PREDICTIVE MODELING OF HOSPITAL READMISSION RATES USING ELECTRONIC MEDICAL RECORD-WIDE MACHINE LEARNING :**A CASE-STUDY USING MOUNT SINAI HEART FAILURE COHORT**

AUTEURS DE L'ARTICLE : KHADER SHAMEER, KIPP W JOHNSON, ALEXANDRE YAHI, RICCARDO MIOTTO, LI LI, DORAN RICKS, JEBAKUMAR JEBAKARAN, PATRICIA KOVATCH, PARTHO P. SENGUPTA, ANNETINE GELIJNS, ALAN MOSKOVITZ, BRUCE DARROW, DAVID L REICH, ANDREW KASARSKIS, NICHOLAS P. TATONETTI, SEAN PINNEY AND JOEL T DUDLEY.

AUTEUR DU RAPPORT : MEDOU Daniel Magloire, étudiant en Master recherche en informatique. IFI, 12 Janvier 2018

SUPERVISEUR ACADEMIQUE : Dr. HO Tuong Vinh

Institut Francophone International (IFI)

***Résumé*—Ce rapport est fait dans le cadre de l'unité d'enseignement intitulée "Bibliographie et Étude des Cas" donc le but est d'amener l'étudiant à pouvoir faire une étude détaillée d'un article. Dans le cas d'espèce, notre article nous fait part des travaux liés à la réadmission hospitalière. Cette dernière, liée aux maladies chroniques ou aiguës telles que dans l'article de référence que nous étudions à savoir "les accidents vasculaires cérébraux, l'insuffisance cardiaque, l'infarctus du myocarde et la pneumonie, le diabète, etc..." de nos jours est un défi que tous les États des pays du monde entier cherchent à réduire le taux de réadmission des patients souffrants de l'une de ces affections trente (30) jours après la première hospitalisation. Dans notre article de référence, les auteurs dans leur recherche, se sont fixés pour objectif de connaître la cause ou les causes liées à la réadmission des patients souffrants des affections chroniques. Sur ce, en collaboration avec les responsables de l'hôpital "MOUNT SINAI" aux USA, décident grâce à la base de données des patients mise à leur disposition de mettre en place un modèle de prédiction du taux de réadmission des patients à partir des informations prélevées au Dossier Médical Électronique (DME) du patients. De cette étude, un vaste répertoire des variables du DME des patients atteints d'insuffisance cardiaque avait été évalué. L'article précise que la cohorte (Ensemble d'individus ayant vécu un même événement au cours d'une période donnée) comprenait 1 068 patients et 178 patients ont été réadmis dans un intervalle de 30 jours (taux de réadmission de 16,66%). Au total, 4 205 variables ont été extraites du DME, y compris les codes de diagnostic (n = 1 763), les médicaments (n = 1 028), les mesures de laboratoire (n = 846), les interventions chirurgicales (n = 564) et les signes vitaux (n = 4). Les auteurs de cet article avaient opté pour une modélisation "multi-étapes" et avaient utilisé pour leur expérimentation juste l'algorithme "NAÏVE BAYES" dans le but de classer les cas (réadmis) et les contrôles (non réadmis) ceci pour la première étapes de la modélisation "multi-étapes". Dans la deuxième étape, les caractéristiques contribuant au risque prédictif des modèles indépendants ont été combinées dans un modèle composite en utilisant une méthode de sélection de caractéristiques basée sur la corrélation. Tous les modèles avaient été formés et testés en utilisant une méthode de validation croisée 5 fois, avec 70% de la base de données pour l'apprentissage et les 30% restants pour les tests.Comparés aux modèles prédictifs existants pour les taux de HF (ASC de l'ordre de 0.6-0.7), les résultats de notre modèle prédictif DME (AUC = 0.78, Accuracy = 83.19%) et les stratégies de sélection des caractéristiques à l'échelle du phénomène sont encourageants et révèlent l'utilité d'un tel apprentissage machine par datadrive.**

Du point de vue de travaux connexes effectués par d'autres auteurs, nous avons étudier deux autres articles donc les auteurs chercheurs se sont aussi penchés dans le cas de la réduction du taux de réadmission hospitalière des patients souffrants des affections chroniques ou aiguës ceci pour les trente (30) jours après une première hospitalisation. Ces chercheurs ont abordés le sujet sous un autre angle. Le premier article connexe, dont évaluation du taux de réadmission avait été fait pour les patients atteints du diabète en Inde, l'objectif de cette étude est d'évaluer l'impact de techniques de pré-traitement sélectionnées sur la prédiction du risque de réadmission de 30 jours chez les patients diabétiques en Inde en se basant sur le DME. Le second article connexe lui à pour objectif d'explorer l'utilisation de la régression logistique conditionnelle pour augmenter la précision de la prédiction de réadmission hospitalière des patients atteints d'insuffisance cardiaque. Une étude synthétique des travaux connexes sera faite dans le contenu du rapport.

