Projet chef d’œuvre

Prédire l’état de santé d’un patient après un examen de scintigraphie myocardique de perfusion

Geoffroy Daumer

Table des matières

[1. Présentation de la problématique 3](#_Toc158369400)

[2. Présentation des données 4](#_Toc158369401)

[2.1 Présentation générale 4](#_Toc158369402)

[2.2 Analyse exploratoire des données 5](#_Toc158369403)

[3. Contraintes réglementaires et techniques 7](#_Toc158369404)

[4. Gestion du projet 9](#_Toc158369405)

[5. Etat de l’art 10](#_Toc158369406)

[6. Stratégie de résolution de la problématique 11](#_Toc158369407)

[7. Conclusion 13](#_Toc158369408)

[Références 21](#_Toc158369409)

[Annexe 22](#_Toc158369410)

# 1. Présentation de la problématique

L’objet de mon projet chef d’œuvre est la prédiction de l’état de santé mentale d'employés

d’un patient en fonction du compte rendu écrit d’un examen : la scintigraphie myocardique de perfusion. Cet examen ne donne pas directement l’état de santé du patient, mais fournit des informations pour déterminer si le patient est susceptible d’être malade.

Si il est susceptible d’être malade, on l’envoie en coronarographie. La coronarographie est un examen invasif (comportant des risques). Elle permet de diagnostiquer si le patient est réellement malade, ainsi que de prévoir un traitement ou une opération adaptée.

L’intérêt du modèle de prédiction serait donc d’éviter aux patients non malades d’être envoyés en coronarographie. Il faudrait en premier lieu maximiser le rappel pour ne pas oublier de faux négatifs, et ensuite maximiser la précision pour minimiser les faux positifs.

Le modèle serait entraîné à partir d’un jeu d’entrainement d’une année. Le jeu se compose des comptes rendus de scintigraphies des patients de l’hôpital de Caen de 2019, ainsi que des coronarographies quand ces patients en ont réalisées. Le fait d’avoir les coronarographies dans le jeu de données permet d’avoir la sortie, malade ou non.

Patient

Scintigraphie positive ?

Oui

Non

Coronarographie positive ?

Gold standard (1)

Non

Oui

Figure : Compréhension de la problématique

1 : Le Gold standard se dit d’un examen qui par convention détermine si le patient est malade.

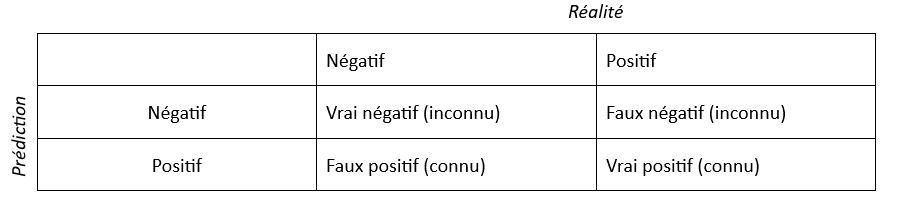
Les patients prédits négatifs à la scintigraphie ne sont jamais envoyés en coronarographie. On ne sait pas si ils sont vraiment négatifs. Cela n’empêche pas de réaliser un modèle de prédiction sur les données existantes.

Figure 2 : Matrice de confusion de la réalité des coronarographies par rapport à la prédiction des scintigraphies

# 2. Présentation des données

### 2.1 Présentation générale

Les données d’entraînement sont un fichier Excel qui comporte :

* Les comptes rendus des patients ayant réalisé une scintigraphie en 2019. (1)
* Des informations extraites de ces comptes rendus par des expressions régulières. Exemple : maladie du cœur connue ? douleur thoracique ? etc. (2)
* Des informations extraites à la main des comptes rendus de coronarographie, lorsque les patients en ont réalisées.

