Identification of Suicide-Related Visits in the Electronic Health Record

Final Project Report
Sarah Arias
https://github.com/sting0131/Synthea_selfharm

1. Introduction

Background/Significance

With the goal of reducing suicide rates by 20% in the next few years, there is an immediate need for research that can inform suicide intervention development. Considering that over 75% of individuals receive some form of healthcare prior to suicide, data in the electronic health record (EHR) can be very informative for suicide detection and prevention efforts. However, leveraging these data for clinical research can be challenging. In addition, the current practice for studying suicide risk involves the use of ICD-9/10-CM codes, which have been found to significantly underestimate suicidal ideation and suicide attempt cases in the EHR. Although there have been numerous analyses of suicide outcomes within EHRs, there is limited information on the critical initial step of developing effective methods for reliable *identification* of suicide-related cases documented at the point of care.

Synthea

Synthea, is an open-source, synthetic patient generator that models the health history of synthetic patients. The approach used by Synthea guarantees fully synthetic output by accepting only publicly available information and health statistics as inputs. Data are generated based on models of clinical workflow and disease progression that can be easily inspected, modified, and refined, facilitating transparency and continuous improvement. The program includes a temporal model that covers a patient's entire lifetime instead of focusing on one health problem or disease.

Target Variables and Features

The identification of healthcare visits as suicide or non-suicide related is a classification problem. The target variable in this case is documentation of suicidal thoughts or behavior.

Current Research

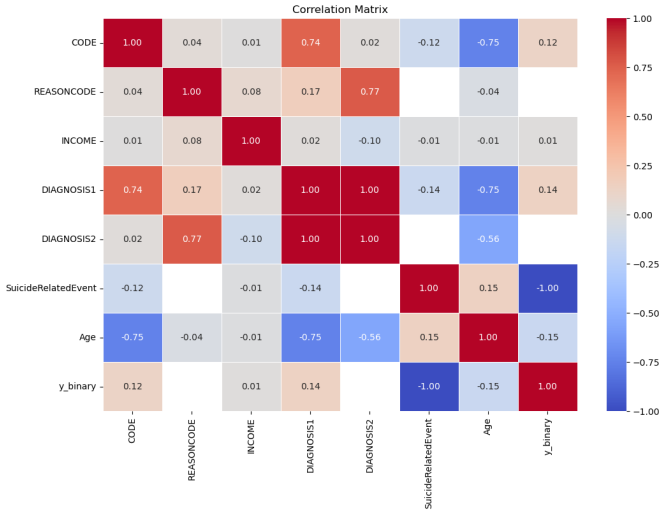
The current study capitalizes on the synthetic healthcare data that can be generated by Synthea to train and test several machine learning (ML) models for classifying healthcare visits as suicide or non-suicide related.

2. Exploratory Data Analysis

Feature Analysis

Case identification for suicide in the electronic health record (EHR) requires consideration of features available within the patient record. As an initial step for classifying visits as either suicide or non-suicide related, Synthea generated records with reason for visit, diagnosis code, and demographic information. The .describe function was used to identify missing variables within the dataset. Bar charts and boxplots were used to visualize the distributions and means of the various groups. A correlation matrix was run to examine relationships between the features. Further consideration of the timing of diagnosis information during clinical workflow led to the

removal of the diagnosis variables from the model (i.e., the clinician would not have access to this data at the time of visit).



Target Variable

The target variable for the current study is presence of suicidal behavior documented during the healthcare visit. To get a clearer picture of the cases within the current dataset, a bar chart was created to visualize the type and frequency of suicide behaviors (Fig.1). Similar to the literature[1], the largest proportion of cases fell into the "deliberate poisoning" category (n=227).