Keywords*—*Réadmission hospitalière . AUC (Area Under the Curve). Cohorte . Accuracy. Data mining . Diabète . Sélection de fonctionnalité . Imputation de la valeur manquante . Prédire les taux de réadmission . Pré-traitement . évaluation des risques . classification binaire . régression logistique conditionnelle (RLC). Comorbidité. Résumé

I. INTRODUCTION

La réadmission hospitalière est considérée comme étant une seconde hospitalisation. Ce phénomène est un défi que bon nombre d'organisations de réglementation comme les Centers for Medicaid and Medicare Services (CMS) aux États-Unis pour ne citer que ceux-là, accordent une grande importance voir capitale car il en est de la survie des êtres humains. De ce faite, l'étude sur la réadmission hospitalière aux États-Unis résulte comme le spécifie les auteurs de cet article des affections chroniques ou aiguës et ont énuméré les maladies entrant dans ce package à savoir les accidents vasculaires cérébraux, l'insuffisance cardiaque, l'infarctus du myocarde et la pneumonie. En effet, l'observation avait été faite sur un cas de maladie bien précis parmi tous ceux qui, cités ci-dessus à savoir les crises cardiaques (insuffisance cardiaque)

et le diabète donc le taux des patients réadmis est relativement élevés. Malgré la mise en œuvre de lignes directrices d'opérations de prestation de soins de santé de haute qualité créées par les autorités de réglementation (CMS), les patients sont réadmis en hospitalisation avant les trente (30) suivant leur hospitalisation initiale. Le problème reste donc à connaître comment diminuer au maximum le taux de réadmission hospitalière de cette pathologie qu'est l'insuffisance cardiaque et le diabète ? Cependant, dans l'article soumis à notre étude dans le cadre de l'unité d'enseignement intitulé ≪ Bibliographie et Étude des Cas ≫, KHADER SHAMEER et al [1], par le biais de l'apprentissage automatique prédire les probabilités de réadmission en se basant des variables du dossier médicale électronique des patients atteints d'insuffisance cardiaque, Le présent rapport nous donnera un résumé détaillé de l'article ainsi que des travaux connexes effectué par d'autres chercheurs [2] et [3], suivant le plan que nous avons arrêté.

II. CONTEXTE, OBJECTIFS ET PROBLÉMATIQUE DE L'ARTICLE

KHADER SHAMEER et al, dans leur article présentent les points essentiels abordés dans la thématique de l'apprentissage automatique et de la fouille des données donc l'objectif principal est de développer une approche de sélection des caractéristiques basées sur des données électroniques, à l'échelle du dossier médical et de l'apprentissage automatique afin de prédire les probabilités de réadmissions hospitalières des affections chroniques ou aiguës. Les deux domaines cités à savoir l'apprentissage automatique et de la fouille des données ont fait l'objet d'une étude sous forme de module respectivement du Master 2 et du Master 1 durant notre parcours académique à L'IFI (Institut Francophone International). Étude faite sur deux niveaux. Le premier étant théorique et l'autre pratique avec la mise en œuvre des travaux pratiques au moins trois par domaine. Il est question (objectif) pour les auteurs de cet article de mettre en place un modèle pour prédire les taux de réadmission chez les patients souffrant d'insuffisance cardiaque et ce modèle est basé sur les données prélevées dans le dossier médical électronique. **La sortie de ce modèle est de classifier les patients Réadmis (RA) et Non-Réadmis (Non-RA) qui est classée comme une tâche de classification binaire**.

