Rapport 1

Ekoué Mawuéna Octave Adama[†], Auguste Aka Tiemele[†], Deoth Guei [†]
Département d'informatique et de génie logiciel, Université Laval

Abstract

Ce rapport se concentre sur l'analyse et le prétraitement de données de polypharmacies dans le but de déterminer les combinaisons de médicaments qui peuvent conduire à une hospitalisation. Nous avons examiné les données pour vérifier s'il n'y a pas de valeurs manquantes ou aberrantes. Nous avons par la suite effectué une analyse approfondie pour tirer des informations intelligibles des données, ce qui nous a conduit à réduire considérablement les bruits dans les données. Ensuite, nous avons proposé des algorithmes d'apprentissage pour résoudre ce problème de polypharmacies nocive, en se concentrant sur les méthodes les plus appropriées pour cette tâche spécifique. Ce rapport sert de base pour les travaux futurs sur la détection de polypharmacie nocives.

Keywords: Traitement de données, polypharmacie

1. Introduction

Selon un article réalisé par Caroline Sirois[1], «un grand nombre de personnes, surtout les aînés, sont exposées à la polypharmacie et ce nombre s'est accru ces dernières années. En 2011, le nombre moyen d'ordonnances par patient de 65 ans et plus au Québec était de 106, correspondant à environ neuf réclamations par mois.» Dans ce premier rapport, nous allons nous concentrer sur l'analyse et le pré-traitement des données du programme de recherche en pharmacie. Nous disposons de données sur 3 millions de patients distincts entre le 1er janvier 2000 et le 31 décembre 2020. Chaque événement de santé d'un patient est noté dans un tuple de la base de données, qui comprend 22 colonnes décrivant les informations du patient, telles que l'identifiant unique, la date de l'observation, la consommation de médicaments et l'hospitalisation. L'objectif de ce projet est de détecter la polypharmacie, qui décrit la situation où un patient prend cinq ou plus médicaments simultanément pour traiter plusieurs conditions. Étant donné que les combinaisons de médicaments peuvent avoir des conséquences néfastes imprévues, il est important de détecter les combinaisons nocives. Cependant, il est également important d'éviter les faux positifs, étant donné que chaque combinaison détectée doit passer par un processus de validation onéreux. Dans cette section, nous allons analyser les données et leurs propriétés statistiques pour déterminer si elles nécessitent un pré-traitement. Nous prendrons en compte les cas problématiques tels que la présence de bruit, le fléau de dimensionalité, les informations manquantes et le déséquilibre des classes. Les techniques de pré-traitement seront alors appliquées pour garantir la qualité des données utilisées pour la suite du projet.

[1]Texte rédigé par Caroline Sirois, B. Pharm., M.Sc., Ph.D., professeure, Département des sciences infirmières, UQAR, https://www.uqar.ca/uqar/uqar-info/2015/05_mai/qp_201406_lespagesbleues.pdf

Département d'informatique et de génie logiciel, Université Laval *

2. Analyse et traitement de données

2.1. Description des données

Les données sont constituées de 22 attributs dont 20 sont binaires, un de type timestamp et un attribut de type int. L'ensemble des données contient au total 30220351 enregistrements. Il n'y a aucune données manquante et le balancement initial des hospitalisations se présente comme suit: 79 % de non hospitalisations et 21% d'hospitalisations.

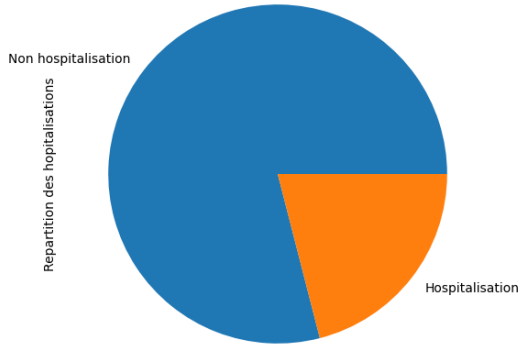

Figure 1. Balancement Initial

Cette proportionnalité de 79% et 21% est exactement la même indépendemment du temps comme on peut le voir dans la figure 2. Dans la figure 2 ci dessous , on a essayé de tracer les graphes du nombre de cas d'hospitalisations et le nombre de cas de non hospitalisations en fonction du temps , en espérant trouver une dépendance entre les attributs «hospit» et «timestamp» qui indiquerait peut-être une autre raison pour l'hospitalisation d'individus (une épidémie par exemple) ou une combinaison de médicaments à une année précise. Après observation, on constate une même variation pour les deux courbes. Le nombre de cas d'hospitalisations croit progressivement jusqu'en 2002 , année à partir de laquelle il y'a une stabilisation de ce nombre de cas d'hospitalisations jusqu'à l'année 2022. A partir de l'année 2022 jusqu'à 2023 on a une décroissance. La courbe de cas de non hospitalisations suit exactement le même schéma avec des valeurs beaucoup plus hautes sauf en 2023. Cette piste, pour une possible détection du bruit lié au signal d'hospitalisation est donc une piste sans succès.

