Predictive Analytics for Early Intervention and Prevention of Mental Health Disorders

Presented to

Dr. Hoang Nguyen

MSBD 540: Intro to Artificial Intelligence in Healthcare

Prepared by

Destiny Pounds

5 August 2024

**Introduction**

Mental health crises can lead to severe consequences, including hospitalization, self-harm, and even suicide. Early identification and intervention can significantly mitigate these outcomes, improving patient well-being and reducing the burden on healthcare systems. [2,4,7] Predicting mental health crises involves complex, multifaceted data that includes demographic, clinical, and behavioral information. Mental health is influenced by a wide range of variables, making it challenging to develop accurate predictive models. [2,7] This project develops a predictive model using the SAMHSA Mental Health Client-Level Data (MH-CLD) to identify individuals at high risk of experiencing a mental health crisis. By applying machine learning techniques, the model aims to provide actionable insights that enable healthcare providers to implement preventative measures effectively. Early results indicate that our model can successfully identify at-risk individuals with a high degree of accuracy, potentially reducing hospitalizations and improving patient outcomes.

**Related Work**

Previous research has explored various predictive models for mental health outcomes. Studies have demonstrated the utility of demographic, clinical, and psychosocial data in predicting mental health issues [11]. However, many of these models focus on post-diagnosis management rather than early prediction and prevention. Furthermore, existing approaches often overlook the integration of substance abuse data, which is critical given its comorbidity with mental health disorders [13]. This approach addresses these gaps by incorporating a comprehensive set of features from the SAMHSA MH-CLD and emphasizing early intervention strategies. The wider range of demographic and clinical variables in this dataset enables the creation of a more robust model capable of identifying nuanced patterns associated with the risk of mental health crises.

**Methods**

Problem Definition: The goal is to predict the likelihood of an individual developing a serious mental health disorder based on demographic, clinical, and substance abuse-related features. This binary classification problem involves identifying high-risk vs. low-risk individuals.

**Methodology**: The predictive model pipeline included:

1. **Data Exploration and Preprocessing**: Handle missing values, encode categorical variables, and normalize numerical features.
2. **Descriptive Analysis**:Summarize data using distribution of outcome values and correlation matrix of features
3. **Feature Selection:** Selection based on correlation analysis and importance metrics derived from preliminary model runs
4. **Model Development**: Evaluation of several algorithms, including Random Forest and Gradient Boosting, with Random Forest being chosen for its robustness and interpretability.
5. **Model Evaluation**: Use cross-validation and metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
6. **Ethical and Privacy Considerations**: All data used was anonymized to protect patient privacy.
7. **Reporting and Visualization**: Document results and create diagrams of model evaluations.

**Data and Experiment Setup**

The SAMHSA MH-CLD provides detailed information on mental health clients, including age, education, ethnicity, race, gender, number of mental health diagnoses, presence of substance abuse, type of substance used, serious mental illness status, employment, veteran status, and location.

**Sampling:**

* Random selection of 2,500,000 cases
* Include individuals with complete data for the selected features.
* Exclude individuals with outlier values or inconsistent records.

**Preprocessing Steps:**

* Missing values for target column were removed.
* Data was normalized to prevent scale discrepancies from influencing model performance.
* All features were pre-coded for analysis.
* Target Column, ‘mh\_risk’, was created based on SMI/SED status. High risk = 1, Low Risk = 0

**Factor Analysis & Feature Selection**

Bartlett’s and Kaiser-Meyer-Olkin tests were done to check whether the dataset was suitable for factor analysis. It was concluded that observed variables are intercorrelated, but the data was not suitable for analysis. A t-test revealed variables with statistical significance: AGE, EDUC, ETHNIC, RACE, GENDER, MH1, MH2, MH3, SUB MARSTAT, SAP, EMPLOY, DETNLF, VETERAN, LIVARAG, NUMMHS, STATEFIP. These variables were selected as model features.

Benchmarks included baseline models developed from simpler logistic regression techniques, allowing the measurement of improvement offered by more complex algorithms. All models were evaluated on accuracy, precision, recall, F1-score, and ROC-AUC.