III. PROPOSITION DE LA SOLUTION

Pour résoudre ce problème de réadmission qui est lié à l'insuffisance cardiaque, les auteurs de cet article ont proposé une solution qui est l'une des l'une des premières tentatives d'utiliser les données à l'échelle du phénomène pour identifier de nouveaux facteurs de réadmission liés à l'insuffisance cardiaque et développer des modèles de prédiction à l'échelle du Dossier Médical Électronique (DME). Pour ce faire, deux étapes sont mises en place pour la stratégie de modélisation. La première qui est une approche basée sur la sélection des caractéristiques basées sur les données (variables) fournies par le DME. Dans ce cas, la sélection des variables qui contribueront à la construction du modèle sont retenus par la méthode de sélection des caractéristiques basées sur la corrélation. La deuxième partie est la construction des modèles en utilisant une méthode de validation croisée 5 fois (k=5).

IV. PRÉSENTATION DES RÉSULTATS

Les résultats de cette études sont liés à plusieurs éléments à savoir ; les caractéristiques de la cohorte, la sélection des caractéristiques à l'échelle de DME et modélisation prédictive à l'aide de cinq modalités de données différentes, la réduction de fonctionnalités et affinement du modèle et enfin la comparaison avec les modèles existants de réadmission pour l'insuffisance cardiaque. Pour une bonne compréhension, nous avons trouvé nécessaire de présenter selon chaque étape mise en place par les auteurs.

A. La cohorte

La cohorte, elle est constituée de 4205 variables extraites du DME. Celle-ci se trouvant être classées en cinq (05) catégories à savoir :

- Les codes de diagnostic (codes de la CIM-9 et IMO), avec un total de n = 1763 variables pour CIM-9. Il est à noter que tous les codes ont été unifiés selon CIM-9 ;
- Les procédures (codes CIM-9, SNOMEDCT et CPT), avec n = 564 variables et toutes les procédures ont été utilisées. Donc celles codées SNOMED-CT ou CIM-9-CM ;
- Les médicaments, avec n = 1028 variables et toutes les données sur les médicaments avaient été normalisées à l'aide de RxNorm ;
- Les mesures du laboratoire, avec n = 846 ;
- Les signes vitaux, avec n = 4 constitués du ≪ pouls, le taux de respiration, la pression artérielle systolique, les battements cardiaques et la température.

Lorsque nous additionnons ces quantités toutes les variables (1763 + 564 + 1028 + 846 + 4 = 4205), nous avons le total exacte de la cohorte qu'avait été utilisé par les auteurs de cet article.

B. La sélection des caractéristiques

La sélection des caractéristiques à l'échelle du DME et la modélisation prédictive à l'aide de cinq modalités différentes est inspirée de la figure 1 de l'article dans sa page 4 comme le montre la figure 1 dans notre rapport.

Cette figure nous montre les différentes étapes à suivre pour mettre en place le modèle qui permettra de faire la prédiction des patients souffrant de l'insuffisance cardiaque ayant déjà été admis en hospitalisation pour au moins une fois. Les auteurs de l'article cherchent à évaluer les taux potentiels de réadmission de ces patients dans les 30 jours qui suivent leur première hospitalisation. **La sortie de ce modèle est LA CLASSIFICATION DES PATIENTS READMIS ET NON-READMIS**. Dans la première étape, K. SHAMEER et al. passent tout d'abord au pré-traitement du jeu de données pour le cas des données manquantes et des doublons si elles existent. Une fois ce pré-traitement terminé, s'en suit la sélection des variables qui permettrons de mettre en place le modèle. Et cette sélection est faite en fonction des cinq variables retenues

du DME, chacune faisant objet d'un modèle généré qui permet de sélectionner les caractéristiques pertinentes. Ces dernières ont été comparées à l'aide de trois méthodes à savoir la régression logistique, PCA (Analyse en Composants Principal) et la mesure orthogonales ceci dans le but de comprendre l'espèce variable et leur relation inhérentes ou leur corrélation. La deuxième étape consiste à générer un modèle qui permettra d'effectuer des prédictions. La troisième partie permet de voir la précision du modèle ainsi que la qualité de ce modèle avec AUC (Area Under the Curve) comprise entre 0.5 et 1. La quatrième concerne la validation du modèle avec les données de tests. La figure 2 est le tableau des résultats du modèle mis en place par KHADER SHAMEER et al en utilisant l'algorithme Naïve Bayes.
(FIGURE 1).

