N° d'ordre NNT : 2023LYO20010

THÈSE de DOCTORAT DE L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512 Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 16 mars 2023, par :

Lijuan REN

A Lifestyle Related Disease Prediction Framework Based on Missing Value Imputation and Stacking Ensemble Method.

Devant le jury composé de :

Hervé PINGAUD, Professeur des Universités, Institut National Universitaire Champollion, Président
Lina SOUALMIA, Professeure des Universités, Université Rouen Normandie, Rapporteure
Keshav DAHAL, Professeur des Universités, University of the West of Scotland, Rapporteur
Haiqing ZHANG, Professeure, Chengdu University of Information Techno, Examinatrice
Tewfik ZIADI, Maître de conférences HDR, Sorbonne Université, Examinateur
Tao WANG, Maître de conférences HDR, Université Jean Monnet Saint-Étienne, Examinateur
Aïcha SEKHARI, Maîtresse de conférences, Université Lumière Lyon 2, Examinatrice
Abdelaziz BOURAS, Professeur des Universités, Université Lumière Lyon 2, Directeur de thèse

Contrat de diffusion

Autre résumé

Les maladies liées au mode de vie (MLT) désignent les maladies dont la physiopathologie est fortement influencée par des facteurs liés au mode de vie, et la modification de ces facteurs étiologiques peut améliorer de manière significative la prévention et le traitement des maladies. Le concept de maladies liées au mode de vie provient principalement de deux aspects. 1) À mesure que les pays s'industrialisent et s'enrichissent, la fréquence des maladies augmente en raison des changements de comportement des gens.2) Elles sont extrêmement liées aux modes de vie ou aux comportements des gens. En général, la plupart des maladies chroniques, y compris les maladies cardiovasculaires, le syndrome métabolique, l'obésité, le diabète de type 2 et certains cancers, sont des maladies liées au mode de vie et sont étroitement liées au mode de vie des gens. Des études ont montré que les maladies liées au mode de vie sont les maladies les plus répandues dans le monde aujourd'hui, en termes absolus et relatifs, et que le nombre de décès dépasse celui du sida, du paludisme et de la tuberculose réunis. Les maladies cardiovasculaires, l'obésité, le diabète de type 2, l'hypertension et certaines tumeurs malignes particulières sont tous devenus des problèmes importants au XXIe siècle. En République d'Irlande, 61 % des adultes sont en surpoids ou obèses, et plus de 40 % des adultes déclarent souffrir d'au moins une affection liée au mode de vie, les plus répandues étant l'hypertension artérielle et l'hypercholestérolémie. En outre, 17,8 millions d'individus dans le monde sont décédés de maladies cardiovasculaires (MCV) en 2017, selon le rapport sur la charge mondiale de morbidité publié en 2018, et le nombre global estimé de décès liés aux tumeurs (principalement le cancer) est de 9,56 millions. L'OMS prévoit que d'ici 2030, 366 millions d'individus dans le monde seront atteints de diabète, contre 175 millions actuellement. Malgré la disponibilité d'une large gamme de médicaments, la fréquence des maladies liées au mode de vie n'est pas maîtrisée en raison des problèmes de sécurité liés à ces médicaments. En résumé, le système de santé mondial est en crise en raison de la prévalence de ces troubles liés au mode de vie.

Ces maladies ont un début sournois, une période d'incubation prolongée et une progression rapide. Il est difficile d'identifier et de traiter de nombreux patients à temps. De plus, comme la majorité des maladies liées au mode de vie ont encore des étiologies et une pathogénie peu claires et des résultats thérapeutiques médiocres, il est important, d'un point de vue pratique, de prévenir le développement des maladies liées au mode de vie. Compte tenu des caractéristiques des maladies liées au mode de vie et des tendances contemporaines en matière de santé, la prédiction précoce des maladies a des ramifications importantes pour la recherche. Il s'agit de l'une des étapes clés de la prévention et du traitement des maladies causées par le mode de vie d'une personne, car l'identification des risques pour la population avant l'apparition des maladies peut aider les gens à modifier leur mode de vie le plus tôt possible, en particulier les comportements de vie des groupes à haut risque, réduisant ainsi le risque de maladie. Le principal outil d'évaluation et de prévention des maladies liées au mode de vie est le modèle de prédiction des maladies . Un modèle de prédiction des maladies établit spécifiquement un modèle intelligent pour prédire la probabilité d'une maladie spécifique à un moment précis dans l'avenir, classe les groupes à haut risque en fonction du seuil de probabilité, et mène des interventions sur le comportement, le régime alimentaire et autres pour prévenir les maladies futures. Elle peut être classée dans la catégorie de la prévention des maladies. En d'autres termes, le modèle de prédiction des maladies peut montrer aux sujets de l'évaluation la probabilité qu'ils tombent malades à l'avenir et anticiper cette probabilité, ainsi que les conseiller

sur la manière de gérer leur propre santé.

