**Using Random Forest to Predict Panic Disorder Diagnoses**

**Executive Summary**

**Statement of the Problem & Hypothesis**

In the post-pandemic world, mental health diagnosis and treatment has become more necessary than ever due to heightened stress and anxiety brought on by major societal changes. Studies have shown that the prevalence of anxiety-related disorders increased during and after the pandemic (Georgieva et al., 2021). Panic disorder in particular is defined by “recurring and unexpected panic attacks, accompanied by anticipatory anxiety about future attacks and their consequences” (Oussi et al., 2023). It can be a debilitating disorder that affects all aspects of one’s life. A previous study found that the random forest algorithm outperformed other machine learning techniques when predicting generalized anxiety disorder among women (Jothi et al., 2018). Building a random forest model to predict panic disorder diagnoses could allow mental health providers to detect the disorder earlier than traditional methods, which could lead to patients receiving much needed and appropriate treatment sooner.

The research question for this capstone study is, “Can the random forest approach be used to predict whether a person has a panic disorder diagnosis?" The null and alternate hypotheses of the study are:

*H0: A random forest model predicting panic disorder with a mean squared error (MSE) of 0.25 or lower cannot be achieved using the panic disorder data set.*

*H1: A random forest model predicting panic disorder with a mean squared error (MSE) of 0.25 or lower can be achieved using the panic disorder data set.*

The MSE metric sets a threshold for model accuracy to ensure that the model is fit for real life application.

**Data Analysis Process**

The first step in the data analysis process was to collect data relevant to the research question. The data set used was publicly available and found on Kaggle.com at the following link: <https://www.kaggle.com/datasets/muhammadshahidazeem/panic-disorder-detection-dataset/data>. Next I imported all the necessary packages for data manipulation, visualization, machine learning, and evaluation metrics. I loaded both available data sets (training & testing) and concatenated them for data preparation. Next I explored the data by viewing the shape, summary, head, and descriptive statistics of the data set. Before data cleaning, there was 120,000 rows and 17 columns. I calculated the data sparsity percentage of the data set which was found to be 5.6%. See below for variables included in the original data set:

|  |  |
| --- | --- |
| ***Field*** | ***Type*** |
| Participant ID | Numeric |
| Age | Numeric |
| Gender | Categorical |
| Family History | Categorical |
| Personal History | Categorical |
| Current Stressors | Categorical |
| Symptoms | Categorical |
| Severity | Categorical |
| Impact on Life | Categorical |
| Demographics | Categorical |
| Medical History | Categorical |
| Psychiatric History | Categorical |
| Substance Use | Categorical |
| Coping Mechanisms | Categorical |
| Social Support | Categorical |
| Lifestyle Factors | Categorical |
| Panic Disorder Diagnosis | Numeric |

Next, I began the data cleaning process by dropping the columns I knew I would not be using in the analysis. This included Participant ID, Family History, and Personal History. I assessed for duplicates and dropped the 187 duplicates that were found. I assessed for missing values, and none were found. I created seaborn boxplots of the quantitative variables (Age & Panic Diagnosis) to assess for outliers. None were found. I obtained the summary statistics of the remaining categorical variables to assess the distribution of each. All had a normal distribution, so no imputation was needed.

To analyze the data, I created univariate visualizations of each variable by plotting histograms for each. I also created bivariate visualizations of the predictor variables with the target variable (Panic Disorder Diagnosis). For the quantitative predictor variable (Age), I plotted a scatterplot. Since Panic Disorder Diagnosis is a binary variable, I chose to create crosstabs of the categorical predictor variables with the target variable.

I continued preparing the data by encoding the ordinal categorical variables, or variables with a natural rank order, with numeric values. Since the Demographics variable reflects population density (Rural/Urban), I chose to classify it as an ordinal variable. I then performed one-hot encoding for the categorical variables without natural ordering and converted the variables to integers. The final encoded data set included 119,813 rows and 32 columns. I saved the cleaned data set to a CSV file. I created a seaborn heatmap to identify correlation between variables. Next I split the data 80/20 into training and testing sets. I saved the split data sets to CSV files.