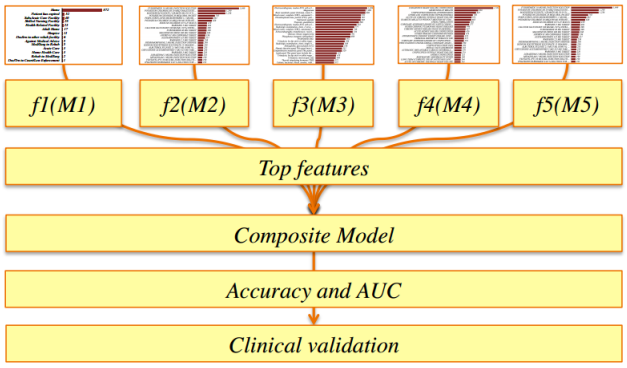

FIGURE 1. Architecture d'apprentissage automatique à l'échelle du DME et stratégie de modélisation prédictive pour trouver les facteurs déterminants des taux de réadmission

Au vue de ce tableau, Le modèle final est développé en utilisant 105 sur les 4205 caractéristiques avec une AUC = 0,78, proche de 1 ce qui montre la bonne qualité du modèle et une précision de test de validation croisée de 83,19 % confirmant le résultat de l'AUC.
(FIGURE 2).

Data-element	Type	Encoding	Accuracy	AUC	Features
Diagnosis	ICD-9 Diagnosis	Binary	70.3297%	0.605	34/1763
Procedures	ICD-9-Procedure	Binary	77.907%	<0.50	4/273
Procedure	CPT-codes	Binary	72.9858%	0.553	8/564
Medications	Medication name and dosage	Binary	81.9048%	0.615	26/1028
Labs	Non-descriptive lab measurements	Continuous	73.9336%	0.535	29/846
Composite model	Combined features	Hybrid	83.9000%	0.780	105

FIGURE 2. Récapitulatif des différents prédicteurs et caractéristiques bayésiens compilés à l'aide de la méthode CFS

C. Comparaison avec les modèles actuels de réadmission pour l'insuffisance cardiaque

Pour une étude comparative, les auteurs ont présenté les résultats obtenus à partir de certains critères de sélection des caractéristiques et avec un algorithme d'apprentissage automatique bien précis (Naïve Bayes) et ont obtenus comme résultat AUC = 0.78 et Accuracy = 83.90 %. Ces derniers avaient été comparés à une modélisation prédictive e réadmissions hospitalières utilisant les données de santé du Québec, canada par Hosseinzadeh et al la méthode de sélection des caractéristiques pour la construction du modèle reste inconnue et les algorithmes utilisés étaient Naïve Bayes (0.65) et Random Forest (0.64). Pour la cohorte de diabétiques d'un hôpital en Inde, Duggal et al ont montré que Naïve Bayes (0,67) présentait des économies associées à la réadmission plus élevées que la régression logistique (0,67), Forêts aléatoires (0,68), Adaboost (0,67) et réseaux neuronaux (0,62) . Futoma et.al ont montré que Random Forests (0,68) et l'apprentissage en profondeur en utilisant des réseaux de neurones (0,67) ont un taux de précision similaire inférieur à 1 million de patients et inférieur à 3 millions d'admission.

V. ÉTUDE DES TRAVAUX CONNEXES : ARTICLE 1

Dans l'article [2], les auteurs dans la même lancée que ceux l'article de référence cherche un moyen d'améliorer le taux de réduction de réadmission hospitalière des trente (30) jours suivant l'hospitalisation initial du patient souffrant du diabète. Dans leurs recherches, leur objectif principal était **d'évaluer l'impacte des techniques de pré-traitement des données sélectionnées sur la prédiction du risque de réadmission de 30 jours chez les patients diabétiques** précisément le cas des patients de l'hôpital en Inde. Dans cet article, **Reena Duggal et al.** précise que le diabète est associé à un risque accru de réadmission à l'hôpital et plusieurs fois dans un court laps de temps [2]. La prédiction du risque de réadmission est une tâche complexe car elle nécessite l'intégration de divers attributs liés aux patients, tels que l'état de santé des patients, les facteurs sociodémographiques et l'utilisation des services de santé. Le premier défi majeur est la compréhension et l'identification des attributs pertinents (facteurs ou caractéristiques) et des valeurs de données existant dans l'ensemble de données sur les soins de santé à haute dimension qui mène à la réadmission des patients atteints de diabète. Les données sur les soins de santé dans le monde réel sont bruyantes, incohérentes, hétérogènes et infestées de nombreuses valeurs manquantes. Ainsi, avant de commencer la tâche de construction du modèle, il est essentiel de pré-traiter les données de manière efficace et de les rendre appropriées pour la modélisation prédictive.

