Joey Lévesque
111-219-341

Sam Chandavong
111-026-119

Corentin Labelle
111-132-133

Analyse et traitement de données massives

GLO-7027

Projet de session

Travail présenté à

Richard Khoury

Département d'informatique et de génie logiciel

Université Laval

Hiver 2023

Sam Chandavong, Corentin Labelle, Joey Lévesque

Abstract

Le jeu de données fourni pour le rapport comporte les informations de 3 millions de patients distincts entre le 1er janvier 2000 et le 31 décembre 2020. Dans le jeu de données, il est possible de retrouver des cas de polypharmacies erronées et réels. Ces cas réels seront les valeurs que l'algorithme sur lequel nous travaillerons tout au long de la session essaiera de détecter. En plus des cas de polypharmacies, des hospitalisations sans polypharmacie se retrouveront aussi dans le jeu de données. Certaines de ces données ne comportant pas de polypharmacie ni d'hospitalisation sont plus difficilement classifiable. Certaines fois ils agiront comme du bruit dans notre algorithme et d'autres fois comme valeurs normales.

1 Analyse des données

1.1 Valeurs normales et cas problématiques

Notre jeu de données comporte 22 colonnes : 19 colonnes pour les médicaments, une colonne pour la date de la prescription, 1 colonne pour l'identifiant du patient, une autre colonne pour l'hospitalisation. Ces données sont répétées un total de 30220351 fois. Dans notre présent rapport, la valeur normale du taux d'hospitalisation sera l'occurrence moyenne qu'une personne ne prenant aucun médicament soit hospitalisé.

1.2 Présence de bruit

Comme mentionné plus haut, il y a plusieurs causes de bruit dans les données fournies pour le projet. La liste suivante n'est bien sûr pas une liste exhaustive. Il est fort probable que tout au long du projet, nous trouvions d'autres causes de bruit dans notre jeu de données.

1. Il est possible de trouver des cas d'hospitalisation qui ne sont pas liés à la prise de médicaments. Pensons par exemple à une personne qui a subi un accident de voiture.
2. Certains patients ont peut-être une santé de fer qui malgré une polypharmacie n'ira pas en hospitalisation.
3. Au contraire du point précédent, certaines personnes pourraient être hospitalisé avec la prescription de plusieurs médicaments pour des raisons externes à celle-ci (hospitalisation liée à la covid par exemple).
4. Certaines périodes de l'année pourraient être plus propices à des hospitalisations sans lien avec la polypharmacie (la période de la grippe au début de l'hiver par exemple).

1.3 Le fléau de dimensionnalité

Notre jeu de données comprend 22 attributs (19 médicaments, hospitalisation, 'id' du patient et date). Les combinaisons sont donc multiples et il y aura forcément du bruit. Il faut donc bien choisir nos algorithmes lors de l'analyse. Une possibilité pourrait être de diminuer le nombre d'attributs (combiner des médicaments fortement corrélées, décomposition en composantes principales), ce qui diminuerait le nombre de combinaisons possibles. À titre d'exemple un jeu de données contenant uniquement le temps et un médicament serait très facile à valider, la présence du fléau de dimensionnalité arrive puisque le nombre de combinaisons disponibles augmente exponentiellement dans le jeu de données pour chaque médicament supplémentaire. Chaque ajout supplémentaire de médicaments dans le jeu de données diminue la certitude de trouver des cas de polypharmacies.

1.4 Les informations manquantes

Certaines informations sont manquantes dans le jeu de données afin de nous permettre de réussir à tirer une conclusion hors de tout doute de la présence d'une polypharmacie.

1. L'âge des personnes n'est pas contenu dans le jeu de données. Il serait logique de penser que la fragilité des personnes âgées serait plus à risque à une hospitalisation.
2. N'ayant pas accès aux antécédents médicaux des patients, il n'est donc pas possible de corréler les possibilités d'aller en hospitalisation avec le nombre de problèmes de santé.
3. Les effets secondaires des médicaments sont aussi absents de notre jeu de données. Certains d'entre eux pourraient avoir un effet (par exemple la perte d'équilibre) qui mènerait à une hospitalisation plus fréquente. Il serait toutefois possible de déterminer cette information sans jamais en être sûr dans notre jeu de données.
4. Considérant qu'il existe plus de 19 médicaments en dehors du jeu de données, il est possible qu'un cas de polypharmacie ne soit pas recensé avec les médicaments de notre jeu de données. Les patients pourraient prendre des médicaments qui ne sont pas listé dans la base de données.
5. En lien avec le point précédent, on n'a pas la posologie des médicaments qui peuvent avoir une influence sur la polypharmacie. Aussi, certains médicaments sont considérés au besoin, et non d'une fréquence régulière.