(1)

(2)

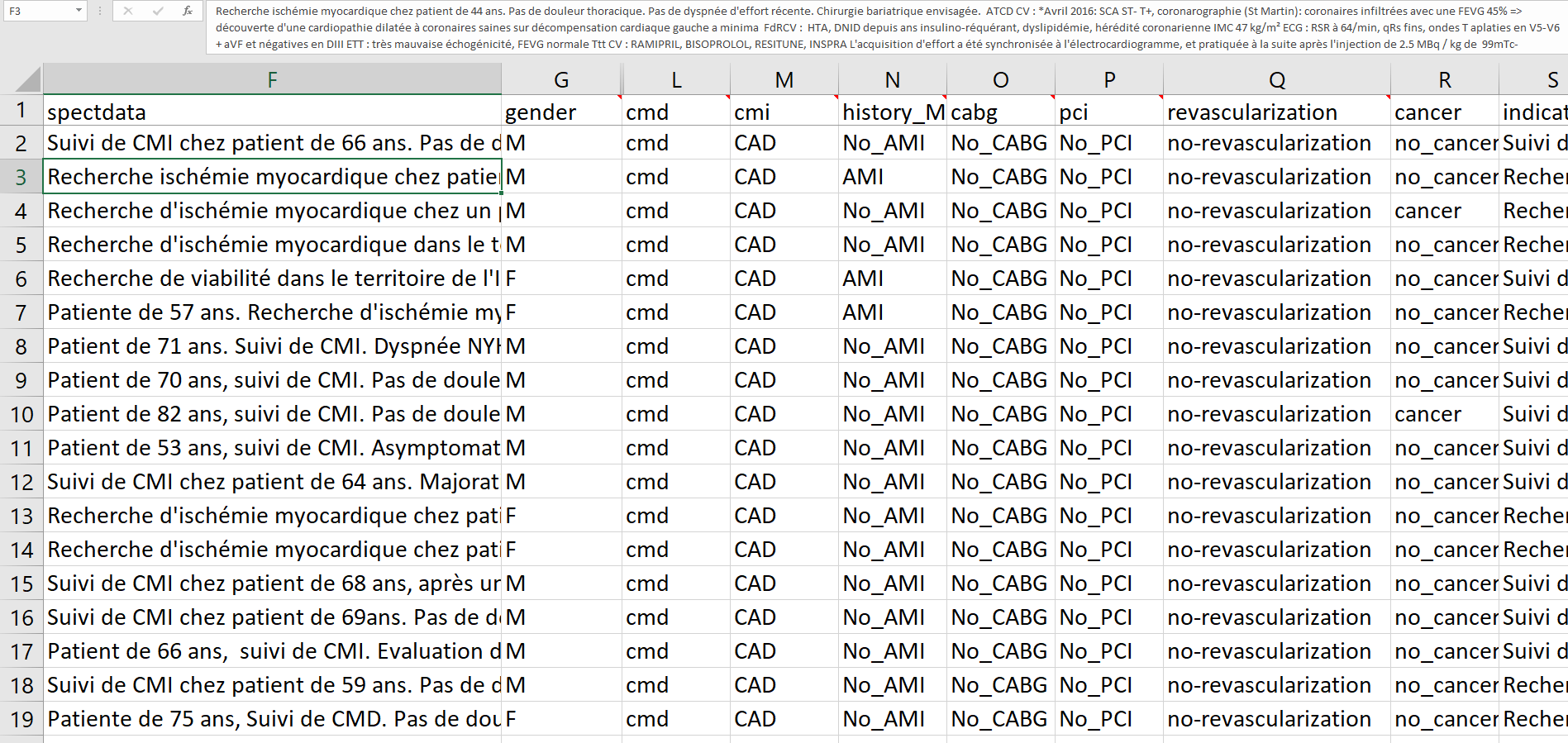


Figure : Fraction du jeu d’entrainement, partie scintigraphies

(3)