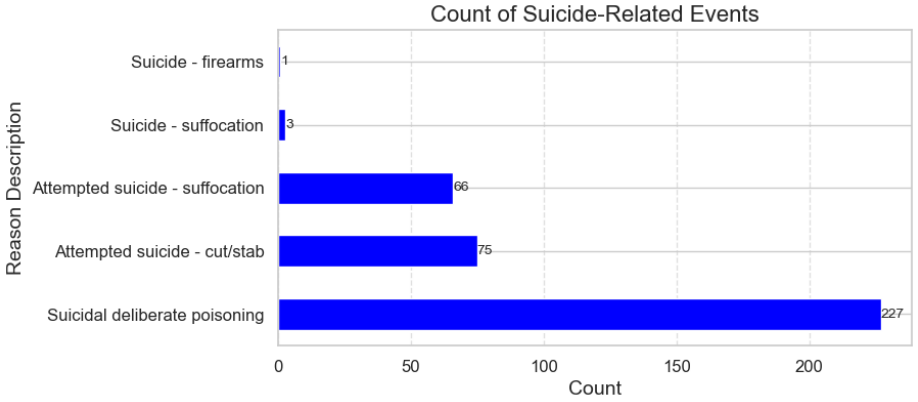


Fig.1. Frequencies of categories of suicide-related events detected within the dataset.

An examination of suicide-related behaviors by patient gender indicated a higher proportion of females within the current dataset (Fig. 2.). This is in line with the literature whereby females are more likely to seek medical treatment for suicide-related issues.[2]

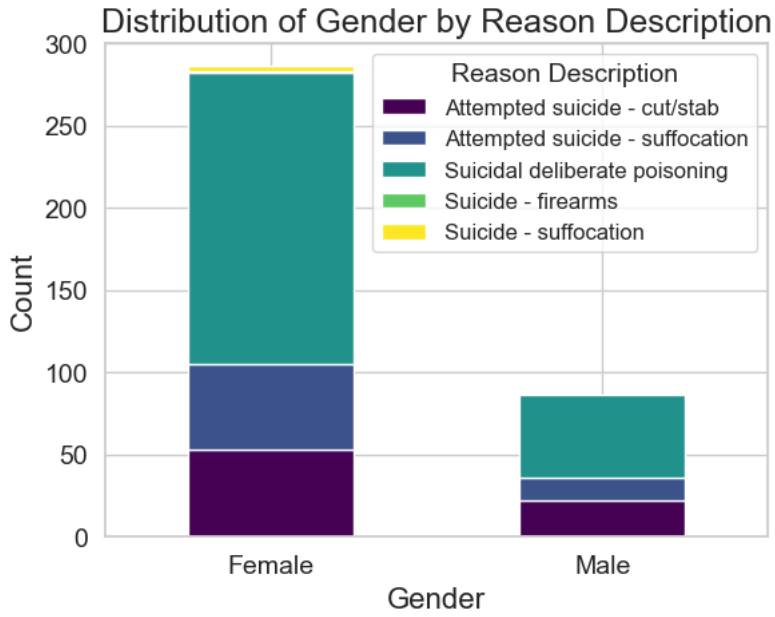


Fig.2. Distribution of suicide-related events by patient gender (female/male).

An additional consideration is the age of the sample. In this case, the mean age is 39.28 years. When looking at the suicide-related cases, individuals with suicide-related visits tend to fall within the older age category (Fig.3.). There is proportionately higher reporting of suicidal thoughts and behaviors in middle-aged individuals[3], so this is not entirely surprising.

