# Case Study: Using Machine Learning Models to Predict Mental Health Crises

Joelle C. Rudolph

Bellevue University

DSC-630: Predictive Analytics

Professor Andrew Hua

14 April 2024

# Case Study: Using Machine Learning Models to Predict Mental Health Crises

# Introduction

For my case study, I chose to examine a machine learning model project developed by Garriga et al. to predict mental health crises from electronic health records (Garriga *et al*., 2022*)*. With the increasing demand for mental healthcare, especially after the COVID-19 pandemic, hospitals are being prompted to work on identifying new methods of anticipating demand in order to better deploy already limited resources to improve patient outcomes and reduce long-term costs. This high caseload makes the manual review of large quantities of data across a multitude of patients impractical, unsustainable and error-prone when trying to make proactive care decisions. Automating this process using machine learning would provide an immense impact on improving patient outcomes by enabling large-scale continuous data review of EHRs without the need for clinicians to be manually keeping track of every patient on a consistent basis.

Data was collected in two separate phases of this project. The first stage involved a retrospective cohort study to actually build and evaluate the mental health crisis prediction model using past electronic health record (EHR) data. The second stage was a follow-up 6-month study with the prediction model implemented in clinical practice as part of a prospective cohort study in order to evaluate the algorithm’s use in applied practice.

**Methods**

Preparation of the data began with compiling EHR data collected over several years (2012-2018) from a total of 17,122 patients. They decided to only include patients with a history of relapse because detecting first crises was not the goal of the project, so patients who had no crisis episode in their records were excluded from the data set. Furthermore, patients with only one crisis episode were also excluded in the data set since the researchers wanted to utilize information from multiple episodes. Patients with three or fewer months of records in the system were also excluded because the amount of historical data would have been insufficient for the algorithm to learn from. The resulting dataset included any predictions for the period after two crisis episodes and those having the first record at least 3 months before querying the model.

Regarding the features chosen for the model, the team defined the features into three categories: (1) static or semi-static patient information (such as age, gender and International Classification of Diseases coded diagnoses); (2) latest available assessments and interactions with the hospital (for example, most recent risk assessments or well-being indicators and severity and number of crisis events in the last episode); and (3) variables representing the time elapsed since the registered events (crisis episodes, contacts, and referrals). In total, they extracted 198 features. Once implemented, the model generated a predicted risk score (PRS) between 0 and 1 for each patient. The team tested a variety of different machine learning techniques but found that XGBoost (eXtreme Gradient Boosting) outperformed most of the other methods evaluated. The Mann–Whitney U-test suggested a significantly better performance of XGBoost (P < 0.01) when compared to the other methods, with the exception of a feed-forward neural network.

# Results

The main metric used when evaluating the results was the area under the receiver operating characteristic (AUROC) curves of the model, and the average precision (AP) was also utilized to better address unbalanced data sets. They obtained an AUROC of 0.797 for the general model and an AP of 0.159 for the XGBoost model, which significantly outperformed the scores of the two baseline models that the team used for comparison. The top 20 features on the model at each data point were also analyzed for their relative effect in the test set according to the mean absolute SHAP (SHapley Additive exPlanations) value. SHAP values for each of these features were then compared to find which held the highest predictive power in the test data. In order to ensure stability of the model and these interpretations, the team also generated 100 different random samples, with 40% of the patients per sample, and repeated the process of training the model and computing the SHAP values for the whole test set. Consistency of these predictors was then evaluated using cosine similarity between the SHAP values of the top 20 features of the original model and the models trained on the 100 samples.

# Conclusion

The team found that model performance scores varied slightly based on the nature of the disorder, with the model performing considerably better for organic disorders (disorders where there is a decrease in mental functioning that is not the result of a psychiatric condition), with an AUROC of 0.890 compared to the overall performance of 0.797. For other diagnostic groups, the performance ranged between 0.770 and 0.814. The lowest performance was observed for mood-affective disorders, followed by schizophrenia and schizotypal and delusional disorders. Data availability was also a factor, and having a longer data history improved the risk prediction performance for a given patient.

In terms of prediction power of different factors in the data, they learned that historical severity of symptoms (specifically, the total number of crisis episodes and the duration of the last episode), interactions with the hospital (including unplanned contacts, missed appointments or a recent crisis), patient characteristics (including age and individual risk indices) and total time since the patient’s first hospital visit held the most predictive power. Additionally, the recency of records had a major effect on the PRS, and unplanned contacts with a patient had the biggest short-term effect on PRS.

Crisis predictions were delivered to clinicians every 2 weeks in the clinical portion of data collection to assess the value of the model in clinical practice. The team asked clinicians to answer questions as to whether the algorithm helped identify patient deterioration and enabled a pre-emptive intervention. The model’s predictions were rated useful in 64% of the presented cases overall, indicating that the model could indeed be of help to clinicians when determining which patients may be at risk. Although the clinicians reported that the prediction model helped to prevent a crisis episode in 19% of cases, this eventuality was not witnessed due to the ethical aspect of clinicians reacting to the predictions, which would have been legally and ethically unacceptable. Further research could consider the feasibility of developing an algorithm to detect first crises in order to prevent such an occurrence from happening in the first place. Another idea to consider would be further analyzing the model’s performance across different demographic factors - they did find an increase in the AUROC for patients aged 65-74 years as well as a lower AUROC for the ‘Black’ ethnic subgroup compared to those in the ‘White’ ethnic subgroup, which wasn’t further explored in the current study. In future studies, it may be beneficial to further examine the potential causes of these disparities, which would involve examining uncontrollable factors as well as both known and unknown biases. If successful, this can help ensure that the model will perform more fairly across all patient populations.

# References

Garriga, R., Mas, J., Abraha, S. *et al.* Machine learning model to predict mental health crises

from electronic health records. *Nat Med* 28, 1240–1248 (2022).

https://doi.org/10.1038/s41591-022-01811-5