Prédiction du taux de remboursement des médicaments en France

Projet de datamining

Chloé Gobé
Venceslas Danguy des Déserts
Eymard Houdeville

Table des matières

Introduction

Ce projet a été réalisé dans le cadre du cours IS3024, *Des données à la connaissance*. Nous avions pour objectif de prédire le taux de remboursement d'un médicament en fonction d'autres données telles que le prix, les avis des différentes commissions, le laboratoire de production, etc.

1 Ensemble de données

Nous nous sommes basés sur la Base de Données Publique et officielle des Médicaments [1] (BDPM).

Voici le schéma complet de cette base de données :

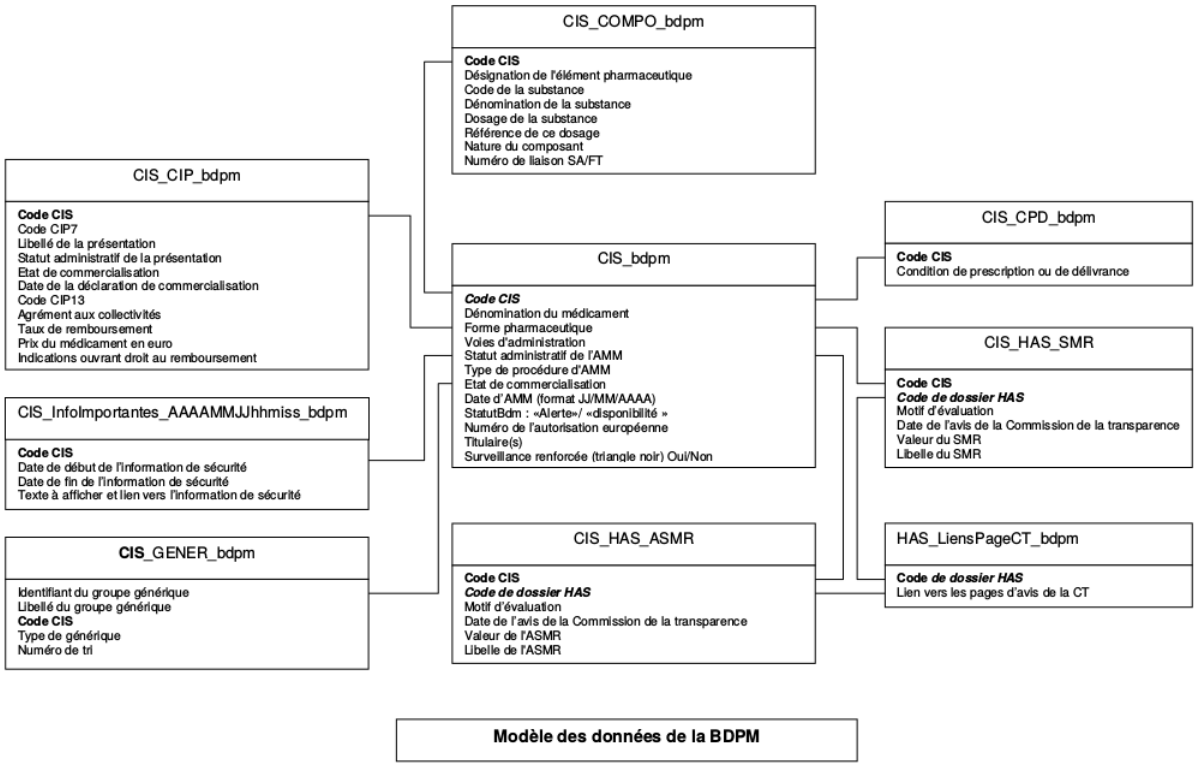

FIGURE 1 – Schéma de la base de données

Il est possible de récupérer un dump récent de cette base de données sur le site *www.data.gouv.fr*.

Le nombre de médicaments présents dans cette base de données est d'environ 14000 références dont les dates d'autorisation de mise sur le marché varie entre 1973 et 2018. Certains médicaments sont commercialisés, d'autres ont été arrêtés.

Les tables les plus importantes sont les suivantes :

CIS_bdpm : Contient une ligne par médicament avec son nom, son identifiant unique (code CIS), sa forme pharmaceutique, ses voies d'administration, ainsi que des informations sur les différentes procédures et status.