L'histogramme de prescription des médicaments montre la répartition des nombres de prescriptions en utilisant cinq catégories ou seaux différents. Cette distribution est caractérisée par une asymétrie positive, ce qui signifie que la majorité des médicaments ont un nombre relativement faible de prescriptions, tandis que peu de médicaments sont très largement prescrits.

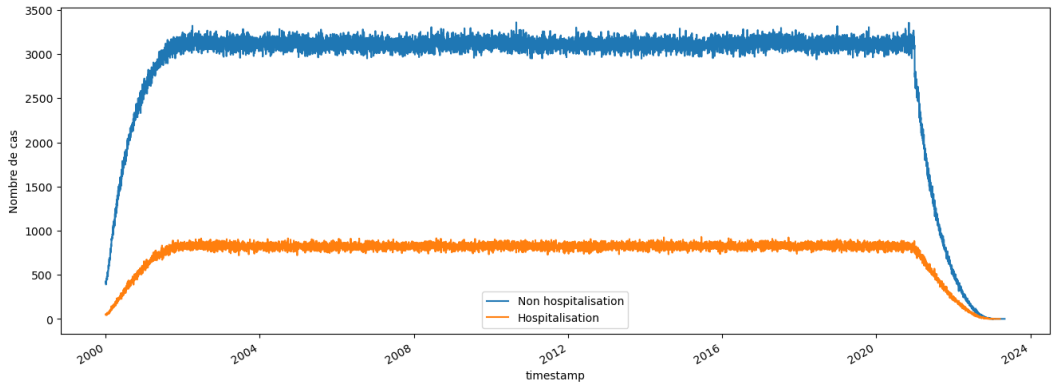

Figure 2. Distribution des hospitalisations en fonction du temps

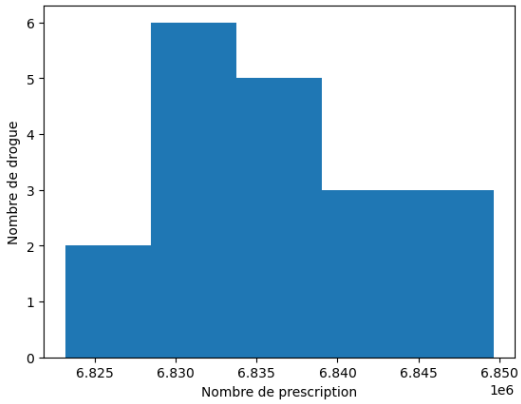

Figure 3. Distribution des prescriptions avec 5 seaux de nombre de prescriptions

En d'autres termes, la plupart des médicaments ne sont que peu utilisés par les médecins, mais il y a une petite proportion de médicaments qui sont largement prescrits. Cette asymétrie positive montre une répartition inégale des prescriptions de médicaments, avec un nombre relativement plus élevé de médicaments peu utilisés et un nombre relativement plus faible de médicaments très utilisés.

Dans la suite de l'analyse , nous avons essayé de trouver des corrélation entre les différents médicaments afin de pouvoir potentiellement réduire le nombre d'attributs. L'intuition derrière cette recherche de pouvoir potentiellement trouver des médicaments qui joueraient des mêmes rôles ou ayant des composantes similaires créant donc des mêmes effets et pouvant être résumés en un seul attribut.Cependant cette corrélation n'est pas apparente car il semble que les apparitions de médicaments soient fortement decorélées .Nous nous sommes servi de la carte de chaleur de la figure 4 pour pouvoir observer ce manque de corrélation.

