Hospital Mortality Reduction: Machine Learning Approaches

CSML1030: Machine Learning Capstone

Project Proposal for David Rotenberg, The Centre for Addiction and Mental Health (CAMH)

April 4, 2020

Pete Gray, Pratik Chandwani, Conrard Giresse Tetsassi Feugmo

This work aims to discover possible approaches to improving hospital care by using machine learning methodologies on both demographic and clinical patient data. A step-by-step analysis will be conducted, and factors that affect deaths or suicide attempts among patients will be identified.

1 Project Background

More than 165,000 Canadians die in hospitals every year [1]. This is approximately 60% of all deaths in Canada. There are wide variations in hospital mortality. Much of this variation remains unexplained and may reflect quality of care. Among the causes of death in Canadian hospitals, the top six are [2]:

- Stroke
- Chronic obstructive pulmonary disease (COPD)
- Sepsis
- Heart attack
- Pneumonia
- Heart failure

According to the Canadian Mental Health Association, there is a fundamental link between mental health and physical health [3]. People with serious mental illness are at higher risk of chronic physical conditions, and those suffering from chronic physical conditions are at higher risk of anxiety and depression.

Frequent visits to emergency rooms may indicate that people are not getting access to the services or the support they need in the community for help with mental health and/or addictions. According to the Canadian Institute for Health information, many people visit an emergency room at least 4 times a year [4]. Among these people 50% are under age 35 (including youth and children under age 20), 32% have poor mental health conditions (anxiety, mood disorder) and 18% have addiction to substances (usually alcohol).

The death rate at hospitals constitutes an indicator of health care quality. By analyzing clinical and demographic data, patterns that are correlated with a higher risk can be discovered. An understanding of these patterns could be applied to reducing patient deaths and improving the quality of care.

2 Problem Statement

The application of machine learning methodologies to healthcare has yielded many great results. Machine learning has gone from the realm of a relatively small number of data scientists to the

mainstream of analysis and business. Today, machine learning is helping to streamline administrative processes in hospitals, map and treat infectious diseases, improve diagnostic precision and personalize medical treatments. Our study aims to apply machine learning approaches to infer knowledge from both clinical and demographic variables. This knowledge could be leveraged in practice by incorporating trained machine learning models into current or future diagnostic applications.

2.1 Special Considerations

Circumstances related to the COVID-19 pandemic prevent CAMH from sharing the clinical data that was to be used for this project at this time. In order to complete this research, and still produce results that can provide valuable insights for decision making, datasets that have been approved or provided by CAMH based on relevance to real-world problems will be used instead. In lieu of producing a model that can provide enhanced diagnostic accuracy to existing clinical tools, exploratory research will be conducted on the alternative datasets. The intention is to identify features and patterns in data, as well as ML methodologies, that can readily be applied to research on CAMH's actual clinical data.

3 Data description

Two datasets have been chosen based on their relevance to CAMH's research on patient mortality.

3.1 Global Women in Data Science (WiDS) Datathon 2020

The WiDS Datathon 2020 focuses on patient health through data from MIT's GOSSIS (Global Open Source Severity of Illness Score) initiative. The challenge is to create a model that uses data from the first 24 hours of intensive care to predict patient survival. This data is part of a growing global effort and consortium spanning Argentina, Australia, New Zealand, Sri Lanka, Brazil, and more than 200 hospitals in the United States. Labeled training data are provided for model development, as well as unlabeled data for predictions. Kaggle will use these predictions to determine the public leaderboard rankings. The dataset has 91,713 entries with 189 attributes. These attributes include:

Ethnicity	Heart rate
Gender	Hemoglobin concentration
Age	Lactate
B	Whether the patient has a Cirrhosis history
Weight	Whether the patient has been diagnosed Diabetes
Height	Whether the patient has been diagnosed Hepatic
Glucose concentration	Creatinine concentration
Temperature measured	Body system

3.2 Centre for Addiction and Mental Health (CAMH)

CAMH is a psychiatric teaching hospital located in Toronto, plus ten community locations throughout the province of Ontario. CAMH provides a wide range of clinical care services for patients of all ages and families. The dataset is unlabeled and has 722,611 entries with 49 attributes. These attributes include:

Death size	Violent death

Cause of death	Work
Gender	Necropsy
Death date	Medication
Birth date	Nationality
Employ	Insurance
Education	Language
Marital status	Relative

4 Methodology

In order to provide CAMH with valuable insights, while using alternative datasets, the methodologies that will be used will prioritize explainability of results and interpretability of models, over the ability of those models to produce accurate inferences using clinical data. The intention is that the knowledge discovered can be applied to real-world problems when clinical data is available.

4.1 Data Exploration

In order to understand the data, descriptive statistics and data visualization will be performed.

4.2 Supervised Approaches: Classification

The MIT's GOSSIS dataset is labeled, allowing us to use supervised learning approaches to improve patient survival. The target variable is 'hospital_death', which indicates whether the patient died in the hospital in the first 24 hours or not.