1.5 Le déséquilibre des classes

Dans notre projet le déséquilibre des classes peut être expliqué par certaines raisons dont entre autres le fait que certaines personnes soit plus propices aux hospitalisations et que certains médicaments ou groupes de médicaments soit plus souvent prescrits que d'autres

1.6 Les valeurs aberrantes

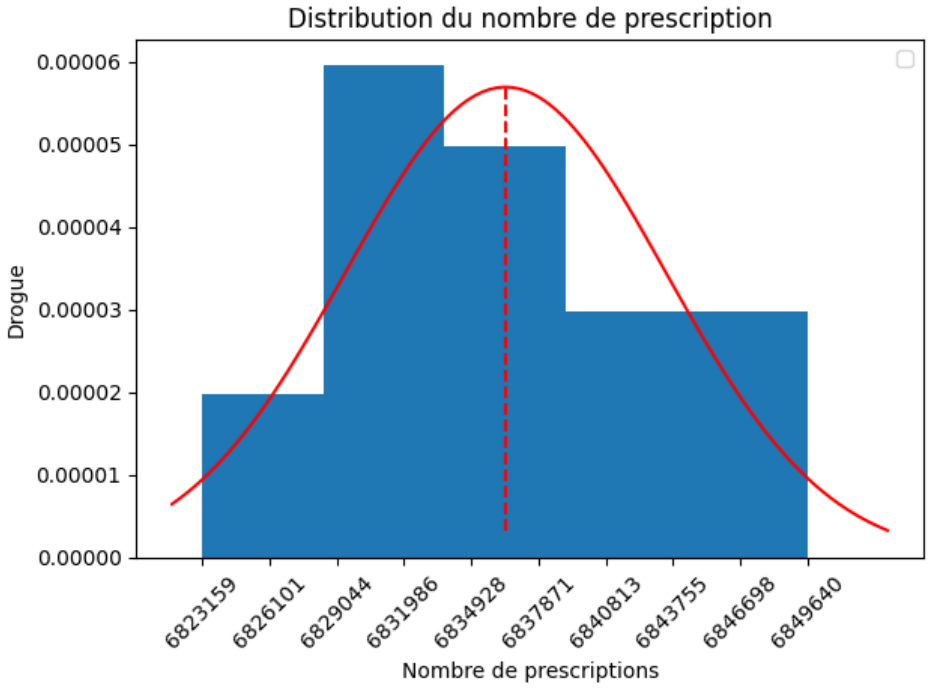
Dans de rares cas, un patient ne prend aucun médicament présent dans la base de données. Plutôt que de les retirer, ces données pourraient être utilisées pour, par exemple, utiliser un nombre d'hospitalisations de base. Si on remarque que les patients ne prenant aucun médicament ont, en moyenne, deux hospitalisations, on peut assigner un nombre d'hospitalisations de base de deux. Ainsi, tout patient ayant un nombre d'hospitalisations supérieur à deux pourra être considéré comme significatif et devra être analysé plus en profondeur.

Aussi, il est possible de penser à des personnes ayant une (ou des) précondition(s) médicale(s) les amenant plus régulièrement à l'hospitalisation et qui n'a pas de lien avec les polypharmacies. De la même façon, une personne avec une santé de fer n'allant jamais à l'hôpital malgré une condition qui sur une personne différente l'aurait amené à une hospitalisation.