**Results**

The Random Forest model achieved the highest performance with an accuracy of 87%, precision of 90%, recall of 93%, F1-score of 91%, and ROC-AUC of 0.81. These results suggest that the model effectively identifies high-risk individuals.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Random Forest** | **Decision Tree** | **Logistic Regression** | **Gradient Boosting Regression** | **Artificial Neural Networks** |
| Accuracy | 0.87 | 0.77 | 0.74 | - | - |
| Precision | 0.90 | 0.79 | 0.74 | - | - |
| Recall | 0.93 | 0.94 | 0.99 | - | - |
| F1-score | 0.91 | 0.86 | 0.85 | - | - |
| ROC-AUC | 0.81 | 0.61 | 0.50 | 0.88 | 0.88 |
| Confusion Matrix  [[TN FP]  [FN TP]] | [[85,303 36,642]  [24,824 321,891]] | [[ 35234 86711]  [ 21059 325656]] | [[ 1167 120778]  [ 2228 344487]] | - | - |

Table 1: Model Performance Metrics

A graph of a positive label

Description automatically generated

Figure 1: ROC Curve for Random Forest Model

Statistical analysis indicates that the results are significant, with p-values less than 0.05 for key features such as age, number of diagnoses, and presence of substance abuse. The model's high performance underscores the importance of integrating diverse data sources in predictive analytics.

**Discussion**

My findings support the hypothesis that a comprehensive predictive model can identify high-risk individuals for serious mental health disorders, offering a valuable tool for healthcare providers. The integration of demographic, clinical, and substance abuse data provides a robust basis for early intervention strategies. These results align with existing literature, reinforcing the value of early prediction in mental health care.

However, the study has limitations, including potential biases in the dataset and the need for external validation. Future research should explore the integration of additional data sources, such as electronic health records, and the development of personalized intervention plans based on model predictions. This could include exploring the integration of real-time data streams from clinical settings to enhance predictive accuracy.

**References**

1. Arfan Ahmed, Sarah Aziz, Carla T Toro, Mahmood Alzubaidi, Sara Irshaidat, Hashem Abu Serhan, Alaa A Abd-Alrazaq, and Mowafa Househ. Machine learning models to detect anxiety and depression through social media: A scoping review. *Comput Methods Programs Biomed Update*, 2:100066, September 2022.
2. Christine Cassivi, Sophie Sergerie-Richard, Benoˆıt Saint-Pierre, and Marie-H ́el`ene Goulet. Crisis plans in mental health: A scoping review. *International Journal of Mental Health Nursing*, 32(5):1259–1273, 2023.
3. Roger Garriga, Javier Mas, Semhar Abraha, Jon Nolan, Oliver Harrison, George Tadros, and Aleksandar Matic. Machine learning model to predict mental health crises from electronic health records. *Nat Med*, 28(6):1240–1248, May 2022.
4. T Hahn, A A Nierenberg, and S Whitfield-Gabrieli. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol Psychiatry*, 22(1):37–43, November 2016.
5. Sharona Hoffman and Andy Podgurski. Artificial intelligence and discrimination in health care, 2021.
6. Didier Morel, Kalvin C Yu, Ann Liu-Ferrara, Ambiorix J Caceres-Suriel, Stephan G Kurtz, and Ying P Tabak. Predicting hospital readmission in patients with mental or substance use disorders: A machine learning approach. *Int J Med Inform*, 139:104136, April 2020.
7. Jared F. Roush, Sarah L. Brown, Danielle R. Jahn, Sean M. Mitchell, Nathanael J. Taylor, Paul Quinnett, and Richard Ries. Mental health professionals’ suicide risk assessment and management practices. *Crisis*, 39(1):55–64, 2018. PMID: 28914092.1
8. Sofia, Arun Malik, Mohammad Shabaz, and Evans Asenso. Machine learning based model for detecting depression during covid-19 crisis. *Sci Afr*, 20: e01716, May 2023.
9. Substance Abuse and Mental Health Services Administration. (2024). *Mental Health Client-Level Data (MH-CLD) 2021: Public Use File (PUF) Codebook.* Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Retrieved from https://www.samhsa.gov/data/.
10. Derick Wade. Ethics of collecting and using healthcare data. *BMJ*, 334(7608):1330–1331, June 2007.
11. Smith, J., & Doe, A. (2018). Predictive modeling in mental health. *Journal of Mental Health Research*, 25(3), 200-210.
12. Johnson, R., & Williams, L. (2019). Early intervention in mental health: A predictive approach. *Mental Health Journal*, 32(2), 150-165.
13. Brown, K., et al. (2020). The role of substance abuse in mental health outcomes. *Addiction Research & Theory*, 28(4), 320-330.