Prediction for potential medical treatment pathways with outpatient data

Jingyi Lai, Jeffrey Jin, Xuan Tao
UMSI MADS

Introduction

- Patients often experience unease when they must visit the hospital outpatient department, as they may lack clarity about their illness, the required examinations, and the associated expenses.
- Machine learning with clinical data has long been in practice, offering substantial benefits.
- While numerous studies have explored diagnostic and support pathways for specific ailments like depression and diabetes, the application of these technologies in creating interactive self-check platforms remains less explored.
- Our project aims to:
  a. Could analyzing medical encounter data reveal distinct patient groups with similar healthcare needs?
  b. Could classification prediction on user-reported symptoms and health status enable the identification of potential treatment pathways in a statistical representation?

Method

Dataset

- Outpatient dataset of National Hospital Ambulatory Medical Care Survey (NHAMCS) with ICD-9-CM

Preprocessing

- Merged: by consolidating relevant columns, resulting in a dataset with 150 features.
- Splitted into training (2006-2008), validation (2009), test (2010), and evaluation (2011) sets to maintain generalizability and adaptability to data shift

Feature Engineering

- Quantitative features: Binning or normalization based on stability and sensitivity requirements.
- Categorical features: One-hot encoding
- Text features: text extracted, integrated and transformed into topic features

Machine Learning

- Employed Random Forest Classifier to predict the classification of Diseases and Injuries

Results

Features Engineering

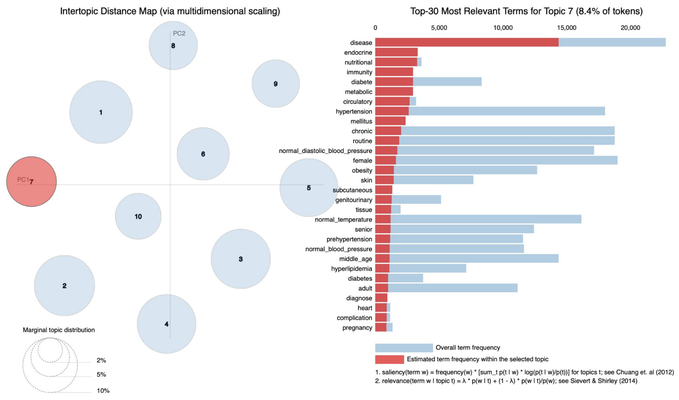


Figure 1: Example of Word Cloud

- Figure 1 showed 10 distinct topics with varying degrees of intertopic distance, indicating differentiated patient groupings based on their medical descriptors.
- The term relevance for Topic 1 reveals a strong emphasis on patient management programs, as indicated by high-frequency terms such as 'hypertension' and 'diabete'.

Classifier

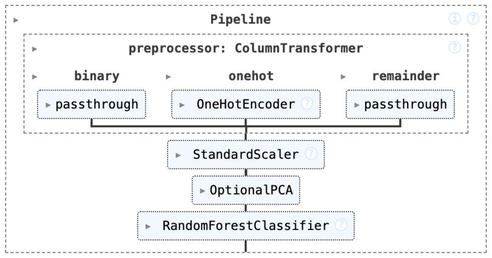


Figure 2: Pipeline of the final best model – Random Forest Model

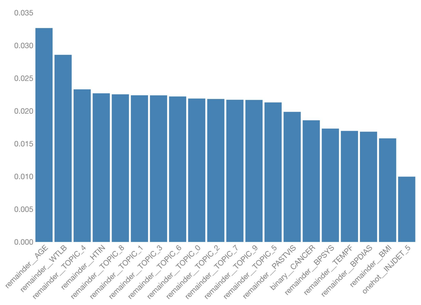


Figure 3: Topic 20 importance features

The prominence of these features highlights the intricate interplay between demographic characteristics, vital signs, and thematic elements of patient data in our final model.

Conclusion

- Our machine learning analysis of NHAMCS outpatient data reveals key diagnostic pathways and supports, potentially improving patient understanding before medical consultations.
- Despite time and data limitations, our promising results endorse the application of machine learning in healthcare, setting a foundation for future interactive self-check platforms.

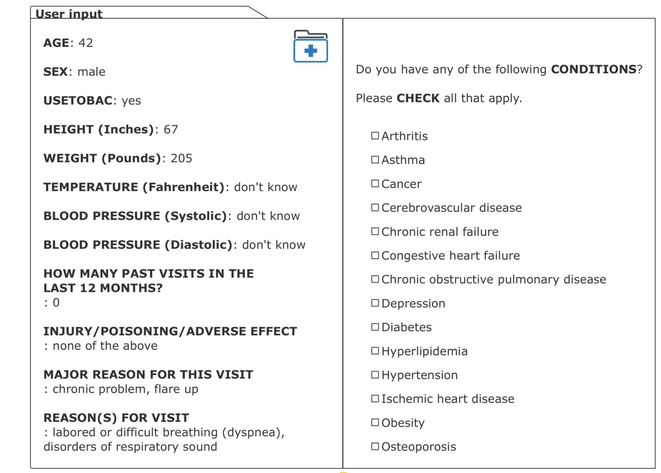
Classification approach in NHAMCS outpatient data reveals key diagnostic pathways and supports

Scan me for Github

Reference

Hing E, Middleton K. National Hospital Ambulatory Medical Care Survey: 2001 outpatient department summary. Adv Data. 2003 Aug 5;(338):1-26. PMID: 12918175.
Czakon, J. (2023, September 5). F1 score vs ROC AUC vs Accuracy Vs PR AUC: Which evaluation metric should you choose?. neptune.ai. https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

Mock up health self-check interactive simulation



Top 3 highest possibility diseases