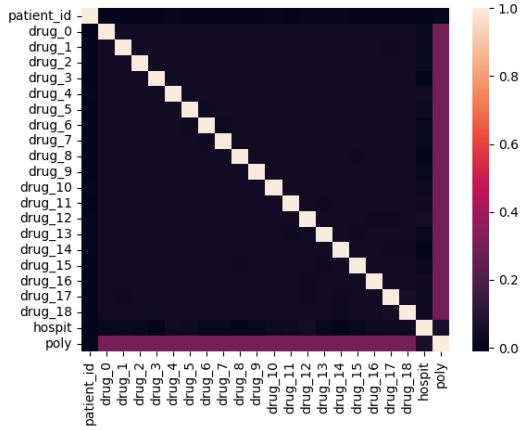

Figure 4. Corrélation entre attributs

2.2. **Problème de scalabilité**

Avec les 30220351 enregistrements que nous avons dans nos données, nous risquons d'avoir lors de l'apprentissage un "problème de scalabilité". Cela signifie que les algorithmes d'apprentissage automatique peuvent devenir lents et inefficaces lorsqu'ils sont confrontés à des ensembles de données très volumineux, ce qui peut rendre difficile la généralisation des tendances dans les données. Cela peut également entraîner un suremballage ou une sur-complexité du modèle, ce qui peut avoir un impact négatif sur la précision des résultats. Nous allons réduire le volume des données en effectuant un nettoyage de ces derniers. Nous allons retirer toutes les informations inutiles qui peuvent entraîner des biais dans l'apprentissage de sorte de ne garder que les informations nécessaires.

2.3. **Élimination des bruits liées a la polypharmacie**

La polypharmacie décrit la situation de patients prenant cinq ou plus médications simultanément pour traiter plusieurs conditions dont ils souffrent. Ceci étant, nous avons dans notre dataset des enregistrements qui comprennent moins de cinq médicaments. Ces enregistrements constituent des bruits puisqu'elles ne sont pas des polypharmacies et donc les hospitalisations lié a ces dernières sont probablement dû a d'autres facteurs autre que la polypharmacie. Elles ne seront pas pertinentes pour la détection des polypharmacies nocives. Pour pouvoir détecter les non-polypharmacies, nous avons créé une nouvelle variable nommée 'poly' qui contient le nombre de médicaments dans chaque prescription. Grâce à cette variable nous pouvons supprimer toutes les enregistrements dont le nombre de médicaments est inférieur a 5. Après élimination des non-polypharmacies, nous sommes passé de 30220351 à 11830229 enregistrements, soit une réduction de plus de la moitié des données.

La distribution des hospitalisations ou non selon le nombre de médicaments prises simultanément est la suivante:

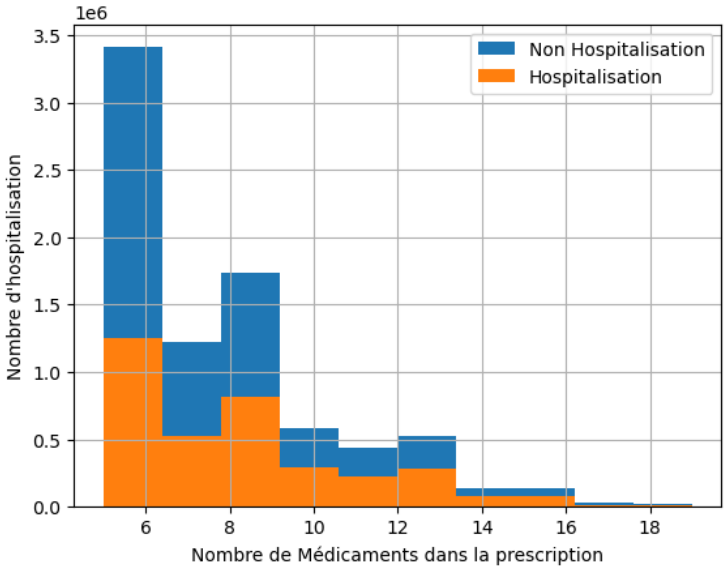

Figure 5. Distribution des hospitalisation selon le nombre de médicaments dans la prescription