A. Critères appliqués pour extraire l'ensemble des données nécessaires pour la visite des patients

Les auteurs de [2], ont données la liste des critères d'extraction des données

1) Admission à l'hôpital pour patients hospitalisés.
2) Durée du séjour (DDS) est minimum 1 jour. Garde de jour non incluse.
3) Les patients de plus de 18 ans.
4) Rencontres exclues en raison de l'accouchement.
5) Visite à l'hôpital d'un patient atteint de diabète.

B. Type et la nature des données

1) 58 625 hospitalisations dans un hôpital indien.

2) Diagnostic extrait, antécédents et procédure informations, médicaments et données de laboratoire. Rencontres du 1er juillet 2013 au 31 juillet 2015.
3) 18 616 admissions en garderie.
4) 3 118 admission avec des patients de moins de 18 ans au moment de l'admission.
5) 994 admissions avec diagnostic grossesse liée ou obstétrique.
6) 26 516 patients n'ayant pas de diagnostic ou d'antécédents de DIABÈTE.
7) 9 381 admissions qualifiées (patients diabétiques), **1 211 réadmis dans les 30 jours** et 8 170 non réadmis dans les 30 jours Nous constatons qu'il y a un déséquilibre très considérable sur la quantité des patients réadmis et non réadmis.

C. Techniques de pré-traitement des données

Cette partie est l'étape objective des auteurs dans [2]. La techniques de pré-traitement des données utilisée trouve sa base sur trois points essentiels et chacun des principes ayant deux approches.

1) Sélection des caractéristiques.
   — L'approche par filtre basé sur la corrélation.
   — Le test de Chi2 de Pearson.
2) Imputation des valeurs manquantes.
   — La technique d'imputation moyenne pour les attributs numériques.
   — Le mode pour l'attribut nominal de tous les cas observés.
3) Réduction du déséquilibre des classes.
   — La technique de sur-échantillonnage.
   — La technique de sous-échantillonnage.

D. Architecture globale du processus de modélisation

(FIGURE 3).

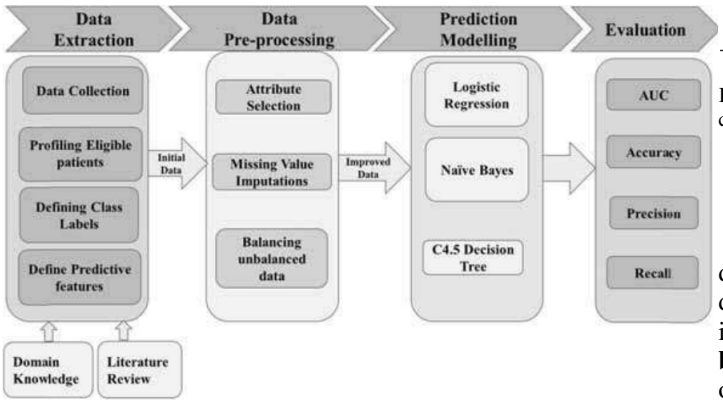

FIGURE 3. L'architecture globale du processus de modélisation

E. Caractéristiques utilisées dans le processus de prédiction

Dans cette partie, il est à signaler que 31 attributs étaient utilisés pour la construction de modèle. Trois types de caractéristiques sont présentés à savoir :

1) Caractéristiques démographiques des patients.
   — Âge.
   — Sexe.
2) Autres caractéristiques.
   — Département de provenance du patient.
   — Source d'admission.
   — Médicament pris contre le diabète.
   — La durée de séjour, etc...
3) Comorbidité.
   — Problème de coeur.
   — Hypertension.
   — Insuffisance rénale, etc...

F. Évaluation expérimentale et outil

1) Outil utilisé : WEKA v3.6.12.
2) Le test du modèle prédictif a été effectué en utilisant une procédure de validation croisée 10 fois supérieure.
3) Quatre paramètres d'évaluation ont été considérés pour évaluer la qualité des différents modèles.
   — Zone sous la courbe (AUC).
   — Precision.
   — Recall.
   — Accuracy.