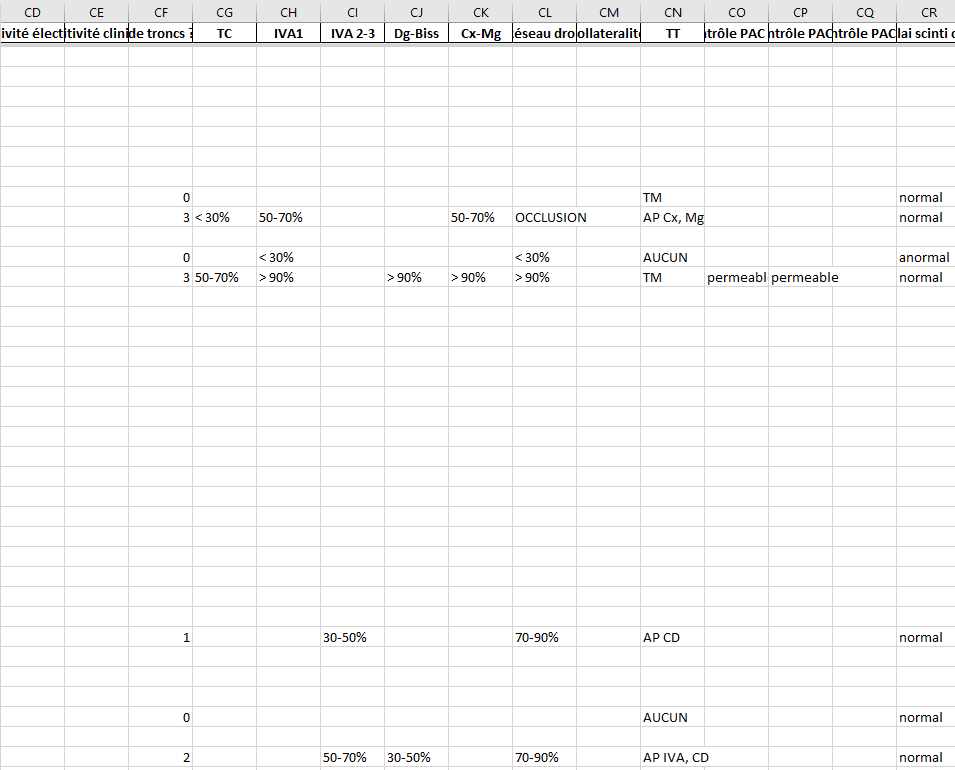


Figure : Fraction du jeu d’entrainement, partie coronarographies

Les colonnes d’entrée du modèle de prédiction seront le texte de compte rendu de scintigraphie (1), ou les caractéristiques extraites du texte (2), selon les performances obtenues.

La sortie est malade / non malade. On génèrera cette information à partir de la colonne traitement (TT) des données de coronarographies (3). Les patients positifs sont les patients ayant réalisé une coronarographie et ayant un traitement (autre que "AUCUN") à la suite de l'opération.

### 2.2 Analyse exploratoire des données

#### 2.2.1 Compréhension des données

Voici les différents types de colonnes pour la partie scintigraphie du jeu de données :

* Compte rendu écrit de scintigraphie (1).
* Des colonnes d'identification (nom, numéro de dossier…)
* Caractéristiques générales : IMC, sexe, âge, date d'examen
* Raison de la venue du patient (suivi, recherche maladie, bilan pré-opératoire)
* Cancer
* Antécédents cardiaques (cardiomyopathie ischémique, infarctus, pontage artéro-coronarien, implantation de stent…)
* Facteurs de risques cardiovasculaires (hypertension, dyslipidémie, hérédité…)
* Données cliniques (bloc de branche gauche, fibrillation atriale)
* Traitements (antiagrégants, anticoagulants…)
* Symptômes (douleur thoracique ? typique ?)
* Type de stress (les patients sont soumis à un stress lors de l’acquisition d'images, via un produit ou via un effort physique)
* Volumes liés à l’activité cardiaque (fraction d’éjection du ventricule gauche…)
* Données liées à l'interprétation des images : présence d’ischémie, d'infarctus, nombre de segments touchés, significativité, envoi ou non du patient en coronarographie.

Pour la partie coronarographie :

* Colonnes d'identification (numéro de dossier, initiales…)
* Date examen
* Délai scintigraphie / coronarographie normal (certains patients ont réalisé leur coronarographie avant leurs scintigraphie) ?
* Origine de l’examen (clinique du CHU / clinique Saint-Martin)
* Antécédents d’interventions cardiaques (stents / pontages)
* Contrôles d’interventions cardiaques (stents / pontages)
* Taux d’occlusion des différentes artères coronaires
* Nombre de troncs significativement touchés
* Traitement proposé à la suite de l’examen
* Une colonne finale qu’on peut créer à partir du traitement : malade / non malade.

Il y a 3698 lignes (nombre de scintigraphies) et 95 colonnes. La majorité des colonnes sont de type objet, quelques-unes sont de type float ou int.

#### 2.2.2 Préparation des données

* [Dropping irrelevant columns and rows
* -> Examine patient outcomes (post-scinti positivity / post-coronary positivity) according to reason for visit
* Renaming columns
* Feature Creation]

Création de la sortie malade / non-malade (explications en 2.1)

Suppression des colonnes en double (gender / gender3, dyspnee / NYHA…), des colonnes identifiantes, des informations relevées sur les coronarographies (ces informations ne rentreront pas dans le modèle car elles ne proviennent pas des scintigraphies).

