Il y a une grande diversité de paramètres à tester, au niveau des prétraitements de textes, des encodages numériques, des modèles et des hyperparamètres. Tout cela sur des dizaines de cibles, qui sont de différents types. J'ai eu le sentiment que j'aurais été plus à l'aise sur une interface graphique pour tester les différents paramètres, je n'aime pas le système des notebooks.

J'aurais aimé savoir comment un expert aurait traité le projet.

7. Références

1. Lam H, Nguyen F, Wang X, Stock A, Lenskaya V, Kooshesh M, Li P, Qazi M, Wang S, Dehghan M, Qian X, Si Q, Polydorides AD. An accessible, efficient, and accurate natural language processing method for extracting diagnostic data from pathology reports. J Pathol Inform. 2022 Nov 8;13:100154. doi: 10.1016/j.jpi.2022.100154. PMID: 36605108; PMCID: PMC9808011.

2. ***https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9751356/.***