2.4. **Élimination des bruits dû à l'insensibilité aux polypharmacies nocives.**

Dans la figure 4, on a pu constater que plus le nombre de médicaments pris simultanément augmente, plus le nombre d'observation diminue, ce qui est tout a fais normal puisque il y a peu de personne à qui on prescrit de tel grands nombre de médicament en même temps. Mais, on peut aussi constaté que même lorsque le nombre de médicaments prescris est très grand (a partir de 13 ou la queue de la distribution semble commencer) le nombre d'hospitalisation reste plus faible que celui de non-hospitalisations. Cela veut dire qu'il y a des personnes qui prennent plus de 10 médicaments simultanément et qui ne sont pas hospitalisées, alors que plus le nombre de médicaments augmente, la probabilité d'avoir des combinaisons nocives devient de plus en plus important. Il s'agirait là donc probablement des personnes résistants aux combinaisons nocives. Leur non-hospitalisation a de fortes chances de constituer un bruit pour le signal d'hospitalisation. Il parait donc judicieux d'éliminer ces cas pour réduire les bruits. Après suppression de ses bruits, les données passent de 11830229 à 11287583 enregistrement.

2.5. **Élimination des enregistrements redondants**

Nous considérons un enregistrement comme redondant lorsqu'il n'apporte aucune information supplémentaire qui pourrait aider à la détection des polypharmacies nocives. Par exemple, pour un même patient qui prend une certaine combinaison de médicaments et qui est hospitalisé a la période A, lorsqu'il prend cette même combinaison de médicaments à la période B et est encore hospitalisé, nous considérons que ce deuxième enregistrement est redondant puisqu'on a les mêmes combinaisons et les mêmes résultats (même cause même effets). Le fait d'avoir une redondance sur les combinaisons des médicaments et la valeur de l'attribut « hospit » associé pourrait induire de fausses combinaisons néfastes. En effet , un individu particulier pourrait être intolérant à un ou une sous combinaison de la combinaison initiale. Si on considère que ce même phénomène peut se produire plusieurs fois pour plusieurs individus dans nos données, le fait d'avoir la même combinaison plusieurs fois

pourrait donc duper le système entraîné qui pourrait classifier comme mauvaise la combinaison initiale entière. Nous supprimons donc les enregistrements redondants. Si le résultat de la même combinaison était différent, on aurait une nouvelle information, de ce fait, on l'aurait gardé. Après suppression des enregistrements redondants, nous sommes passés de 11287583 à 8069307.

2.6. **Réduction de la dimensionnalité**

Nous avons au total 23 attributs en comptant celle que nous avons ajouté pour la détection des polypharmacie ('poly'), ce qui est considérable en terme de dimensionnalité. Notre modèle gagnera en simplicité si nous pouvons supprimer certains attributs. Mais il est crucial de nous assurer que les attributs que nous allons supprimer ne contiennent pas d'informations importantes pour la détection des polypharmacies nocives.

- **L'attribut artificielle 'poly'** Il va de soit que l'attribut 'poly' que nous avions créé peut être supprimer sans aucune crainte puisqu'elle nous avait seulement permis de détecter et d'éliminer les non-polypharmacies.
- **L'attribut 'timestamp':** Lorsque nous considérons la figure 2 plus haut (*Figure 2. Distribution des hospitalisations en fonction du temps*) on peut remarquer que les hospitalisations sont identiquement proportionnelles durant toute la période visualisé. Le nombre d'hospitalisations et celui de non-hospitalisations augmentent et diminuent ensemble toujours avec la proportionnalité de 21% et 79% respectivement. On peut donc éliminer l'attribut «timestamp».
- **Les autres attributs:** Visualisons le nombre de prescriptions de chaque médicament et le nombre d'hospitalisations où sont prescrits chacun de ces médicaments. Avec un tel graphique, si nous pouvions observer par exemple quelques attributs qui ont comparés aux autres un nombre très faible nombre d'hospitalisations, on pourrait conclure que leurs combinaisons ont des chances de ne pas être nocives.

