الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 08 ماي 1945 بسيدي بلعباس
École Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbès

THESIS

To obtain the diploma of **Master**
Field: **Computer Science**
Specialty: **Artificial Intelligence and data Science**

Theme

Adressing Small and Incomplete Clinical Datasets: Data Imputation and Synthetic Data Generation

---

Presented by:
BOUCHOUKA Zohra Hidaya

Submission Date: **Sept, 2024**
In front of the jury composed of:

Mr. KHALDI Belkacem	President
Ms. DIF Nassima	Examiner
Ms. Julia FLECK	Supervisor
Mr. Sidi Mohamed BENSLIMANE	Co-Supervisor
Mr. Xie XIAOLAN	Co-Supervisor

Academic Year : 2023/2024

Abstract

In the age of data, healthcare has become a vital field where enormous amounts of information are collected, analyzed, and used to improve patient outcomes. Yet, the practical application of big data in healthcare faces significant challenges, especially when it comes to tabular data.

Medical datasets often suffer from challenges such as small sample sizes and missing values, which hinder the application of traditional machine learning models that typically require large, complete datasets for accurate predictions. This research addresses these challenges, focusing on the development and application of innovative data imputation and augmentation techniques to enhance the predictive reliability of healthcare models.

The study aims to identify and evaluate state-of-the-art approaches for handling missing data specially for small datasets to improve model performance. By applying these techniques to healthcare scenarios, the research seeks to improve predictive accuracy and support better clinical decision-making, emphasizing the importance of identifying critical variables influencing patient outcomes.

While his thesis is focused on research, with no practical implementation undertaken at this stage, eventually the findings are expected to significantly enhance clinical decision-making, showcasing the potential of advanced data processing techniques to overcome common challenges in healthcare data analytics.

Keywords — Tabular Data, Missing Data, Data Processing, Small Datasets, Data Imputation, Machine Learning, Healthcare, Classification Model, Supervised Learning.

ملخص

في عصر البيانات، أصبح مجال الرعاية الصحية مجالاً حيوياً يتم فيه جمع وتحليل كميات هائلة من المعلومات لتحسين نتائج المرضى. ومع ذلك، تواجه التطبيقات العملية لعلوم البيانات في الرعاية الصحية تحديات كبيرة، خاصةً فيما يتعلق بالبيانات الجدولية.

تعاني مجموعات البيانات الطبية غالبًا من تحديات مثل أحجام العينات الصغيرة والقيم المفقودة، مما يعيق تطبيق نماذج التعلم الآلي التقليدية التي تتطلب عادةً مجموعات بيانات كبيرة وكاملة للحصول على تنبؤات دقيقة. تتناول هذه الدراسة هذه التحديات، مع التركيز على التقنيات المبتكرة لاستكمال البيانات وتوليد البيانات الاصطناعية لتعزيز موثوقية التنبؤات في نماذج الرعاية الصحية، خاصةً في مجالات مثل أداء نقل الدم والتشخيصات.

تهدف الدراسة إلى تحديد وتقييم أحدث الأساليب للتعامل مع البيانات المفقودة، خصوصًا في مجموعات البيانات الصغيرة، لتحسين أداء النماذج. من خلال تطبيق هذه التقنيات في سيناريوهات الرعاية الصحية، تسعى الدراسة إلى تحسين دقة التنبؤات ودعم اتخاذ قرارات سريرية أفضل، مع التركيز على أهمية تحديد المتغيرات الحاسمة التي تؤثر على نتائج المرضى.

ورغم أن هذه الأطروحة تركز على البحث، دون تنفيذ عملي في هذه المرحلة، فمن المتوقع أن تسهم النتائج في تحسين اتخاذ القرارات السريرية في مجال الصحة، مما يبرز إمكانات تقنيات معالجة البيانات المتقدمة في التغلب على التحديات الشائعة في تحليلات بيانات الرعاية الصحية.

الكلمات المفتاحية: لبيانات الجدولية، البيانات المفقودة، معالجة البيانات، مجموعات البيانات الصغيرة تعويض البيانات، تعلم الآلة، الرعاية الصحية، نموذج التصنيف، التعلم الموجه,

Résumé

À l'ère des données, le secteur de la santé est devenu un domaine crucial où d'énormes quantités d'informations sont collectées, analysées et utilisées pour faciliter les tâches pour les experts et améliorer les résultats pour les patients. Cependant, l'application pratique des sciences des données dans le secteur de la santé rencontre des défis importants, notamment lorsqu'il s'agit de données tabulaires.

Les ensembles de données médicales souffrent souvent de défis tels que les petites tailles d'échantillons et les valeurs manquantes, ce qui entrave l'application des modèles d'apprentissage automatique traditionnels qui nécessitent généralement des ensembles de données volumineux et complets pour des prédictions précises. Cette recherche aborde ces défis en se concentrant sur le développement et l'application de techniques innovantes d'imputation et d'augmentation des données pour améliorer la fiabilité prédictive des modèles de santé, en particulier dans des domaines tels que la performance des transfusions sanguines et le diagnostic.

L'étude vise à identifier et évaluer les approches de pointe pour gérer les données manquantes, et les petites base de données, afin d'améliorer la performance des modèles. En appliquant ces techniques aux scénarios de santé, la recherche cherche à améliorer la précision des prédictions et à soutenir une meilleure prise de décision clinique, en mettant l'accent sur l'importance d'identifier les variables critiques influençant les résultats des patients.

Bien que cette thèse soit axée sur la recherche, sans mise en œuvre pratique à ce stade, les résultats devraient finalement améliorer de manière significative la prise de décision clinique, démontrant le potentiel des techniques avancées de traitement des données pour surmonter les défis courants dans l'analyse des données de santé.

Mots-clés — Données Tabulaires, Valeurs Manquantes, Traitement des Données, Petits Ensembles de Données, Imputation des Données, Apprentissage Automatique, Santé, Modèle de Classification, Apprentissage Supervisé.