2 Utilisation de nos données

2.1 Distribution des médicaments

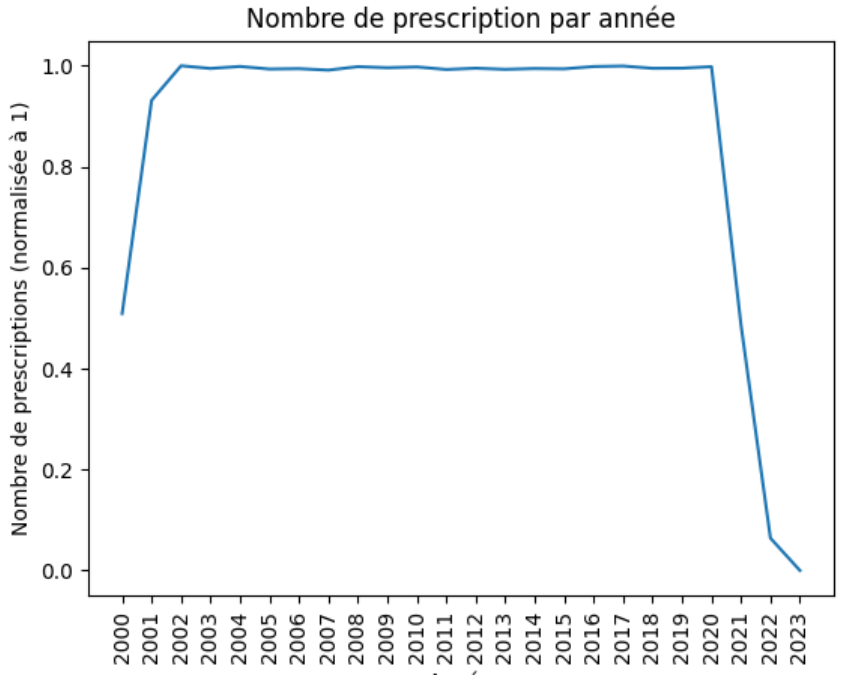
Le premier élément important à analyser est la distribution du nombre de prescriptions de chaque médicament. La moyenne de la distribution est de 6 836 436.58 et l'écart-type est de 7007.88 (environ 0.1% de la moyenne). Le médicament la moins souvent prescrite est le médicament 16 (6 823 159 fois), ce qui est aussi environ 0.19% inférieur à la moyenne. Le médicament la plus souvent prescrite est le médicament 8 (6 849 640 fois), ce qui est aussi environ 0.19% supérieur à la moyenne. On en comprend que la distribution des médicaments est très concentrée autour de la moyenne.



À la vue du graphique, il semble qu'il y ait un peu plus de médicament dans les sceaux de fréquences élevées que dans les sceaux de fréquences faibles. La distribution des médicaments serait donc légèrement asymétrique négative.

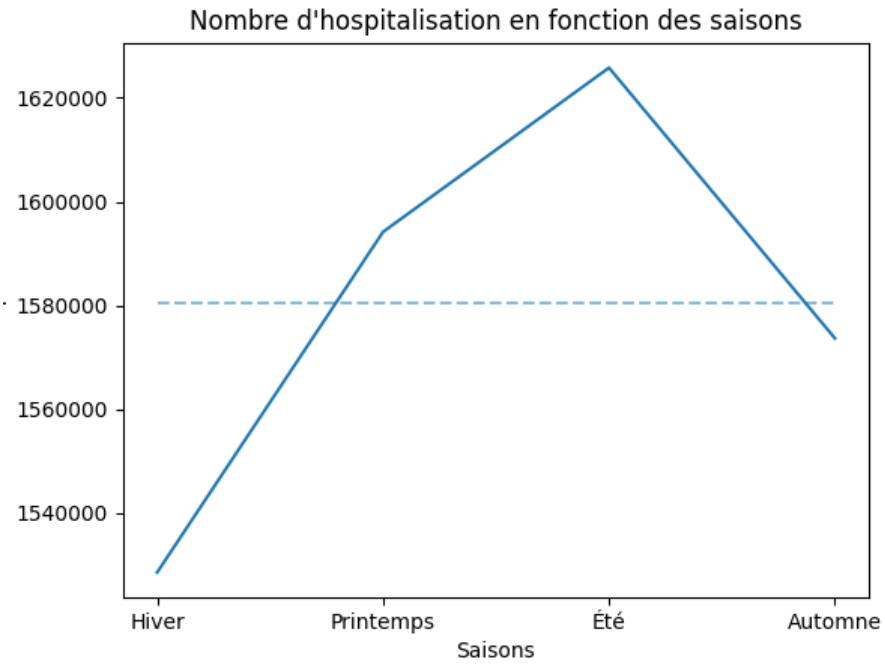
2.2 Distribution temporelle

En moyenne, 1 259 181.29 prescriptions sont prescrites chaque année. Mis à part la première année et les trois dernières années, le nombre de prescriptions est uniformément distribué dans le temps.