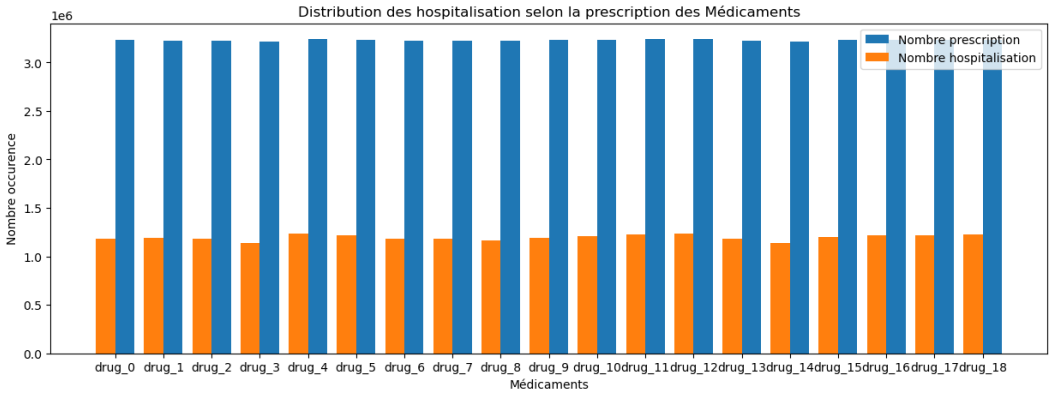

Figure 6. Nombre d'hospitalisation à l'apparition de chaque médicament

On constate qu'à l'apparition de tout les médicaments, il y a presque les mêmes nombre d'hospitalisations. Cela ne nous apporte aucune information décisionnelle. Ce n'est donc pas une bonne piste à explorer

2.7. **Balancement final des données**

Après l'analyse et le nettoyage des données, nous avons au final 8069307 enregistrement et 21 attributs dont les 19 médicaments, 'patient_id' puis 'hospit'. Le balancement final des classe se présente comme suit: 65% de non hospitalisation et 35% d'hospitalisation

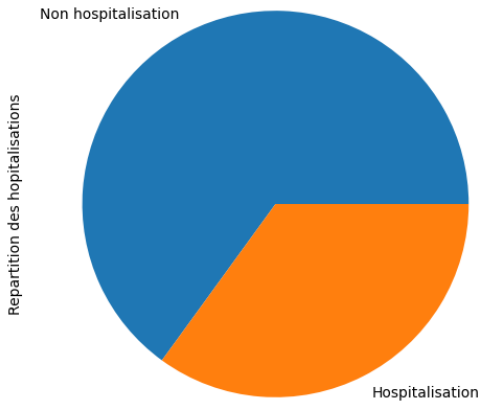

Figure 7. Balancement final des classes

3. **Les Algorithmes d'apprentissage proposés.**

3.1. **Présentation de l'algorithme Random forest**

L'objectif de notre projet est d'identifier les combinaisons de médicaments nocives à partir de nos données. Étant donné la nature du problème et la structure de nos données, nous avons choisi de traiter ce problème comme une classification binaire. Les classes sont les patients qui ont été hospitalisés et ceux qui ne l'ont pas été. Ensuite nous sélectionnerons les combinaisons d'attributs qui ont le plus d'influence sur la précision de notre modèle de décision.

Dans notre cas, où nous n'avons que des attributs binaires, nous avons décidé d'utiliser l'algorithme de "Forêt aléatoire", qui est une méthode d'apprentissage automatique basée sur des arbres de décision. Un arbre de décision est une structure d'attributs en forme d'arbre qui permet de prendre des décisions en suivant un chemin à partir de la racine jusqu'à la feuille, en fonction des valeurs des variables de la donnée à prédire. L'apprentissage d'arbres de décision consiste à sélectionner, à chaque étape, la variable qui sépare le mieux les données en deux sous-groupes homogènes en termes de la variable à prédire. Cette sélection est basée sur une mesure de l'impureté, telle que l'indice de Gini ou l'entropie de Shannon.

Cependant, cette méthode peut souffrir de sur-ajustement et peut être sensible aux données manquantes ou bruyantes. C'est pourquoi nous nous sommes tournés vers l'algorithme de "Forêt aléatoire", qui construit chaque arbre de décision à partir d'un échantillon aléatoire des données d'entraînement, avec des variables choisies au hasard pour chaque arbre, afin de réduire le sur-ajustement et la corrélation entre les arbres individuels. Le modèle de "Forêt aléatoire" combine ensuite les prédictions de tous les arbres pour obtenir une prédiction finale plus fiable et plus stable.

Pour notre problème, nous allons utiliser l'importance des attributs de "Forêt aléatoire" pour identifier les attributs qui contribuent le plus à la prédiction. Nous pourrons ensuite sélectionner un sous-ensemble d'attributs basé sur cette importance et entraîner un nouveau modèle de "Forêt aléatoire" en utilisant uniquement ces attributs. Enfin, pour vérifier la fiabilité de nos attributs sélectionnés, nous allons valider les performances de notre modèle avec ces attributs sur des données de test.

