Projet chef d’œuvre

Application d’extraction de données sur du texte utilisant l’IA

Geoffroy Daumer

Table des matières

[Introduction 3](#_Toc154670490)

[Partie 1 4](#_Toc154670491)

[I. Compréhension du besoin client 4](#_Toc154670492)

[I.1. Enjeux réglementaires 5](#_Toc154670493)

[I.2. Utilisateurs projet 6](#_Toc154670494)

[I.3. Schéma fonctionnel utilisé pour le projet 7](#_Toc154670495)

[I.4. Macro-planning initial 8](#_Toc154670496)

[II. Etat de l’art de l’extraction de données sur du texte 9](#_Toc154670497)

[III. Eléments de conception technique 10](#_Toc154670498)

[IV. Choix techniques liés au projet 12](#_Toc154670499)

[V. Réponse finale apportée 13](#_Toc154670500)

[V.1. Analyse exploratoire 13](#_Toc154670501)

[V.2. Résultats 15](#_Toc154670502)

[Partie 2 16](#_Toc154670503)

[Environnement de développement 16](#_Toc154670504)

[La gestion de projet 16](#_Toc154670505)

[Retours d’expérience sur les outils, techniques et compétences à l'œuvre tout au long du projet 16](#_Toc154670506)

[Partie 3 bilan de projet et améliorations envisageables 17](#_Toc154670507)

[Conclusion 18](#_Toc154670508)

[Références 19](#_Toc154670509)

# Introduction

Mon alternance s’est déroulée au CHU de Caen, en médecine nucléaire.

Le but de ce stage était d’utiliser l’IA pour structurer des comptes rendus médicaux en jeux de données analysables. Ces comptes rendus proviennent d’examens de scintigraphies myocardiques de perfusion (fonctionnement des artères du cœur), et de coronarographie (anatomie des artères du cœur). L’analyse des comptes rendus permet de distinguer l’importance de différentes caractéristiques pour la détermination d’une maladie des artères coronaires.

Pour les coronarographies, les données d’entraînement n’ont pas été générées à la main, mais avec un programme, cela soulève plusieurs questions quant à l’utilisation de l’IA.

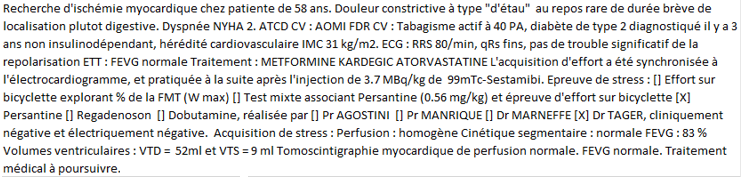
L’extraction de données s’est concentrée sur les scintigraphies, question de praticité. Des modèles « scikit-learn » ont été utilisés, en combinaison avec des expressions régulières.

# Partie 1

## I. Compréhension du besoin client

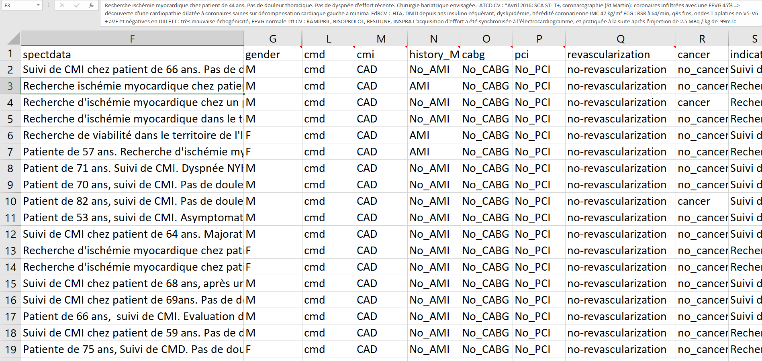
Notre but est de développer un programme d’IA performant d’extraction de caractéristiques sur des comptes rendus de scintigraphies. Le but final est de contribuer à un projet de recherche de comparaison des performances diagnostiques de la scintigraphie avec la coronarographie.