Suppression des patients adressés en coronarographie et n'en ayant pas réalisées (valeurs manquantes).

Il y a 619 positifs (malades) et 2970 négatifs dans le jeu.

Parmis les 690 patients ayant réalisé une coronarographie sur les 3700 scintigraphies, 620 sont des vrais positifs (ont un traitement ou une intervention). Ce chiffre dénote une excellente précision pour les scintigraphies.

# 3. Contraintes réglementaires et techniques

* La confidentialité des données rend impossible l’utilisation de serveurs distants, comme Google Colab, les modèles doivent donc être entrainés et utilisés en local.
* L’ordinateur n’a pas de GPU, donc les modèles demandant trop de ressources ne peuvent pas être utilisés.
* Le système d’exploitation est Windows.
* Il y a des contraintes pour utiliser des nouveaux logiciels car il y a une sécurité réseau.

Il y a donc plusieurs restrictions, il est difficile de tester les technologies les plus récentes dans cette configuration.

Pour réaliser un travail de recherche en santé, il faut obtenir une autorisation d’un comité d’éthique : le CLERS (Comité Local d’Ethique pour la Recherche en Santé).

Voici le tableau récapitulant les obligations de conformités concernant les travaux de recherche sur des données de santé :

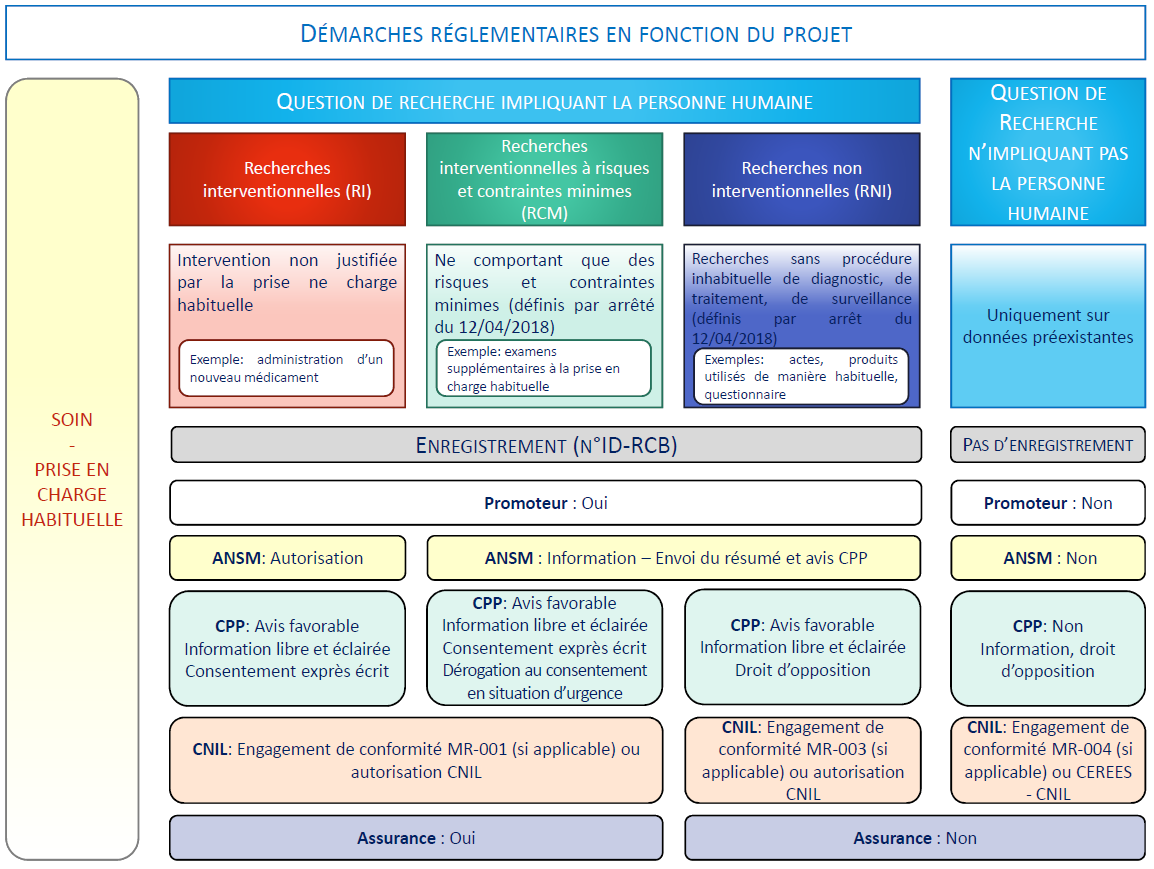


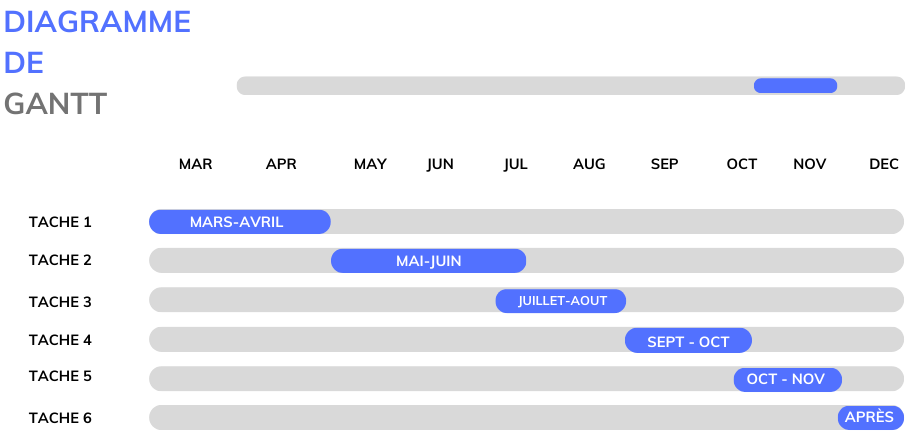
Figure : Schéma des réglementation pour les travaux de recherche impliquant les données des patients

Comme on le voit sur la figure 2, dans le cas de recherches portant uniquement sur des données préexistantes, pas besoin de demander une autorisation au CPP (Comité de Protection des Personnes). Il faut être conforme à la MR-004, qui est une liste de règles établies par la CNIL (Commission nationale de l'informatique et des libertés).

Obligations MR-004 :

* Données collectées strictement nécessaires aux objectifs de la recherche.