CIS_CIP_bdpm : Cette table contient une ligne par présentation, une présentation étant en quelque sorte une instance d'un médicament. Par exemple le médicament *Doliprane en comprimés* peut être vendu par boîte de 8 ou 16 comprimés, et à une concentration de 500 mg ou 1000 mg. Ce sont ces présentations qui ont un prix et un taux de remboursement fixés par les différentes commissions.

CIS_HAS_SMR et CIS_HAS_ASMR : Contiennent les avis des commissions sur le *Service Médical Rendu* (SMR, l'utilité du médicament pour une pathologie donnée), ainsi que sur l'*Amélioration du Service Médical Rendu* (ASMR, utilité du médicament pour une pathologie donnée par rapport aux traitements préexistants). La commission rend un avis pour chaque pathologie pour laquelle le médicament peut avoir un intérêt, ainsi un médicament peut-il avoir un SMR bon pour une pathologie et mauvais pour une autre.

La base de données contient aussi des informations sur l'existence ou non de génériques, ainsi que sur la présence d'informations "de sécurité", mais nous avons choisi de ne pas les inclure pour garder un modèle simple. Il doit toutefois être possible d'améliorer nos résultats en les utilisant (la présence d'un générique doit par exemple fortement influencer le taux de remboursement du médicament original).

2 Nettoyage des données

Bien que cette base de donnée soit officielle, la qualité des données laisse à désirer. Nous avons donc du remanier une bonne partie des champs pour proprifier les données.

2.1 Médicaments - CIS_bdpm

Formes galléniques

Les formes galléniques sont très nombreuses : le dataset en compte au départ 401 différentes. En regardant le champ de près, on s'aperçoit que sont distingués par exemple les *comprimés*, les *comprimés sécables* ou les *comprimés enrobés*. Il y a aussi des erreurs d'écriture, ainsi l'un des médicaments a-t-il pour forme gallénique *capsule molle ou* (littéralement). Nous avons estimé que nous n'avions pas besoin d'un tel niveau de détail et avons donc fusionné les différentes formes de façon à n'en garder que 67. L'idée est de générale est de mettre ensemble tous les comprimés, toutes les pommades, etc, et de regarder "à la main" les formes restantes pour vérifier qu'il ne reste pas d'aberration ou de formes pouvant être fusionnée.

Utilisation de types "catégories"

Pandas possède un type *category* utilisé pour représenter un champ correspondant à une unique catégorie parmis plusieurs. L'utilisation de ce type de données ne change pas le contenu de la cellule en soi mais facilite le traitement ultérieur pour calculer des statistiques ou transformer l'entrée de façon à pouvoir la donner à un modèle d'apprentissage. Nous avons donc taggés comme étant des catégories les champs suivants :

— galenic_form
— route_of_administration
— owners
— commercialisation_status

— clearance_status
— clearance_type
— bdm_status
— enhanced_monitoring

Voies d'administration

Les différentes voies d'administration utilisables pour un médicament donné sont présentées dans le dataset comme une liste d'éléments séparés par des point-virgules.

Exemple : *infiltration ;intra-articulaire ;périarticulaire ;péridurale ;périneurale*

Cette liste n'est pas utilisable par pandas en l'état.

Nous avons contacté un étudiant en médecine que nous connaissons et qui nous a aidé à catégoriser plus simplement ces voies très nombreuses en trois grandes catégories susceptibles d'être intéressantes pour notre prédiction : - Injectables (os, veines, muscles, articulations...) - Per os (globalement par la bouche, le tube digestif ou autre formes associées dans la pratique de la médecine) - Autre : les inclassables ou les formes qui ne rentrent pas dans les deux premières grandes catégories.

Nous avons ensuite défini un booléen pour chacune de ces grandes catégories (A,P,I). Ce booléen est bien plus facilement exploitable dans la suite de nos opérations qu'une liste longue et complexe.

Dates

Il faut enfin parser tous les champs contenant des dates de façon à ce que pandas les reconnaisse comme tel et non plus comme des chaînes de charactères. Cela permet d'établir une relation d'ordre (temporelle) entre différentes valeurs.