G. Tableau récapitulatif des résultats expérimentaux

(FIGURE 4).

Feature selection	Missing value imputation	Balancing	Predictive model	AUC	Precision	Recall	Accuracy
Baseline							
–	–	–	Logistic regression	0.68	0.56	0.08	0.87
–	–	–	Naïve Bayes	0.68	0.33	0.23	0.84
–	–	–	Decision tree	0.56	0.42	0.02	0.87
Top 3—AUC, precision, and recall							
Chi-square	Mean/mode imputation	Oversampling	Decision tree	0.86	0.82	0.85	0.83
Correlation-based	Mean/mode imputation	Oversampling	Decision tree	0.83	0.77	0.80	0.78
Chi-square	Mean/mode imputation	Oversampling	Logistic regression	0.83	0.74	0.78	0.75
Top 3—accuracy							
Chi-square	Mean/mode imputation	–	Logistic regression	0.68	0.56	0.08	0.87
Correlation-based	Mean/mode imputation	–	Naïve Bayes	0.66	0.56	0.10	0.87
Chi-square	Mean/mode imputation	–	Decision tree	0.58	0.48	0.04	0.87

FIGURE 4. Tableau des tests avec et sans avec et sans extraction des caractéristiques et pré-traitement des données

VI. ÉTUDE DES TRAVAUX CONNEXES : ARTICLE 2

Dans l'article [3], les auteurs abordent le cas de réadmission des patients atteints d'insuffisance cardiaque comme le cas de l'article de référence [1]. K. Zhu et al., dans leur article intitulé **Predicting 30-day Hospital Readmission with Publicly Available Administrative Database**, publié en 2015, ont aussi mené leurs recherches aux USA précisément en Californie et ont pour objectif é**d'explorer l'utilisation de la Régression Logistique Conditionnelle (RLC) pour augmenter la précision de la prédiction de réadmission hospitalière**

des patients atteints d'insuffisance cardiaque. Pour ce faire, les étapes ci-dessous ont été nécessaire pour l'aboutissement de leur recherche.

A. Critères appliquée pour extraire l'ensemble des données nécessaires pour la visite des patients

1) Analyse des données sur les dossiers de congés pour patients hospitalisés. Les éléments retenus pour cette analyse.
   — Les données démographiques des patients (Âge 65 ans et plus).
   — Les données cliniques.
   — Les données sur l'utilisation des soins en Californie.
2) Extraction des dossiers des bénéficiaires de l'assurance maladie ayant ne expérience hospitalière de 11 mois.
3) Le groupe d'âge était compris entre 65 ans et plus.
4) La décharge s'est produite entre janvier et novembre 2010.
5) Patient qui n'a pas été transféré immédiatement après l'hospitalisation initiale et la réadmission à l'hôpital.
6) Le patient qui a été renvoyé chez lui pour recevoir des soins personnels, des soins à domicile ou des soins infirmiers à domicile.

B. Type et la nature des données

1) Données de laboratoire des patients de California Medicare libérés de janvier à novembre 2010.
2) Ensemble de 36 variables retenues pour le développement du modèle de prédiction.
3) L'ensemble de données original contenait 3 970 921 enregistrements.
4) Après avoir effectué tous les éléments d'extraction de données, les dossiers de 22 410 patients sont restés.
5) 17 434 n'ont pas été réadmis dans les 30 jours suivant leur sortie. Classe majoritaire.
6) 4976 ont été réadmis dans les 30 jours suivant leur sortie.
7) Prédicteurs communément utilisés tels que le nombre de maladies chroniques et les procédures de soins actifs.

C. Techniques appliquées aux données

1) Réduction du déséquilibre des classes.
   — La technique de sur-échantillonnage.
   — La technique de sous-échantillonnage.
2) Le régression logistique standard.
   — Obtenir des variables influentes.
   — Déduire des règles de décisions.
3) L'arbre de décisions. Avec comme objectifs.
4) Ensemble de données stratifié et appliqué une régression logistique sur chaque strate de données.
5) **Une validation croisée avait été appliquée dans le but d'évaluer la performance de prédiction globale de la régression Logistique Conditionnelle (CLR)**.

D. Techniques de pré-traitement des données

La techniques de pré-traitement des données utilisée dans [3] se base sur trois points essentiels.