Compte rendu

Patient



Données structurées

Figure Processus de transformation des données

Un jeu d’entraînement fait avec un programme d’expressions régulières est réalisé pour 2019 sur les scintigraphies, pour les coronarographies le jeu est réalisé à la main sur cette même année. Ces jeux sont supposés être vrais, et servir à entraîner des modèles d’IA.

### I.1. Enjeux réglementaires

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 :

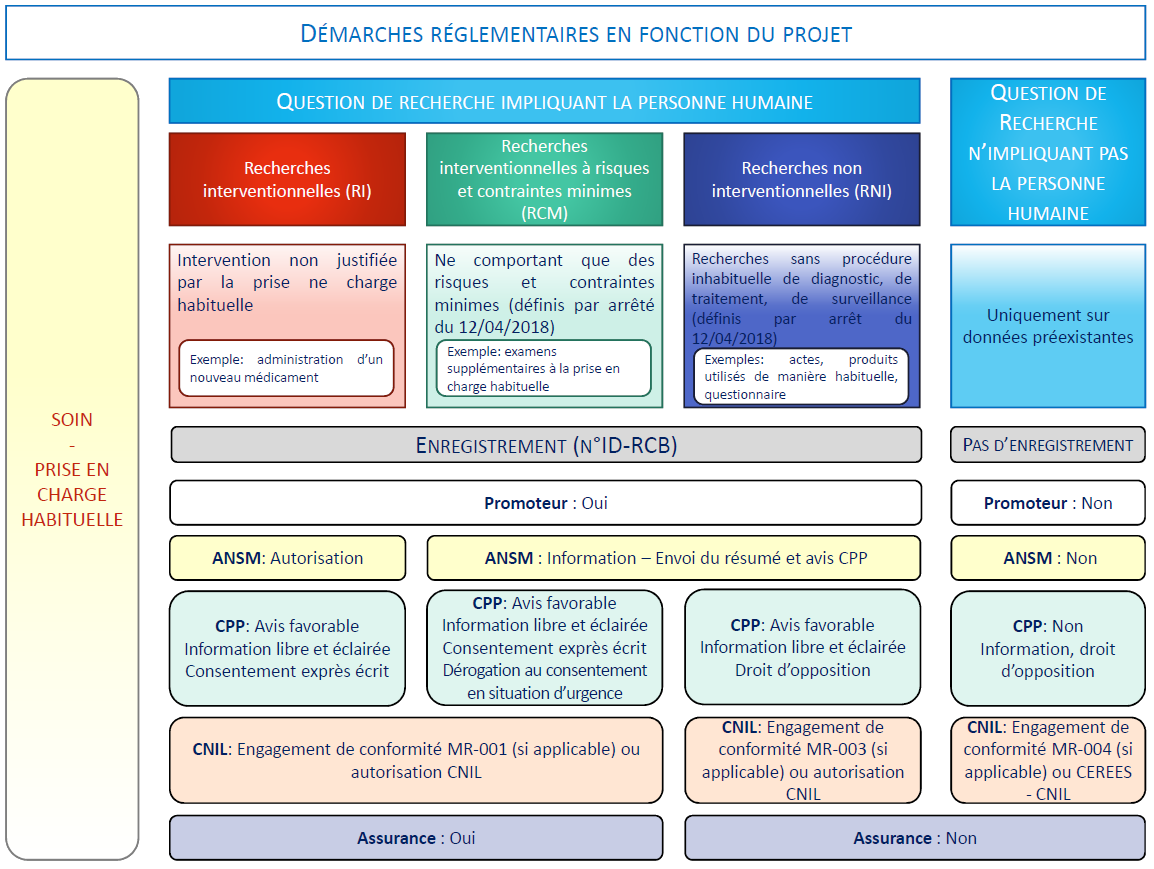


Figure : Schéma des réglementation pour les travaux de recherche impliquant les données des patients

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.

### I.2. Utilisateurs projet

L’objectif côté utilisateurs, serait de développer une application qui permette aux médecins chercheurs de structurer les données en sortie de la base de données.