2.2 Présentations - CIS_CIP_bdpm

En plus des traitements sur les dates et les types catégories que nous appliquons aussi, il y a quelques champs spécifiques à traiter.

Prix

Nous avons décidé de supprimer tous les médicaments n'ayant pas de prix de vente. Après vérification nous sommes arrivés à la conclusion qu'une absence de prix (qui va de pair avec une absence de taux de remboursement) dans la base de donnée signifie simplement que le médicament n'est pas remboursé par la Sécurité Sociale. Nous avons choisi de ne pas inclure ces médicaments dans notre ensemble de données, mais il aurait pu être intéressant de tenter de prédire aussi un taux de 0%.

Taux de remboursements

Nous nous sommes aperçus que s'il existe peu de taux différents (15%, 30%, 65% et 100%), ces taux pouvaient être écrits dans deux formats différents dans la base : avec et sans espace entre le dernier chiffre et le %. Ainsi trouve-t-on des médicaments remboursés à *15%* et d'autres à *15 %*. Nous avons donc enlevé les espaces partout...

2.3 Les tables présentant les résultats SMR et ASMR

Les avis SMR et ASMR sont fixés par des commissions et rendent compte du service rendu ou de son amélioration [2]. Ils sont normalement déterminants dans le choix du taux de remboursement [3]. Les laboratoires déposent une requête d'évaluation pour un médicament donné et pour un ensemble d'indications thérapeuthiques. Les commissions rendent ensuite un avis pour chacune de ces indications. La base de données ne contient malheureusement que les avis rendus après 2002... Comme il est compliqué d'analyser les commentaires textuels de la commission, nous avons choisi de ne garder que les notes qui vont de un à cinq. Un médicament donné à donc plusieurs notes : une par indication thérapeutique. Afin de simplifier le modèle, nous avons regroupés ces notes en fonction de leurs valeurs et les avons comptées. Ainsi, dans notre dataset final, les notes d'un médicament se présentent comme suit :

SMR - insuffisant : un avis

SMR - faible : zéro avis

SMR - modéré : zéro avis

SMR - important : deux avis

SMR - majeur : zéro avis

Ce médicament rend donc un service médical insuffisant pour l'une des indications évaluées, mais important pour les deux autres.

3 Modèles et entraînements

3.1 Choix des données

Les taux de remboursement étant très peu uniformes, nous avons décidé de simplifier les valeurs en faisant la prédiction d'un taux de remboursement "faible" et d'un taux de remboursement "élevé".

Taux de remboursement	Nombre de médicaments
15%	564
30%	975
65%	10779
100%	678

FIGURE 2 – Répartition du taux de remboursement dans le dataset

Comme on peut le constater, la grande prégnance des taux de remboursement à 65% rend l'exercice de prédiction peut intéressant. Nous avons au contraire choisi de nous concentrer sur les médicaments au taux de remboursement extrême : pourquoi tel médicament est-il entièrement remboursé et tel autre très faiblement ?

Nous choisissons d'intégrer les colonnes contenues dans le vecteur "features" du code et générons plusieurs fois l'arbre en prenant en compte ou non certaines features pour mieux comprendre leur importance relative.

3.2 Choix du modèle

Nous choisissons de travailler avec les Decisions Tree de ScikitLearn. La grande facilité d'interprétation des arbres de décision, leur caractère heuristique simple nous semble un avantage dans un projet que nous pourrions être amenés à présenter à des médecins. Toute décision de classification de notre arbre sera ainsi facilement expliquée par une suite de choix booléens.

Nous disposons en outre à la fois de données catégoriques et de données numériques.

Nous savons que ce choix nous expose néanmoins à un certain nombre de dangers : - Une petite variation dans les données peut avoir des conséquences plus ou moins importantes dans la structure du résultat. Il va donc falloir nous assurer de la stabilité de notre modèle. - Notre arbre peut vite s'avérer très touffu et complexe. Il va falloir jouer avec les paramètres de Scikit pour garder une certaine lisibilité

3.3 Méthode

Nous utilisons un framework très classique en travaillant avec un test set et un training set.

3.4 Paramétrage de l'arbre

Le crière que nous utilisons est le critère de Gini : c'est un indice qui permet de rendre compte de la disparité des individus au sein d'une même classe.

