**[Open Source] Template de readme + tags**

**Faire une copie, ne pas compléter ce template**

| Suivi de modification / validation : rédaction d’un readme à ajouter à la page BOAS | | |
| --- | --- | --- |
| Fonction | Qui | Date |
| Porteur de la demande |  |  |
| Direction scientifique |  | JJ/MM/2023 |

# Informations synthétiques

| Plan détaillé | |
| --- | --- |
| Définition d’un angle | Lister les différents éléments présents sur un format de readme standardisé pour les résultats de la BOAS |
| Date de publication | À déterminer |
| Proposition d’une structure | * Objectifs de l’algorithme * Titre du projet * Auteur * Méthodologie * Données utilisées * Validation * Date de mise à jour * Maintenance * Comment installer l’algorithme * Comment utiliser l’algorithme * Support * Contributions * Crédits * Licence et conditions d’utilisation * Autres |

## 

# Gabarit du readme

*La documentation pourra être renseignée dans ce template ou directement sur git dans le format markdown (format natif du read.me), selon les préférences et compétences git du porteur de projet.*

*Un exemple est rédigé dans les cases pour HEPAVIRALGO à titre indicatif, penser à bien créer une copie de ce document et ne pas écrire par-dessus*

## Objectifs de l’algorithme :

## *Élaborez sur l’algorithme en expliquant ce qu'il fait, dans quel contexte il peut ou non être utilisé et sur quelles données ou quelles populations cibles, quelle région géographique*

| Les algorithmes HepavirAlgo sont des algorithmes de ciblage appliqués au SNDS afin de constituer des cohortes et sous-groupes spécifiques de malades en France. Les pathologies et sous pathologies sont ici : hépatite virale chronique B et C, cirrhose pour les patients VHB et VHC chroniques repérés, variant hépatique Delta et porteurs sains du VHB. Ces algorithmes doivent permettre d’estimer la prévalence de ces pathologies ou sous-pathologies au niveau national et régional, les trajectoires de soins associées, les consommations de soins et de biens médicaux résultantes.  De précédents travaux appliqués à des bases médico-administratives afin d’identifier les hépatites virales chroniques ont déjà été réalisés. La cartographie des pathologies de la CNAM propose un certain nombre de règles de classification de maladie pour les hépatites virales chroniques. Plus récemment, la validité de certains de ces algorithmes a pu être testée sur le SNDS, en les comparant au statut virologique issu d’une cohorte HEPATHER (Lam, 2023). Les algorithmes les plus simples intègrent essentiellement la nomenclature CIM-10 codée lors des séjours. Ces algorithmes présentent une excellente spécificité mais des sensibilités modérées. La sensibilité des algorithmes était généralement améliorée après prise en compte de tests biologiques relatifs aux malades avec une hépatite virale, la délivrance de médicaments spécifiques à l’hépatite B ou C, la prise en compte des notifications d’ALD. |
| --- |

## 

## Titre du projet :

*Donnez le nom du projet en une phrase pour expliquer son objectif principal*

| Développement d’algorithmes de ciblage de patients atteints d'hépatite virale B et C chronique. |
| --- |

## Auteur(s) :

## 

## *Précisez*

| Nom, Structure (adresse mail) |
| --- |

## Méthodologie :

*Élaborez sur la méthodologie utilisée pour développer l'algorithme, pourquoi vous avez utilisé certaines approches, le raisonnement derrière vos choix, les défis auxquels vous avez été confrontés, et les fonctionnalités futures prévues. Indiquez les références d’articles scientifiques si l’algorithme a fait l’objet d’une publication*

| Les objectifs et enjeux du projet HepavirAlgo sont multiples. Tout d'abord, il consiste en l’utilisation des méthodes de machine learning afin d'améliorer les performances prédictives des travaux cités précédemment. L'intérêt de ces méthodes réside dans (i) la détection agnostique, automatique et sans expertise métier des variables d'intérêt (feature selection) et (ii) la combinaison de ces variables de manière non linéaire, traduisant des relations complexes entre variables, et ce afin d’améliorer les performances prédictives. Ensuite, les codes pour requêter en R les bases Oracle, pré-traiter le SNDS, le transformer en une base unique, puis y appliquer un algorithme de machine learning, sont disponibles librement en open source pour la réutilisation à d'autres projets. |
| --- |

## Données utilisées :

*Précisez les données ou millésimes de données utilisées pour le développement de l’algorithme. Indiquz les références d’articles scientifiques si l’algorithme a fait l’objet d’une publication ou vers la source de la base.*

| En amont de la mise en open source des algorithmes HepavirAlgo, un travail d'investigation et d'évaluation des méthodes a eu lieu en population restreinte (de manière similaire à (Lam, 2023)). En effet, avant de pouvoir appliquer ces algorithmes "en conditions réelles" au SNDS entier sur le portail de la CNAM, ils ont d'abord été établis et affiner en utilisant la cohorte HEPATHER (gold standard des véritables statuts de pathologies) appariée au SNDS. La documentation précise de l'entraînement du modèle et de ses performances est également disponible à la racine de ce repo. Cette documentation donne notamment plus de détails sur la feature selection via des régressions Lasso, le data management du SNDS, le choix de l'algorithme-package XGBoost et les performances par phénotype. |
| --- |

## Validation :

*Précisez sur quelles données l’algorithme a été développé, comment sa performance a été validée, les limites et les biais identifiés, si des sources de données externes ont été utilisées et indiquez la période associées aux données de validation. Indiquez les références d’articles scientifiques si ces étapes de validation ont fait l’objet d’une publication*

| Ces tableaux de performances sont disponibles dans la documentation détaillée expliquant l'entraînement du modèle : |
| --- |