Cependant, en raison de l'incomplétude des données obtenues et du bruit important provenant des valeurs manquantes et des facteurs incertains liés au mode de vie, les cadres actuels de prédiction des maladies sont souvent insuffisants pour prédire de manière robuste le risque de maladies liées au mode de vie. Les deux principaux objectifs de cette étude sont d'établir une méthode améliorée pour traiter les valeurs manquantes dans les ensembles de données déséquilibrées et mixtes, et de créer un modèle d'ensemble fiable pour prédire l'apparition potentielle de maladies liées au mode de vie. L'évaluation et la prévention des LRD sont basées sur ces deux objectifs. Le premier objectif est de traiter l'incomplétude des données réelles, et le second objectif est d'améliorer la capacité de généralisation des modèles de prédiction des maladies sur des ensembles de données bruitées. Par rapport aux modèles traditionnels de prédiction des maladies liées au mode de vie, cette étude repose sur une méthode complète d'imputation des valeurs manquantes et une approche d'ensemble robuste, ce qui rend le cadre de prédiction des maladies proposé plus puissant.

Spécifiquement, en raison du bruit important dans les ensembles de données et des valeurs manquantes, il est difficile d'utiliser les méthodes classiques d'apprentissage automatique pour construire des modèles de prédiction des LRD à partir de données médicales. En particulier, certaines causes inéluctables, notamment le retrait précoce des sujets de la recherche médicale, peuvent rapidement entraîner des valeurs manquantes dans les données de recherche. De nombreuses approches pour faire face aux valeurs manquantes ont été proposées, car la présence de valeurs manquantes rend plus difficile l'extraction de données pertinentes. Les ensembles de données à grande échelle avec des types mixtes et des caractéristiques non équilibrées sont néanmoins courants dans l'industrie médicale. Seules quelques approches peuvent être utilisées pour les données de types mixtes et de caractéristiques non équilibrées en même temps, malgré le fait que les méthodes de pointe existantes peuvent diminuer les erreurs d'imputation et augmenter la qualité des données manquantes. Pour y parvenir, nous proposons une nouvelle technique d'interpolation des valeurs manquantes basée sur l'Adaptive Laplacian Weighted Random Forest (ALWRF) et la technologie de suréchantillonnage SMOTE-NC. Cette méthode peut améliorer les caractéristiques de précision de la prédiction non équilibrée en ajustant de manière adaptative les poids des caractéristiques lors de la construction des forêts aléatoires.

En outre, la robustesse de l'algorithme sera affectée par la présence de bruit. Cependant, comme le bruit est fréquemment présent dans les données médicales, de nombreuses études se sont concentrées sur la façon de le traiter. Étant donné que certains des ensembles de données des troubles liés au mode de vie analysés correspondent à des patients réels, il est difficile en pratique de supprimer directement les valeurs aberrantes. La combinaison de méthodes d'ensemble avec des techniques au niveau des algorithmes est une excellente stratégie pour minimiser la variance, le biais et le bruit. Les performances de chaque modèle de l'ensemble varient en fonction des circonstances. Par conséquent, une approche d'ensemble a été utilisée dans notre travail pour réduire le bruit des données et augmenter la précision de la prédiction des maladies liées au mode de vie. Nous proposons une technique de sélection de modèles itérative multi-objectifs (MoItMS) pour maximiser la variété et la précision des modèles d'ensemble en même temps. Le cadre d'intégration multi-objectif proposé, basé sur l'empilement, peut offrir des méthodologies utiles axées sur les

données pour catégoriser les patients en vue de la gestion de la santé de la population, promouvoir le contrôle des maladies et soutenir la détection des maladies liées au mode de vie lorsqu'il est appliqué à de grands ensembles de données cliniques.

Enfin, nous utilisons un cas de la Chine pour appliquer le cadre de prédiction proposé. Deux modèles importants - les modèles d'imputation des valeurs manquantes et les modèles de prédiction des maladies - sont produits après traitement par les trois modules primaires de valeurs manquantes, de sélection des caractéristiques et de prédiction des maladies. Le cadre de prédiction proposé peut également améliorer les performances de prédiction des LRD pour une meilleure prévention de la santé publique, selon les résultats expérimentaux.