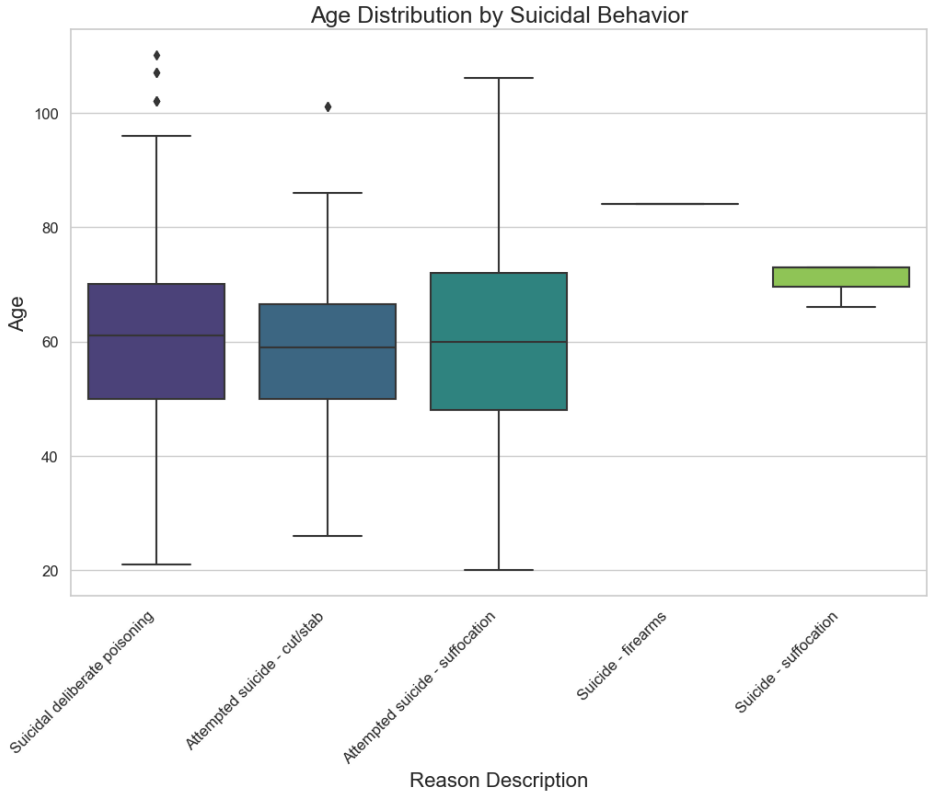


Fig.3. Boxplot of suicide-related events by patient ages (in years).

3. Methods

Splitting the Data

Due to the non-IID and imbalanced nature of the data, sklearn's GroupShuffleSplit was used to split the data. This was chosen because it allows for specification of the groups, so we can be sure that each group appears in only one split. It also helps ensure that each split maintains the class distribution that was in the original dataset. This is particularly relevant for the current project focuses on records that may contain the same patient information (i.e., patient returned for multiple visits and has multiple records in the EHR).

Preprocessor

There features in this dataset include numerical (age, income, code (visit), reasoncode), categorical (race, ethnicity, gender, description (visit), encounter class (visit), reasondescription (visit)), and ordinal (income_range). StandardScaler was used for the numerical features, OneHotEncoder for categorical, and OrdinalEncoder for ordinal.

Features

There were initially 12 features in the dataset, but after preprocessing, there were 32 features.

ML Pipeline

Four machine learning algorithms were tested for this analysis: Logistic Regression, Random Forest Classifier, SVC, and XGBoost. See **Table 1** for the parameters tuned for each algorithm.

Logistic Regression	C: [0.0001, 0.01, 0.1, 1, 10, 100, 1000] Penalty: [l2]
Random Forest Classifier	n_estimators: [50, 100, 150] max_depth: [None, 10, 20, 30] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4]
SVC	C: [0.1, 1, 10, 100] kernel: ['linear', 'rbf', 'poly', 'sigmoid']
XGBoost	max_depth: [3, 5, 7] learning_rate: [0.01, 0.1, 0.2] n_estimators: [50, 100, 200]

Table 1. Tuned parameter values for each machine learning algorithm

Evaluation metrics selected included f1_score, area under the receiver operating curve (AUC-ROC), precision, and recall. These were selected over accuracy as this dataset's target variable has one class that significantly outnumbers the other. Specifically, approximately 3.7% of the cases fall in Class 1.

During the splitting process, each model's algorithms are assigned five different random state values. For Logistic Regression, Random Forest Classifier, and SVC, a 10 k-fold method was applied to the models. The best model and test score are saved for each of the algorithms.

4. Results

Prior to testing the models, baseline accuracy was calculated on the test set. The overall accuracy was 0.96. As accuracy does not sufficiently account for imbalance in the dataset, f1 score (0.00), auc-roc (0.50), precision (1.00), and recall (0.00) were also calculated. Adjustments were made to ensure that the majority class was defined and that Class 1 cases were being included for the calculations.

Even after removing variables that were unlikely to be available at the time of visit (e.g., current diagnosis), the models still showed perfect evaluation metric scores (1.00).