## Date de dernière mise à jour :

*Précisez*

| Il s’agit de la version 1.0 pour l'implémentation R et Python : octobre 2023 |
| --- |

## Maintenance :

*Précisez si l’algorithme est maintenu*

| Maintenu par XX et l’équipe data HDH (adresse.s mail.s) |
| --- |

## Comment installer l’algorithme :

*Fournissez des étapes détaillées pour l'installation du projet et les dépendances nécessaires le cas échéant*

| Une fois établie la méthodologie et les modèles de machine learning entraînés sur la cohorte HEPATHER, les coefficients et poids de ces modèles ont été exportés pour utilisation en population générale. Pour lancer les algorithmes sur le portail de la CNAM il suffit de :  Lancer sa session Rstudio normalement depuis la plateforme CNAM en se plaçant dans un dossier/projet R  Créer un fichier R à la racine du dossier/projet R - peu importe le nom donné à ce fichier - et y copier l'intégralité du contenu installation\_hepaviralgo.R puis de le lancer. Ce programme va créer toute l'arborescence et le contenu des fichiers présents dans le dossier algorithme.  Lancer le programme principal hepaviralgo.R en choisissant le phénotype d'intérêt, les années d'intérêt, et en ajustant éventuellement le seuil de détection selon l'objectif souhaité (afin de calibrer précision, sensibilité et spécificité désirées) Ce programme crée alors deux résultats : l'ensemble des scores de probabilité de pathologie pour l'ensemble des individus étudiés, et surtout la création des BEN\_NIR\_ANO correspondant à la cohorte finale d'intérêt |
| --- |

## Comment utiliser l’algorithme :

*Donnez des instructions et des exemples pour permettre aux utilisateurs/contributeurs d'utiliser l’algorithme, y compris des captures d'écran si nécessaire. Si des prérequis sont requis, y compris en termes de pré-traitement des données, précisez les*

| Pour les personnes intéressées par le fonctionnement du projet, le script hepaviralgo.R consiste lui-même en l'appel à des fonctions issus de deux programmes :  func\_oracle.R comprend toutes les fonctions permettant d'importer les tables Oracle du SNDS, de les joindre puis de les formater en une table unique utilisable par les algorithmes de machine learning. Plus précisément l'objectif est de créer une table unique par année, donnant le décompte de chaque événement durant l'année.  fun\_predictions.R applique ensuite simplement les poids des modèles XGBoost estimés (via les données HEPATHER-SNDS) aux bases annuelles précédemment créées. En raison de non disponibilité du package à la bonne version, le modèle implémenté est uniquement une régression linéaire à ce jour. |
| --- |

## Support :

*Indiquez où les utilisateurs peuvent obtenir de l'aide, que ce soit via un suivi des problèmes, un chat, une adresse e-mail, etc*

| Point de contact : adresse mail |
| --- |

## Contributions :

*Dites si vous acceptez les contributions et spécifiez vos exigences à cet égard. Fournissez de la documentation pour aider les contributeurs à démarrer*

| N/A |
| --- |

## Crédits :

*Si vous avez collaboré avec d'autres personnes, listez-les avec des liens vers leurs profils GitHub et médias sociaux. Incluez également des liens vers les tutoriels ou les ressources que vous avez utilisés*

| N/A |
| --- |

## Licence et conditions d’utilisation :

*Incluez une licence pour préciser ce que les autres peuvent ou ne peuvent pas faire avec votre projet. Préciser les conditions de crédits*

| N/A |
| --- |

## Autre :

*lien vers de la documentation ou autre*

| N/A |
| --- |

Tags/filtres à inclure dans la fonction de recherche :

| **Catégories** | **Tags possibles (non-exhaustif)** |
| --- | --- |
| **Objectifs de l’algorithme** | * Outils de ciblage dans la base principale du SNDS * Outils de manipulation / transformation de la base principale du SNDS * Outils de pseudonymisation / anonymisation * Outils d’identification de patients / dépistage biologique * Outils de cartographie des pathologies * Algorithmes de requête à la demande * Autre |
| **Domaines médicaux** | * Cancers * Maladies cardio-vasculaires * Diabète * Maladies neurodégénératives * Santé mentale et psychiatrie * Périnatalité et santé reproductive * Maladies respiratoires * Maladies infectieuses * Enfants et pédiatrie * Personnes âgées et gériatrie * Autre |
| **Données d’application** | * Base principale * Données d’imagerie * Données tabulaires * Texte libre * Autre |
| **Langage de programmation** | * Python * SAS * R * SQL * Autre |
| **Auteurs** | * Établissement de santé * Industriel santé * Établissement d’enseignement supérieur * CLCC * Associations et ONG * Organisme de recherche * Bureau d’étude * Agence * Administrations et ministère * Autorités administratives indépendantes et consultatives * Collectivités territoriale * Entreprise * Start-up * Fondation * Société savante * Institution européenne ou internationale * Assurance * Pôle de compétitivité * Plateforme de données * Organisme de presse * Association de patients * Établissement / Fédération de santé (CHU, CHRU, CLCC, clinique, etc…) * Autre |
| **Statut de validation scientifique de l’algorithme (sur des données de validation)** | * Validé * En cours de validation * Non validé |
| **Maintenance** | * Régulière * Ad-hoc (en fonction des remontées de problèmes, suggestions) * Pas de maintenance |