Il est aussi envisagé de développer un modèle de prédiction de risque de maladie en fonction du compte rendu, pour assister les médecins dans leurs diagnostics.

### I.3. Schéma fonctionnel utilisé pour le projet

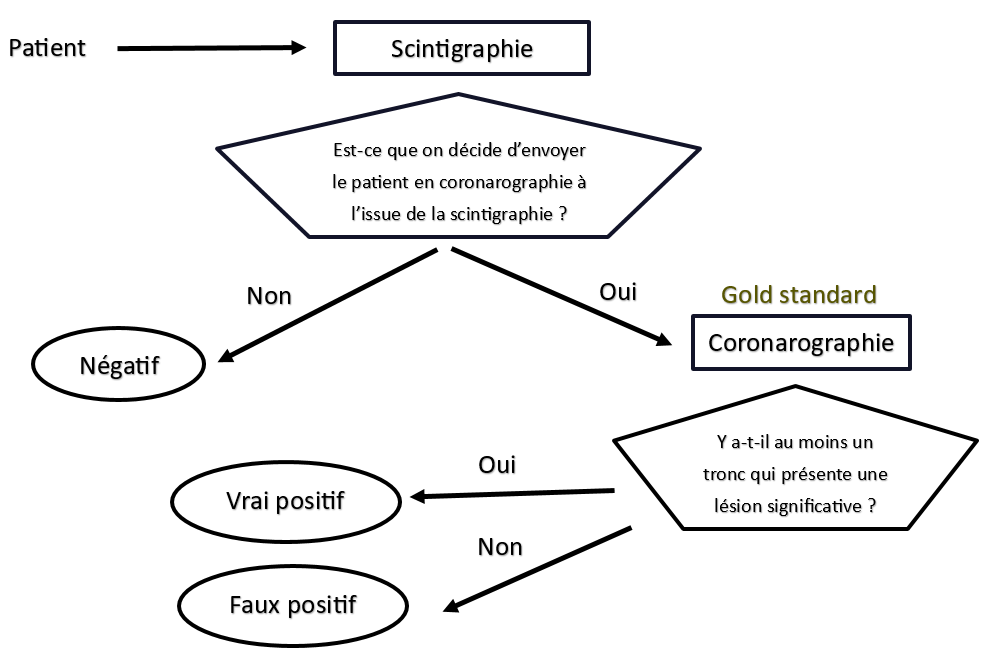


Figure : Schéma explicatif de la comparaison des performances diagnostiques de la scintigraphie myocardique de perfusion

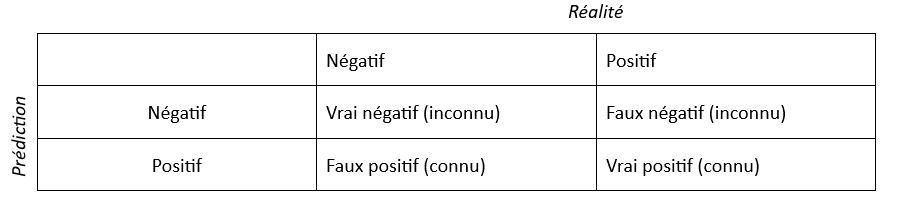
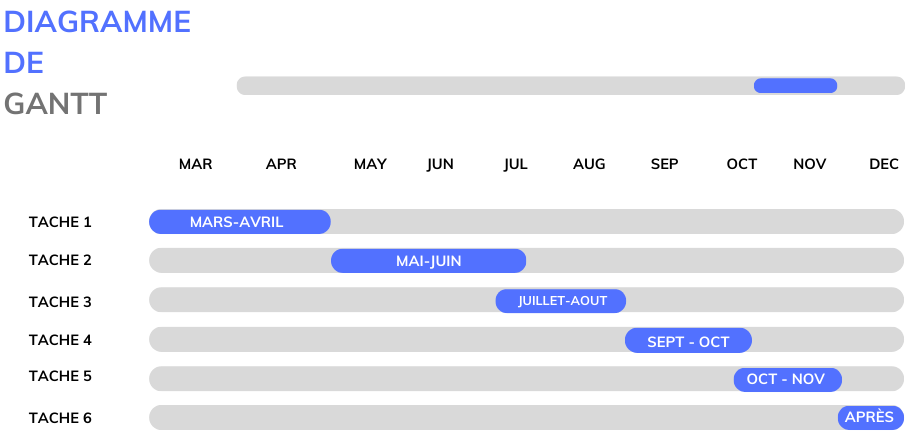


Figure 4 : Matrice de confusion

Comme nous le voyons sur la figure 3, la coronarographie est le « gold standard » : c’est par convention l’examen qui vérifie la présence de la maladie chez le patient.