* Transfert hors UE doit-être strictement nécessaire. Le cas échéant, les données doivent êtres pseudonymisées au minimum.
* Le projet doit-être enregistré sur le « Health Data Hub » (plateforme d’intérêt public de données de santé).
* Données des patients pouvant 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 patients et leurs noms et prénoms.
* Une information générale et individuelle concernant la réutilisation des données doit être effectuée vis-à-vis des patients (possibilité de demander une dérogation en cas d’effort disproportionné).
* Anonymiser au maximum les données lors de la publication.

# 4. Gestion du projet



Tache 1 : Recherche sur les réglementations des projets de recherche

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.

Il n’y a pas eu vraiment de gestion de projet, je faisais ce que me disaient mes tuteurs. Certaines choses peuvent prendre beaucoup de temps quand on apprend à coder et on ne sait pas toujours quoi.

# 5. Etat de l’art

L’extraction de données sur du texte se fait majoritairement avec des expressions régulières dans les articles scientifiques. C’est une méthode relativement simple à déployer. Cependant, elle se limite rapidement avec la diversité du texte et la complexité des éléments à extraire. Quant aux algorithmes de deep learning, ils nécessitent d’annoter une grande quantité de données.

Jusqu’à maintenant, il n’y a pas vraiment de méthodologie développée pour ces applications. On remarque une absence de tâches communes qui permettraient de comparer différentes méthodes de développement, ainsi qu’une absence de métriques communes.

De nouvelles pipelines basées sur des modèles pré-entraînés comme GPT-3 pourraient résoudre en partie le problème du manque de données d’entrainement (1).

Expressions régulières :

Exemple d’application :

Recherche du groupe de chiffres après « volume télé-systolique » dans un compte rendu. Pour connaitre ce volume, qui sera classé dans une colonne appropriée.