I began to build my random forest model by first specifying the hyperparameters for the parameter grid for cross-validation. Then I used the RandomizedSearchCV function to evaluate multiple random combinations of hyperparameters and select those that yielded the best cross-validated score. I fit the best model to my training set and printed the parameters of the best model. I then used the model to make predictions on the testing set. I evaluated the model by calculating the accuracy, mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and r-squared value. See below for evaluation results:

A computer screen shot of a program code

Description automatically generated

**Findings**

The univariate analyses showed that all variables had a normal distribution. The bivariate analyses showed that those with a positive panic disorder diagnosis were more likely to have: high current stressors, panic attacks as a symptom, a severity rating of 'Severe', a signficant impact on life, low social support, and sleep quality as a lifestyle factor. Those with a positive panic disorder diagnosis were also less likely to use exercise as a coping mechanism. Aside from the expected correlation between the one-hot encoded variables, there was low correlation identified in the seaborn boxplots.

According to the evaluation metric calculations utilized in this study, the accuracy score was 1.0, the MSE was 0.0, the RMSE was 0.0, the MAE was 0.0, and the r-squared value was 1.0. This implies that the model has perfectly predicted all samples in the testing set. Since the MSE was under 0.25 threshold, I reject the null hypothesis.

**Limitations**

One limitation of this analysis was that there was no available data dictionary with column descriptions. Due to this lack of clarity, several variables had to be dropped. The historical variables may have been beneficial to my model; however, I would not have been able to draw conclusions from these fields. One disadvantage to the heatmap technique used is that it doesn't include the correlation values in the visual, only the heat color of the correlation. This made it slightly trickier to interpret the heatmap.

**Proposed Actions**

A perfect accuracy score can sometimes be caused by overfitting of the model. This occurs when the model has memorized the training data. One recommended course of action would be to investigate the model's hyperparameters further and use cross-validation to check the consistency across different subsets of the data and verify the model's generalization ability. Another recommended course of action would be to move forward with implementation of the model and pilot the program with a mental health provider, given its perfect accuracy score. This would allow patients predicted by the model to have a panic disorder diagnosis to receive their diagnosis sooner and be prescribed an appropriate treatment plan.

**Expected Benefits**

The primary expected benefit of the study is earlier identification of individuals at risk of developing panic disorder based on the available data. Early detection can lead to timely intervention and treatment, potentially preventing the disorder’s progression. 4.2% of the total study population had a panic disorder. While that percentage may seem low, it amounted to 5,120 individuals. A model with a perfect 1.0 accuracy score can improve diagnostic accuracy and reveal patterns that may be overlooked by mental health providers. It can also guide the customization of treatment interventions through understanding which features (i.e. stress levels, lifestyle factors) are most predictive of the disorder. Implementing this model can enhance the ability of clinicians to make informed decisions for their patients.

**Sources**

Azeem, M. (2023). Panic disorder detection dataset. Kaggle.com. Retrieved January 2, 2025 from <https://www.kaggle.com/datasets/muhammadshahidazeem/panic-disorder-detection-dataset/data>

Georgieva, I., Lepping, P., Bozev, V., Lickiewicz, J., Pekara, J., Wikman, S., Loseviča, M., Raveesh, B. N., Mihai, A., & Lantta, T. (2021). Prevalence, new incidence, course, and risk factors of PTSD, depression, anxiety, and panic disorder during the Covid-19 pandemic in 11 countries. Healthcare, 9(6), 664. <https://doi.org/10.3390/healthcare9060664>

Jothi, N., Husain, W., Rashid, N., Xin, L. (2018). Predicting generalised anxiety disorder among women using decision tree-based classification. International Journal of Business Information Systems, 29(1), 75. <https://doi.org/10.1504/IJBIS.2018.093998>

Oussi, A., Hamid, K., Bouvet, C. (2023). Managing emotions in panic disorder: A systematic review of studies related to emotional intelligence, alexithymia, emotion regulation, and coping. Journal of Behavior Therapy and Experimental Psychiatry, 79. <https://doi.org/10.1016/j.jbtep.2023.101835>

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193. <https://doi.org/10.3390/info11040193>

Shafi, A. (2024). Random forest classification with scikit-learn. DataCamp.com. Retrieved January 2, 2025 from <https://www.datacamp.com/tutorial/random-forests-classifier-python>

U.S. Food & Drug Administration. (2015). FDA.gov. Retrieved January 2, 2025 from <https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/fda-bioresearch-monitoring-information/definitions#:~:text=Good%20Laboratory%20Practice%20(GLP)%20is,%2C%20recorded%2C%20archived%20and%20reported>