1) Sélection des caractéristiques.
2) Imputation des valeurs manquantes.
   — La technique d'imputation moyenne pour les attributs numériques.
   — Le mode pour l'attribut nominal de tous les cas observés.
3) Réduction du déséquilibre des classes.
   — La technique de sur-échantillonnage.
   — La technique de sous-échantillonnage.

E. Les modèles de classification expérimentés

1) Régression Logistique Standard (ensemble de données complet).
2) Régression Logistique par étapes (ensemble de données complet).
3) Forêt Aléatoire (ensemble de données complet).
4) Machine à Vecteurs de Support (ensemble de données complet).
5) Régression Logistique Conditionnelle (avec deux strates basées sur chacune des trois premières variables).
6) Régression Logistique Conditionnelle (avec quatre strates dérivées d'un ensemble de règles basé sur un arbre de décision).
7) Régression Logistique Conditionnelle (embellissement de (6) avec incorporation de variables supplémentaires d'interaction par paires).

La comparaison des modèles de classification avait été effectué via la validation croisée et évalué l'exactitude de la classification de chaque modèle.

F. Tableau Variables influentes identifiées par régression logistique

(FIGURE 5).

	Decision Rules	Total # of Subjects (# of admitted, # of not admitted)	Influential Variables identified for interactions from CLR2 ($p \leq 0.05$)	
			Variable Name	*p*-value
Stratum 1	DISPUniform = 5; NPR ≤ 4	1660 (1039, 621)	CM_COAG	0.04674
			CM_LYTE	0.02462
			DNR	0.00139
			FEMALE	0.01807
Stratum 2	DISPUniform = 5; NPR > 4	646 (462, 184)	RACE3[a]	0.0455
Stratum 3	DISPUniform = 1 or 6; NCHRONIC ≤ 7	3278 (1309, 1969)	CM_LYTE	0.0245
			NPR	0.0173
			RACE2[b]	0.0251
Stratum 4	DISUniform = 1 or 6; NCHRONIC > 7	4368 (2166, 2202)	CM_METS	0.00046
			CM_TUMOR	0.0207
			CM_WGTLOSS	0.0185
			DNR	0.0348
			MEDINCSTQ3[c]	0.0458

FIGURE 5. Variables influentes identifiées par régression logistique

G. Tableau Variables influentes identifiées par régression logistique

(FIGURE 7).

Classification Model		Prediction Accuracy*
LR: standard logistic regression		0.547
SLR: stepwise logistic regression		0.539
RF: random forests	all original variable	0.577
	only variables selected via SLR	0.574
SVM: support vector machine		0.560
CLR1: conditional logistic regression with 3 influence prediction variables	DISPUniform	0.548**
	NCHRONIC	0.564**
	NPR	0.576**
CLR2: conditional logistic regression with 4 data strata based on the first two layers of the decision tree		0.608**
CLR3: CLR2 + consideration of interacting variables based on identified influence ones in CLR2		0.615**

FIGURE 6. Comparaison du modèle de classification

VII. TABLEAU RÉCAPITULATIF DES ARTICLES

(FIGURE 7).

	Article de référence	Article 1	Article 2
Problème abordé	Réadmission Hospitalière dans les 30 jours	Réadmission Hospitalière dans les 30 jours	Réadmission Hospitalière dans les 30 jours
Objectif	Mise en œuvre d'un modèle prédictif basé sur les données du DME pour prédire les taux de réadmission chez **les patients souffrants d'insuffisance cardiaque aux USA (Californie).**	Evaluer l'impacte des techniques de prétraitement sélectionnées sur la prédiction du risque de réadmission de 30 jours chez **les patients diabétiques en Inde**.	Explorer l'utilisation de la régression logistique conditionnelle pour augmenter la précision de la prédiction de réadmission hospitalière **des patients atteints d'insuffisance cardiaque aux USA (Californie).**
Architecture du modèle	Oui	Oui	Non
Caractéristiques et taille du jeu de données	-1068 patients -178 patients réadmis -4205 variables extraites -105 variables retenues.	-58625 hospitalisations. - 26 516 non diabétique -9381 Patients diabétique -1211 réadmis dans les 30 jours - 8 170 non réadmis dans les 30 jours	- 3 970 921 enregistrements. - 22 410 patients sont restés après extraction des caractéristiques - 17 434 non réadmis dans les 30 jours - 4976 réadmis dans les 30 jours
Méthode de prétraitement des données	-Traitement des données manquantes (Méthode inconnue) - Extraction des caractéristiques	- Sélection des caractéristiques -Imputation des valeurs manquantes - Réduction du déséquilibre des classes	- Réduction du déséquilibre des classes - Le régression logistique standard - L'arbre de décisions - Ensemble de données stratifié **-Une validation croisée**
Algorithmes expérimenté	Naïve Bayes	- Naïve Bayes -Arbre de décision -Régression Logistique	-Régression Logistique standard -Forêt aléatoire -Régression Logistique Conditionnelle.
Résultats	Satisfaction	Satisfaction	Satisfaction
Sortie attendue des modèles	-RA = «Patients réadmis » -NonRA = « Patient non réadmis »	-RA = «Patients réadmis » -NonRA = « Patient non réadmis »	-RA = «Patients réadmis » -NonRA = « Patient non réadmis »