Les patients prédits négatifs (à la scintigraphie) ne sont jamais envoyés en coronarographie. On ne peut pas savoir si on a fait une erreur de prédiction dans ce cas-ci.

### I.4. Macro-planning initial



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.

Figure : Diagramme de Gantt du projet

### II. Etat de l’art de l’extraction de données sur du texte

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)

## III. Eléments de conception technique

Jeux d’entrainement :

Les jeux d’entraînement générés englobent les examens de 2019.

Scintigraphies :

Généré avec un programme de fouille d’expressions régulières en langage R.

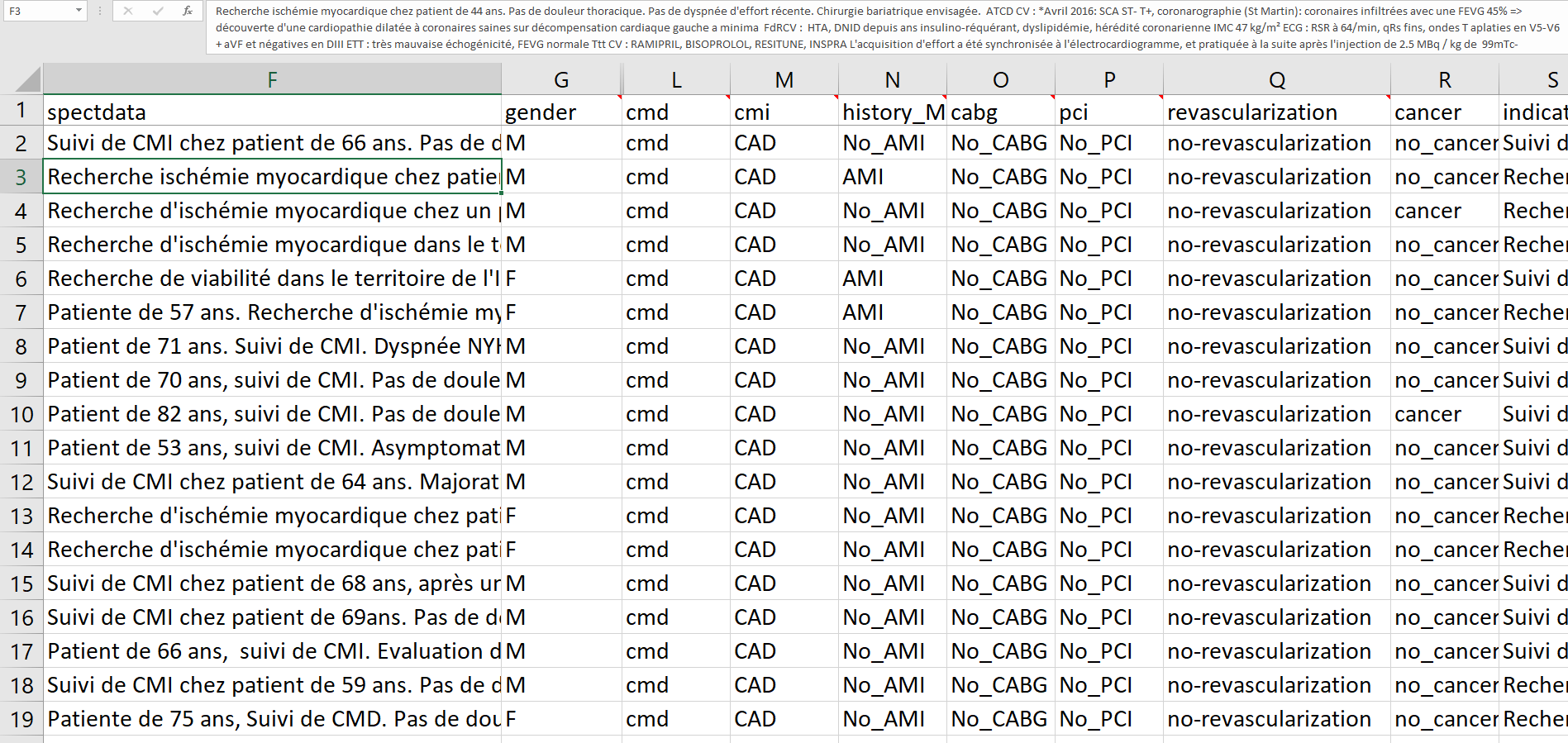


Figure : Fraction du jeu d'entrainement des scintigraphies