Algorithm	Mean F1	Mean ROC-AUC	Mean Precision	Mean Recall
Random Forest	1.00	1.00	1.00	1.00
Logistic Regression	1.00	1.00	1.00	1.00
SVC	1.00	1.00	1.00	1.00
XGBoost	1.00	1.00	1.00	1.00

Table 2. Evaluation metrics for performance of the machine learning models

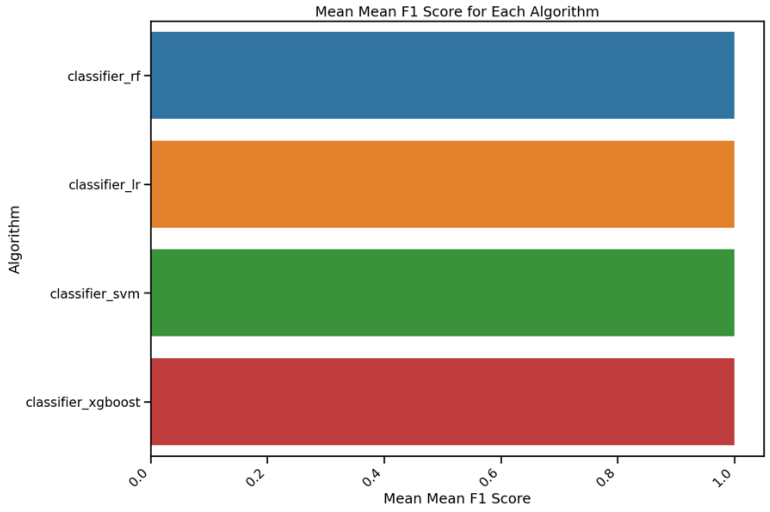


Fig. 4. Mean F1 scores for Random Forest (rf), Logistic Regression (lr), SVC, and XGBoost

SHAP values were calculated to examine the local feature importance for the models. The top five values based on mean SHAP values included income range, presentation for a wellness visit, encounter for a problem, being in the ambulatory encounter class, and the reason for visit code.

It is surprising that income level played an important role in the model performance, but this may be linked to the research findings that individuals who are experiencing financial crises or are from a lower income bracket, tend to be at higher risk for suicide.[4]

5. Outlook

Although the models seemingly performed perfectly in classifying suicide and non-suicide related healthcare visits, it is clear from the additional evaluation metrics (e.g., f1, recall) that the overall performance of the model needs to be further assessed. Future work could benefit from engaging in additional feature engineering. It would be beneficial to return to the synthetic datafile and see if there are any history variables (e.g., history of self-harm, mental health history) that could be used to inform the models.

It is also possible that a combination of machine learning approaches could be better suited for classifying these types of cases within the EHR. In addition, consideration and testing of different hyperparameters may help fine tune the overall model fit and performance.

6. References

1. Kalankesh LR, Farahbakhsh M, Fein RA, Moftian N, Nasiry Z. Exploring Complexity of Deliberate Self-Poisoning through Network Analysis. Psychiatry J. 2017;2017:3619721. doi: 10.1155/2017/3619721. Epub 2017 Jan 29. PMID: 28251146; PMCID: PMC5303583.
2. Ivey-Stephenson AZ, Crosby AE, Hoenig JM, Gyawali S, Park-Lee E, Hedden SL. Suicidal thoughts and behaviors among adults aged ≥18 years – United States, 2015-2019. MMWR Surveill Summ 2022;71(No. SS-1):1-19.
3. Stoliker, B.E., Verdun-Jones, S.N. & Vaughan, A.D. The relationship between age and suicidal thoughts and attempted suicide among prisoners. *Health Justice* **8**, 14 (2020). https://doi.org/10.1186/s40352-020-00117-3.
4. Eric B Elbogen, Megan Lanier, Ann Elizabeth Montgomery, Susan Strickland, H Ryan Wagner, Jack Tsai, Financial Strain and Suicide Attempts in a Nationally Representative Sample of US Adults, *American Journal of Epidemiology*, Volume 189, Issue 11, November 2020, Pages 1266–1274.