2.3 Influence des saisons sur les hospitalisations

Au total, il y a 6 322 242 hospitalisations, donc en moyenne 1 580 560.5 hospitalisations par saison. La différence entre le plus grand nombre d'hospitalisations par saison (en été - 1 625 800 hospitalisations) et le plus petit nombre d'hospitalisations par saison (hiver - 1 528 574 hospitalisations) est de 97 226 hospitalisations. L'écart entre les saisons est non significatif et on peut conclure que les saisons n'influencent pas les hospitalisations.



3 Traitement de nos données

Dû à la nature discrète de nos variables, la régression logistique pourrait être intéressante. L'utilisation de la librairie scikit-learn est aussi primordiale dans le projet.

Plusieurs algorithmes d'apprentissage automatique pourront être implémentés dans notre projet.

Sachant que nous avons l'hypothèse que certains médicaments seront groupés avec quelques patients, l'algorithme de K [1] plus proche voisin (k-NN) serait un modèle intéressant à implémenter, car elle a l'avantage d'utiliser la proximité afin d'effectuer des classifications ou des prédictions sur le regroupement classifications ou des prédictions sur le regroupement des points de nos données. Aussi, la distance euclidienne sera probablement la plus appropriée, car nous avons l'hypothèse que la distance de nos données sera très proche les unes aux autres, ce qui est un avantage avec la distance euclidienne, permettant de mieux comprendre les petites différences comparativement aux autres métriques. Il se peut toutefois que ce modèle ne soit pas aussi optimal que d'autres modèles, puisque le k-NN est sensible au fléau de la dimensionnalité.

Un autre modèle que nous pourrions utiliser pour ce projet serait l'algorithme de Naive Bayes avec Bernoulli [2], étant donné que les valeurs de nos médicaments sont booléennes. Cet algorithme utilise le théorème de Bayes pour calculer la probabilité de chaque combinaison possible de nos médicaments pour finalement calculer la probabilité de notre sortie, ce qui est dans notre cas, une hospitalisation ou non. De plus, considérant que nos données sont volumineuses, cet algorithme est considéré l'un des moins complexes à exécuter.

Sachant qu'il est important d'avoir un modèle qui prédit bien l'hospitalisation par la cause de certaines combinaisons de médicaments, l'arbre de décision [3] serait un modèle très intéressant à implémenter dans ce projet. Elle est simple à implémenter et facile à comprendre, cependant, il faudra espérer que nos données n'aient pas de petites variations, ce qui rendra le modèle moins précis. C'est un modèle plus coûteux que les autres algorithmes, donc il faudra surveiller si ce modèle est applicable avec nos données massives. Il sera possible de faire de l'étalage dans notre arbre de décision pour améliorer la précision si notre modèle permet de s'exécuter.

4 Procédure de tests

Pour la procédure de tests, nous comptons diviser les données en 3 groupes. À la vue du projet, le premier groupe qui contiendra 60% des données servira à entrainer nos algorithmes, le deuxième groupe qui contiendra quant à lui 20% des données servira à valider l'algorithme et le dernier 20% sera utilisé pour les tests finaux. Tout au long du projet nous essayerons différents algorithmes comme mentionnés plus haut et ajusterons les pourcentages de répartition des groupes afin de trouver ce qui permettra d'atteindre le résultat le plus prometteur.

Notre procédure de tests nous permettra de nous assurer de limiter les faux négatifs puisqu'il est couteux d'avoir des résultats erronés à la fin de notre étude.

Reference

[1] Qu’est-ce que l’algorithme des k plus proches voisins ? | IBM. (n.d.). Retrieved February 13, 2023, from https://www.ibm.com/ca-fr/topics/knn

[2] What is Naïve Bayes | IBM. (n.d.). Retrieved February 13, 2023, from https://www.ibm.com/topics/naive-bayes

[3] Qu’est-ce qu’un arbre de décisions | IBM. (n.d.-b). Retrieved February 13, 2023, from https://www.ibm.com/fr-fr/topics/decision-trees