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

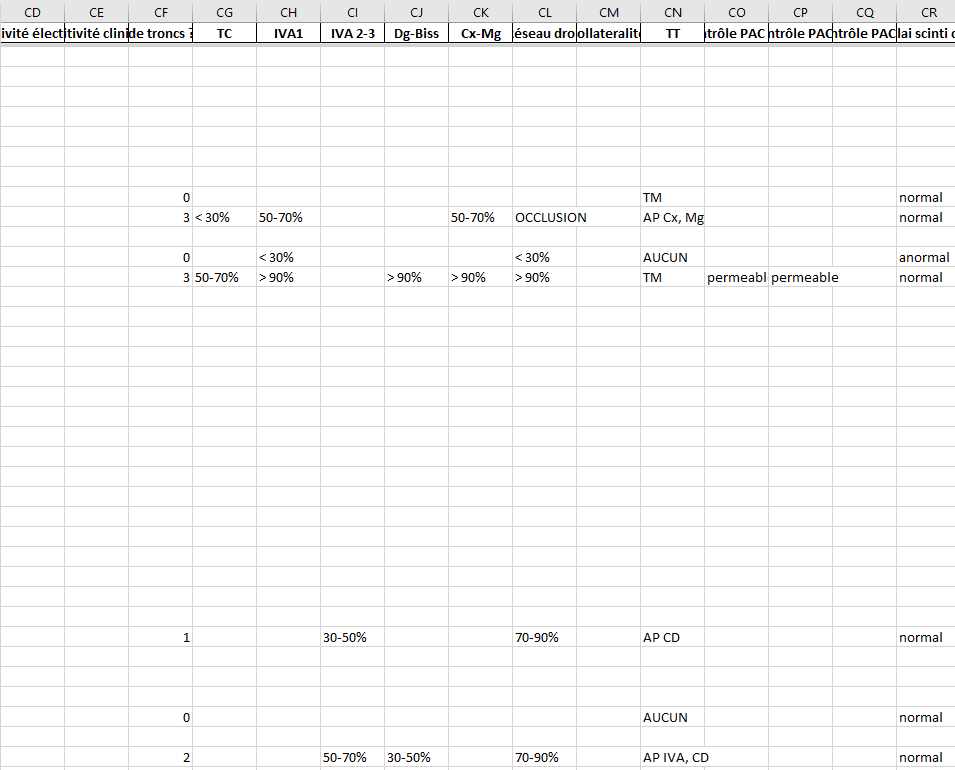


Figure : Fraction du jeu d’entrainement des coronarographies

La colonne TT signifie le traitement prescrit, elle permet de déterminer si le patient est réellement malade, et sera utile à des potentiels modèles de prédiction de risques d’être malade.

## IV. Choix techniques liés au projet

* 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).
* 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.

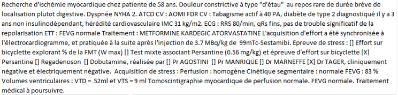
Certaines données à extraire sont de type catégoriques (du type cmd oui ou non), d’autres de type continues (du type valeur du volume télé systolique).

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).