FIGURE 7. Récapitulatif des articles

VIII. CONCLUSION ET PERSPECTIVES

Parvenu au terme de notre analyse des travaux de recherche de **KHADER SHAMEER et al.**[1] qui, nous présentent dans cet article une nouvelle procédure de développement d'une approche de sélection de caractéristiques basées sur les données électroniques à l'échelle du DME et l'apprentissage automatique pour prédire le taux de réadmission hospitalière des patients atteints de l'insuffisance cardiaque dans les 30 jours. Les données provenant de la base de données MySQL de l'hôpital du Mont Sinaï avaient tout d'abord été soumis à un pré-traitement ensuite, de chaque modalité avait été généré un modèle dans le but de sélectionner les caractéristiques inhérentes. Pour cette sélection des caractéristiques, avaient été appliquées des méthodes comme la régression logistique, ACP (Analyse des Composants Principaux). Seules celles des caractéristiques corrélées avaient été retenues pour former le modèle prédictif à mettre en place. **105 sur 4205 variables** (confère FIGURE 2) que forment la cohorte sont retenues après l'extraction des meilleures caractéristiques pour former le modèle et **La sortie de ce modèle est la classification des patients Réadmis (RA) et Non-Réadmis (Non-RA)**. Cependant, **KHADER SHAMEER et al.** pour leurs expérimentations avaient divisé les données en deux jeux « train et test » avec 70 % pour le train et 30 % pour le test. Ce modèle soumis à un algorithme d'apprentissage automatique « **Naïve Bayes** » a donné un taux de précision de test de validation croisée de **83.90 %** et **AUC de 0.78**. Ce qui présente un meilleur résultat comparativement à ceux des autres chercheurs présentés dans cet article. L'étude des autres articles nous donnent d'autres éléments je veux dire approches pour améliorer le taux de prédiction de la réadmission hospitalière des patients atteints des affections chroniques ou aiguës. Dans [2], **Reena Duggal et al.** accordent une importance capitale au pré-traitement des données avant la mise en oeuvre du modèle prédictif. Une fois le pré-traitement effectué et modèle en place, les tests avaient été faits et les valeurs des taux données à retrouver dans la **figure 4** selon les tests effectués sans pré-traitement et ceux effectués avec. Dans cette article, les auteurs ont été satisfaits de leur recherche. Dans [3], **K. Zhu et al.** avaient appliqués la Régression Logistique Conditionnelle pour améliorer leur résultat. Les auteurs ont été satisfait de leurs résultats car la Régression Logistique Conditionnelle (CLR) donne de meilleurs taux de précision comparativement aux autres modèles appliqués. confère **figure 6**.

RÉFÉRENCES

[1] K. SHAMEER, K. W. JOHNSON, A. YAHI, R. MIOTTO et. al. ***Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-wide Machine Learning : a case-study using Mount Sinai heart failure cohort*** . Pacific Symposium on Biocomputing 2017. Page 1-10.

[2] Reena Duggal, Suren Shukla, Sarika Chandra, Balvinder Shukla, Sunil Kumar Khatri. ***Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India*** . International Journal of Diabetes Developing Countries (October–December 2016). Page 95-102.

[3] K. Zhu ; Z. Lou ; J. Zhou ; N. Ballester ; N. Kong ; P. Parikh. ***Predicting 30-day Hospital Readmission with Publicly Available Administrative Database*** . A Conditional Logistic Regression Modeling Approach. Schattauer 2015. Page 1-8.