Ou, recherche de plusieurs motifs : exemple « cmd / cardiopathie dilatée / cardiomyopathie dilatée », si l’un d’eux est présent : mettre « oui » dans la colonne « cmd » (cardiomyopathie dilatée)

* Les expressions régulières sont très pratiques et obtiennent de bonnes performances, elles peuvent notamment extraire des entités textuelles, là où des modèles complexes doivent intervenir dans l’IA (réseaux de neurones, LLMs).

# 6. Stratégie de résolution de la problématique

Pour prédire l’état de santé du patient à partir du jeu d’entrainement, on a deux possibilités à comparer :

1. Prédiction à partir du texte

2. Prédiction à partir des informations extraites du texte

Si on entraîne des modèles d’IA sur un jeu de données réalisé par un programme d’expressions régulières, les modèles ne seront pas meilleurs. Pour que l’IA soit utile, il faut qu’elle apporte des réponses soit en apprenant sur des données totalement vérifiées, soit en ayant une intelligence propre, c’est le cas des modèles de langues.

Les Comptes rendus de scintigraphies sont requêtables par le logiciel GERA, on obtient une série d’examens sur une fourchette de date données, en format Excel. Il faut ensuite rechercher un groupe de mot spécifique aux scintigraphies dans une colonne, pour trier les examens scintigraphiques.

Coronarographie :

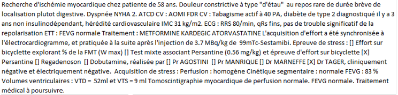
Informations relevées à la main

Certaines données à extraire sont de type catégoriques (exemple : cardiopathie dilatée ? oui ou non), d’autres de type continues (du type volume correspondant à la fraction d’éjection du ventricule gauche).

Les données continues sont des valeurs extraites du texte, les algorithmes de classification et de régression ne fonctionnent pas dans ce cas, je suis donc resté sur des expressions régulières.

Pour les données catégoriques, on peut utiliser des modèles de classification. Du machine learning a été fait sur les modèles catégoriques de « scikit-learn », le texte étant au préalable formaté numériquement avec des fonctions qui comptent l’occurrence des mots dans le texte (Tf-Idf, Countvectorizer).

Compte rendu



Génération d’une colonne :

Donnée catégorique -> modèle scikit-learn

Donné continue -> expression régulière

Idéalement : modèle de langue

Jeu de donnée structuré

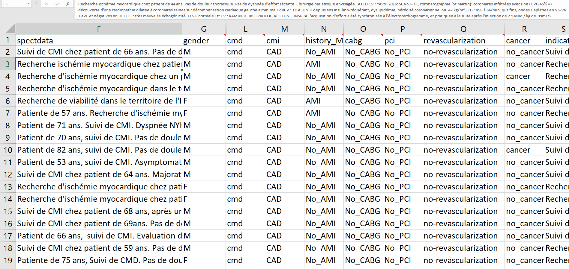


Figure : Choix technique

# 7. Conclusion

Ce qui a pris beaucoup de temps c’est le machine learning pour les modèles d’extraction de texte. Il fallait trouver les bons compromis, compte tenu des contraintes de performance de l’ordinateur.

Faire du machine learning en soi est aussi compliqué, il n’y a pas de méthodologie toute faite pour résoudre ce type de problèmes.

L’entrainement de modèles d’IA sur des données générées avec un programme ne m’a jamais paru logique. Pour apporter une contribution dans l’état actuel des choses, il faut utiliser les modèles de langues pour faire les prédictions, et vérifier les résultats.

Il y a également une application de l’IA intéressante qui reste possible : Prédire le risque d’être malade sur un compte rendu de scintigraphie.

J’ai constaté que le machine learning est un domaine très complexe, le gros soucis est mon besoin de méthode claire pour faire avancer mon travail. L’objectif était d’entraîner de l’IA sur un jeu d’entraînement que l’on peut obtenir avec des expressions régulières, l’idée était que l’IA développe sa propre compréhension du texte, mais elle sera forcément liée à la logique des expressions régulière, et donc sera obsolète.

Les améliorations envisageables seraient d’utiliser des modèles de langue pour extraire les données du texte, car ils comprennent le langage et ne nécessitent pas de surentrainement.

Il faudrait aussi faire un modèle de prédiction du risque d’être malade et surtout l’implémenter aux logiciels médecins.

# Références

1. **Art, Process Extraction from Text: Benchmarking the State of the.** [En ligne] https://arxiv.org/pdf/2110.03754.pdf.

# Annexe