4.2.1 Feature Engineering

Feature selection, the process of finding and selecting the most useful features in a dataset.
There is always a deal when working in the field of heatlcare 0.5 % or error could lead to the death of a patient. We certainly need accuracy but we also need interpretability to help decision making. Unnecessary features decrease training speed, decrease model interpretability, and, most importantly, decrease generalization performance on the test set. For this study we used the following approaches:

- Features with a high percentage of missing values (more than 75%)
- Collinear (highly correlated > 0.9) features
- Data Imputation (Interative_Imputer, KNN_Imputer)
- Features with zero importance in a tree-based model
- Features contributing to 95% of the cumulative importance

4.2.2 Model Selection

We are dealing with a binary classification problem. In addition we are not only interested in the accuracy but we are also looking for a model interpretable as possible. For this two reason and considering the fact there is three persons in the group we will train six model (2 per person)

- *Logistic regression (LR)*
- *Naive Bayes (NB)*
- *Support Vector Machines (SVM)*
- *K-nearest neighbor algorithm (KNN)*
- *Random Forest Classifier (RFC)*

- *Gradient boosting Classifier (GBC)*

LR and **NB** are commonly used for binary classification there linear model and simple for interpretation. **SVM** and **KNN** can be used to train linear-model problems. Finally, **RFC** *,* and **GBC** are ensemble methods, useful to reduce bias. In addition, plotting a decision tree (RFC) really improves decision making. These models will be trained and evaluated and compared based on their accuracy, precision, and recall.

4.2.3 Tuning the Hyperparameters

The models with the best performance will be improved further through hyperparameter tuning.

4.2.4 Testing and Deployment

Finally, selected and tuned models will be tested on unlabeled data and the result will be submitted on Kaggle to evaluate the real performance of our models.

4.3 Unsupervised Approaches

The datasets provided by CAMH are unlabeled. Unsupervised Learning will be used to find descriptors and patterns in the data.

4.3.1 Clustering

We will use clustering algorithms to find clusters and their classes. According to the data exploration the data has 722611 entries and categorical features (especially death_cause) have more than 4000 unique variables. Due to this large size, we will likely use Label Encoding rather than One-Hot Encoding. We will use the following clustering algorithms:

- Affinity Propagation
- Hierarchical Cluster Analysis (Agglomerative Clustering)
- KMeansClustering

4.3.2 Visualization

Visualization is the process of creating diagrams, images, graphs, charts, etc., to effectively communicate information derived from raw data or engineered features. One example of a visualization algorithm is t-distributed Stochastic Neighbor Embedding (t-SNE).

4.3.3 Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration by getting a set of principal variables.

- Principal Component Analysis (PCA)
- Locally-Linear Embedding

4.3.4 Finding Association Rules

This is the process of finding associations between different parameters in the available data. It discovers the probability of the co-occurrence of items in a collection, such as people that buy X also tend to buy Y. These are some of the commonly used algorithms for association rule learning:

- Apriori
- Eclat

4.3.5 Anomaly Detection

Anomaly detection is the identification of rare items, events or observations which brings suspicions by differing significantly from the normal data. In this case, the system is trained with a lot of normal instances. So, when it sees an unusual instance, it can detect whether it is an anomaly or not.

4.4 Reinforcement Learning

Reinforcement learning (RL) aims at empowering one's capabilities in behavioural decision making by using interaction experience with the world and an evaluative feedback. RL helps with solving sequential decision making problems [5]

Relevant descriptors will be optimized with reinforcement learning to maximize survival rate.

Summary

This project will use machine learning methodologies to identify possible approaches to reducing mortality in hospital patients. Supervised learning, unsupervised learning, and reinforcement learning will be used to identify features and patterns in data that are correlated with patient mortality. The datasets have been chosen in collaboration with CAMH for their relevance to real-world problems. Explainability of results and interpretability of models will be given exceptional priority over model accuracy, allowing the results to be leveraged into further machine learning research using actual clinical data.

References

[1] Statistics Canada. (2018). table 13-10-0715-01 deaths, by place of death (hospital or non-hospital). https://doi.org/10.25318/1310071501-eng

[2] Hospital Deaths (HSMR).
https://yourhealthsystem.cihi.ca/hsp/inbrief?lang=en#!/indicators/005/hospital-deaths-hsmr/;mapC1;mapLevel2;/

[3] The Relationship between Mental Health, Mental Illness and Chronic Physical Conditions.
https://ontario.cmha.ca/documents/the-relationship-between-mental-health-mental-illness-and-chronic-physical-conditions/

[4] Frequent Emergency Room Visits for Help With Mental Health and/or Addictions.
https://yourhealthsystem.cihi.ca/hsp/inbrief;lang=en?lang=en#!/indicators/078/frequent-emergency-room-visits-for-help-with-mental-health-and-or-addictions/;mapC1;mapLevel2;/

[5] A review of recent reinforcement learning applications to healthcare.
https://towardsdatascience.com/a-review-of-recent-reinforcment-learning-applications-to-healthcare-1f8357600407