Il est prévu de tenter les modèles de langues en local, ils se sont beaucoup démocratisés en 2023. Cela permettrait d’apporter une contribution au projet avec l’IA.

Compte rendu



Génération d’une colonne :

Donnée catégorique -> modèle scikit-learn

Donné continue -> expression régulière

Idéalement : modèle de langue

Jeu de donnée structuré

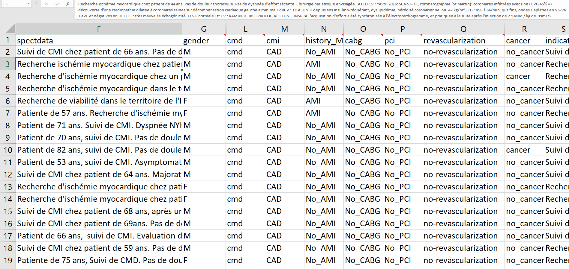


Figure : Choix technique

## V. Réponse finale apportée

### V.1. Analyse exploratoire

Mon jeu d’entraînement est un tableau Excel de 3699 patients, ce sont les patients ayant effectués une scintigraphie sur l’année 2019, les coronarographies réalisées quand il y en a y sont répertoriées.

#### V.1.1. Création de la colonne de risque de maladie

On a créé une colonne « maladie », qui servira à la prédiction de risque d’une maladie, cette colonne est donc créée à partir des informations obtenues via les coronarographies. On a défini trois valeurs possibles: malade, athérosclérose, pas malade.

Coronaropathie significative lorsque le patient a une artère avec une lésion supérieure à 50%

Coronaropathie non significatives (présence de plaques d’athéromes) lorsque le patient n’a pas de lésion significative, mais qu’il a eu un traitement médical de prescrit.

Sain, pour les patients ayant aucun traitement en sortie de coronarographie, ou n’ayant pas été envoyés en coronarographie.

Répartition des malades :

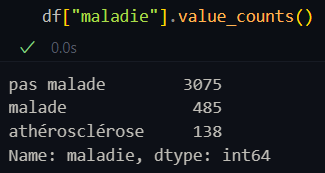


Figure : Répartition des cas par rapport aux patients ayant réalisé une scintigraphie

Ces données serviront à un modèle de prédiction de risque de maladie (pour malade et athérosclérose il vaut mieux envoyer le patient en coronarographie).

#### V.1.2. Compréhension des données

Pour la partie scintigraphies, il y a :

* Des colonnes connues de la base de donnée, ces informations pourraient être des informations en plus pour l’entraînement : colonnes identifiantes, texte, IMC, sexe, âge.
* Des colonnes inconnues de la base de donnée (issues de la fouille de texte)

Dans ces colonnes issues de la fouille, on a :

* Raison de la venue du patient (suivi, recherche maladie…)
* Caractéristiques du patient (cancer ?)
* Antécédents cardiaques (cardiomyopathie ischémique, infarctus, pontage artéro-coronarien, implantation de stent)
* Facteurs de risque 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)
* Volumes liés à l’activité cardiaque (télé-diastolique…)
* Résultats examen (normal / anormal, ischémie significative ?)

Pour la partie coronarographie, il y a :

* Date examen, délai entre coro et scinti normal ? (certaines coronarographies sont anciennes), origine (CHU / Saint-Martin)
* Raison venue patient
* Antécédents d’interventions cardiaques (stents / pontages)
* Symptômes (douleur)
* Facteurs de risque (diabète)
* Données cliniques (sidération myocardique, positivité électrique)
* Résultats (localisation ischémies / infarctus, normalité des volumes, nombre de troncs présentants une lésion significative, pourcentage d’attente de chaque artères principales, traitement, contrôle de stent / pontages

### V.2. Résultats

# Partie 2

## Environnement de développement

Le développement de l’application et des modèles s’est fait en Python sur l’éditeur VSCode.

## La gestion de projet

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.

## Retours d’expérience sur les outils, techniques et compétences à l'œuvre tout au long du projet

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.

# Partie 3 bilan de projet et améliorations envisageables

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.

# Conclusion

# Références

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