La profondeur maximale que nous permettons à notre arbre d'avoir est de 6 : il nous semble que l'arbre devient ensuite difficilement lisible.

Nous fixons les paramètres min_samples_leaf et min_samples_split à 10 afin de nous assurer que notre arbre reste assez général.

Enfin, nous utilisons le module graphviz de python afin de disposer d'une présentation graphique à montrer à des médecins.

3.5 Résultats

En utilisant un test set contenant 25% des médicaments et en entrainant notre arbre sur les 3/4 restants notre arbre classifie dans 89% des cas les médicaments dans la bonne catégorie de remboursement.

Si l'on essaie de générer à nouveau l'arbre en ne tenant pas compte du prix on observe que l'accuracy tombe à 72%.

La matrice de confusion que nous obtenons alors :

$$\begin{bmatrix} [129 & 5 & 38] \\ [30 & 65 & 40] \\ [34 & 10 & 204] \end{bmatrix}$$

Une tentative de prédiction sans les avis SMR et ASMR qui semblent si importants fait tomber l'accuracy à 63%

3.6 Observations et interprétations

L'arbre est en annexe de ce rapport au format pdf. Les couleurs des feuilles indiquent les catégories dans lesquelles sont rangées les médicaments respectant ces critères. L'intensité de la couleur représente l'homogénéité de la classe.

Nous observons que :

— Le SMR est le critère le plus discriminant puisqu'il intervient le plus haut dans l'arbre. Autrement dit, et heureusement, il semblerait que ce soit l'appréciation du service rendu au patient qui prime dans le choix du taux de remboursement d'un médicament.

— Le prix joue le second rôle le plus discriminant dans notre arbre

— On voit occasionnellement apparaitre d'autres critères comme le clearance type ou la forme galénique. La voie d'administration ne semble pas jouer dans le remboursement du médicament.

Si nous ré-entrainons le modèle sans prendre en compte le prix, nous pouvons confirmer que le clearance-type (procédure nationale, décentralisée ou reconnaissance mutuelle) possède une importance discriminative importance.

Si nous ré-entrainons le modèle sans les avis SMR et ASMR nous observons que les voies d'administration prennent une importance insoupçonnée jusqu'ici et remontent systématiquement dans les premières branches de l'arbre, au dessus de la forme galénique (ce qui semble logique : la voie d'administration implique une certaine forme galénique : par exemple une injection ne peut prendre qu'un certain nombre de formes galéniques).

Conclusion

Les données fournies dans la base de données publiques de médicaments sont réparties en plusieurs tables, parfois redondantes ou contradictoires, où les données sont elles mêmes parfois incohérentes ou tout simplement manquantes. Ce projet nous a permis de nous confronter à une base de données réelle en expérimentant un algorithme vu en cours de Datamining pour prédire une des catégorie fournie :le taux de remboursement des médicaments.

Nos résultats confirment les intuitions des médecins et le fonctionnement institutionnel du système de santé français : ce sont les avis SMR et ASMR qui comptent le plus lors de la fixation du taux de remboursement du prix d'un médicament.

Il serait intéressant pour un autre projet de considérer ces taux de remboursement comme des séries temporelles et de mettre ces dernières en corrélation avec d'autres indicateurs macro-économiques : y a t-il globalement une baisse ou une hausse des médicaments fortement remboursés ? Cette baisse touche t-elle certaines catégories de médicaments en particulier ?

Références

[1] *Base de données publique des médicaments (base officielle) - Data.gouv.fr*. URL : `/fr/datasets/base-de-donnees-publique-des-medicaments-base-officielle/` (visité le 09/05/2018).

[2] *Haute Autorité de Santé - Le service médical rendu (SMR) et l'amélioration du service médical rendu (ASMR).* URL : https://www.has-sante.fr/portail/jcms/r_1506267/fr/le-service-medical-rendu-smr-et-l-amelioration-du-service-medical-rendu-asmr (visité le 09/05/2018).

[3] *La fixation des prix et du taux de remboursement.* Ministère des Solidarités et de la Santé. 13 juin 2016. URL : http://solidarites-sante.gouv.fr/soins-et-maladies/medicaments/le-circuit-du-medicament/article/la-fixation-des-prix-et-du-taux-de-remboursement (visité le 09/05/2018).