3.2. **Sélection des hyperparamètres**

Les performances d'un modèle d'apprentissage automatique peuvent être affectées par le choix des hyperparamètres. Pour l'algorithme de "Forêt aléatoire", il est courant de choisir la taille d'échantillonnage des dimensions comme étant la racine carrée du nombre total d'attributs (N), et de prendre 2/3 des données pour chaque arbre. Le nombre optimal d'arbres peut varier en fonction du problème et des données spécifiques, mais il est courant de commencer par un petit nombre et d'augmenter progressivement jusqu'à ce que les performances se stabilisent. Il est important de noter que les valeurs optimales des hyperparamètres peuvent nécessiter des expérimentations pour trouver les meilleures performances.

3.3. **Techniques d'apprentissage par Forage de patrons fréquents**

En sachant que le but de l'entraînement est de trouver un ensemble de combinaisons menant à une hospitalisation , on peut ramener notre problème a une découverte de patrons fréquents. On va considérer comme un itemset l'ensemble des valeurs des différents médicaments ainsi que la valeur de l'attribut hospit . Un exemple de patrons serait «0 0 1 1 0 1 1 0 1 0 1 1 0 0 0 0 1 1 0 » où les items sont les valeurs des médicaments. Ainsi notre but serait de trouver progressivement les 6 , 7 , 8 , 9 , 10...15-itemsets fréquents et ensuite utiliser ces itemsets fréquents pour déterminer des règles d'association de la forme : «0 0 1 1 0 1 1 0 1 0 1 1 0 0 0 0 1 1 0 » —> 1 où 1 est le signal d'hospitalisation. Comme seuil du support on pourrait initialement considérer 70 pourcent pour s'assurer d'avoir suffisamment vu la combinaison avant de statuer sur la confiance du fait que la combinaison soit effectivement nocive. Cependant nous pourrons faire varier ce seuil pendant la résolution du problème. De même, nous pourrons ensuite utiliser une confiance de 90 pourcent des règles d'association et faire varier cette proportion pendant l'étude selon le besoin. Vu que nous nous trouvons dans un contexte de bases de données massives si les algorithmes de bases tels que 'A priori' et ces différentes variantes s'avèrent être inapplicables , nous utiliserons les algorithmes tels que Wolf Search Algorithm , ou encore Dragonfly Algorithm qui sont plus pour la découvertes de patrons dans ce contexte.

3.4. **Techniques d'apprentissage par de réseaux de neurones**

Nous avons aussi l'intention d'explorer l'utilisation des réseaux de neurones pour découvrir des liens cachés entre les médicaments. Pour cette technique , on aura en entrée du réseau un vecteur qui contient les différentes valeur des différents médicaments.Après un entraînement supervisé où nous associerons chaque vecteur à son label qui est le signal d'hospitalisation , nous testerons le pourcentage de réussite sur les données test. On utilisera en sortie du réseau une fonction softmax qui donnera la probabilité selon laquelle le signal d'hospitalisation soit 0 ou 1 .Pour avoir un résultat sur ,on détectera toute les combinaisons menant à une hospitalisation selon un seuil de pourcentage raisonnable selon l'avancée de l'apprentissage ainsi que la comparaison avec les premières techniques mises en place.

4. Choix initial de procédure de tests.

4.1. Séparation du jeu de données

Une bonne pratique de machine learning est de tester le modèle qu'on a entraîné. Le but de cette partie est de savoir comment notre système généralise bien sur de nouvelles données. Il y a plusieurs techniques de test ainsi que plusieurs métriques d'évaluation de performance dont on peut se servir dans le but de jauger cette généralisation. Nous avons environ 8 millions d'exemples à envoyer à l'algorithme d'apprentissage en équilibrant le nombre de données par classe. Nous allons diviser le jeu de données en un sous-ensemble d'entraînement (80%), un sous-ensemble de validation et un sous-ensemble de test. Le jeu de données sera mélangé de manière aléatoire pour éviter les biais potentiels liés à l'ordre des enregistrements. Nous utiliserons également la méthode de validation croisée K-folds (K [5,10]) pour évaluer le modèle en entraînant itérativement sur K-1 plis et testant sur le pli restant, et en moyennant les performances sur les K plis de test pour une évaluation plus précise et invariante. Bien que nous pensions pouvoir entraîner le modèle sans cette méthode, nous l'utiliserons pour obtenir des métriques plus précises et invariantes.

4.2. Métriques pertinentes pour random forest

Lorsque nous évaluons notre algorithme, plusieurs mesures sont disponibles pour juger de sa capacité à prédire l'intention de vote des personnes sondées. Tout d'abord, nous utilisons une matrice de confusion de 8x8 pour les 8 classes à prédire. Après que notre algorithme ait effectué ses prédictions sur les 10% d'individus de test, nous évaluons les résultats en examinant combien de fois notre algorithme a prédit une classe donnée pour un enregistrement, et comparons ces prédictions avec les vraies classes. En utilisant cette matrice, nous pouvons calculer les vrais positifs, les vrais négatifs, les faux positifs et les faux négatifs pour chaque classe, en faisant la somme des cases appropriées. Les vrais positifs sont déterminés par la case où l'algorithme a prédit correctement la classe de la donnée, tandis que les vrais négatifs sont la somme des cases représentant les classes qui n'ont pas été choisies par l'algorithme et qui ne sont pas la vraie classe. Les faux positifs sont obtenus en sommant les cases où l'algorithme a prédit la classe incorrecte, tandis que la vraie classe était une autre. Enfin, les faux négatifs sont obtenus en faisant la somme des cases où l'algorithme a prédit une autre classe que celle évaluée.
À partir de ces mesures initiales, nous pouvons calculer la précision, le rappel et le f score. La précision est le nombre de vrais positifs divisé par le nombre total de prédictions positives de l'algorithme, et elle indique le pourcentage de prédictions d'une certaine classe qui appartiennent réellement à cette classe. Le rappel est le nombre de vrais positifs divisé par le nombre total de vraies classes positives, et il mesure la proportion de vrais positifs que l'algorithme a réussi à prédire parmi tous les cas. Le score F est la moyenne harmonique de la précision et du rappel et se situe entre 0 et 1. Cette mesure est utilisée pour évaluer l'algorithme de manière globale, car elle prend en compte à la fois la précision et le rappel sans prendre en compte les vrais négatifs.

4.3. Métriques pertinentes pour forages de patrons

Pour l'évaluation des patrons découverts , en plus du support et la confiance qui sont les métriques qui vont nous permettre de rouler nos algorithmes , nous allons utiliser d'autres métriques qui mesurent la corrélation entre les évènements qui constituent les règles d'association en vue d'éliminer les patrons inintéressants. Nous allons privilégier les mesures qui ne sont pas influencées par le nombre de transactions telles que la métrique de mesure cosinus $cos(A,B) = \frac{P(A \cap B)}{\sqrt{P(A)*P(B)}} = \frac{support(A \cap B)}{\sqrt{support(A)*support(B)}}$

5. Conclusion

Ce rapport présente une analyse de données médicales pour identifier les combinaisons de médicaments qui peuvent causer une hospitalisation. Les données ont été nettoyées et analysées, et 8069307 enregistrements avec 21 attributs ont été conservé. Le balancement final des classes est de 65% pour les non-hospitalisés et de 35% pour les hospitalisés. L'algorithme principalement choisi pour résoudre le problème est la Forêt aléatoire. C'est une méthode d'apprentissage automatique basée sur des arbres de décision qui réduit le surajustement et la corrélation entre les arbres individuels. L'importance des attributs de la Forêt aléatoire sera utilisée pour identifier les attributs qui contribuent le plus à la prédiction. Un sous-ensemble d'attributs sera sélectionné en utilisant cette importance, et un nouveau modèle de la Forêt aléatoire sera entraîné avec ces attributs. En plus de la Forêt aléatoire, les réseaux de neurones ont également été proposés comme alternative pour résoudre ce problème, en raison de leur capacité à apprendre des modèles complexes et à généraliser à de nouvelles données. Nous prévoyons de tester les deux algorithmes pour comparer leurs performances. En outre, nous avons proposé la découverte de patrons fréquents comme une méthode supplémentaire pour trouver un ensemble de combinaisons de médicaments menant à une hospitalisation. Les itemsets fréquents seront découverts progressivement pour ensuite être utilisés pour déterminer des règles d'association. Dans le prochain rapport, nous allons tester tous les algorithmes proposés et évaluer leurs performances en utilisant des mesures d'évaluation appropriées. Nous prévoyons également d'explorer plus en détail les attributs qui ont le plus d'importance pour la prédiction et de comprendre comment ils sont associés aux hospitalisations. Nous espérons que ces résultats pourront aider les professionnels de la santé à mieux comprendre les risques associés aux combinaisons de médicaments et à prévenir les hospitalisations inutiles.