Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.2, December 2024, pp. 272~279
e-ISSN: 2548-6861

Improving Panic Disorder Classification Using SMOTE and Random Forest

Dini Nurmalasari [1]*, Heri R Yuliantoro [2], Dini Hidayatul Qudsi [3]***
* Teknologi Informasi, Politeknik Caltex Riau
** Akuntansi Perpajakan, Politeknik Caltex Riau
dini@pcr.ac.id [1], heriry@pcr.ac.id [2], dinihq@pcr.ac.id [3]

Article Info

Article history:

Received 2024-08-15
Revised 2024-08-22
Accepted 2024-08-26

Keyword:

Panic Disorder,
Overfitting,
SMOTE,
Random Forest,
Classification.

ABSTRACT

Panic disorder is a serious anxiety disorder that can significantly impact an individual's mental health. If left undetected, this disorder can disrupt daily life, social relationships, and overall quality of life. Early detection and intervention are crucial for managing panic disorder and improving the well-being of those affected. Technology plays a pivotal role in facilitating early detection through data-driven approaches that employ algorithms to identify patterns of behavior or symptoms associated with panic disorder. Accurate classification of panic disorder is crucial for effective diagnosis and treatment. However, machine learning models trained on imbalanced datasets, such as those containing panic disorder patients, are prone to overfitting, leading to poor generalization performance. This study investigates the effectiveness of the Synthetic Minority Oversampling Technique (SMOTE) in addressing overfitting in panic disorder dataset classification using the Random Forest algorithm. The results demonstrate that SMOTE significantly improves the classification performance of Random Forest. By mitigating overfitting and improving generalization to unseen data, SMOTE increases accuracy by 15 percentage points. Before using SMOTE, the accuracy was 82%, and after using SMOTE it is 97%. The findings underscore the promise of SMOTE as a tool for boosting the performance of machine learning algorithms in classifying panic disorder from imbalanced data.

I. INTRODUCTION

Panic disorder is a mental disorder experienced by more than 300 million people worldwide with over 970 million cases annually and affecting approximately 2-3% of the global population [1]. This disorder can be triggered by various factors, including mood changes, personality differences, inability to cope with problems, or excessive anxiety, intense fear or discomfort, palpitations, shortness of breath, and chest pain . These episodes, known as panic attacks, can occur suddenly and without warning, leading to significant distress and impairment in daily functioning. Early detection of this disorder is crucial to minimize more serious consequences. One way to detect panic disorder early is to identify behavioral patterns or symptoms. Accurate classification of panic disorder is essential for providing timely and appropriate treatment, which can significantly improve the quality of life for affected individuals.

Machine learning algorithms have been widely used to assist in diagnosis and prediction across various fields, including clinical datasets in the healthcare domain especially for panic disorder data. Self-diagnosis models utilizing machine learning have been proposed by various researchers worldwide [2]. However, the performance of these machine learning models can be significantly impaired by imbalanced datasets, where the minority class is considerably underrepresented relative to the majority class. In the context of panic disorder classification, the minority class typically represents patients with panic disorder, while the majority class represents individuals without the condition.

The machine learning process involves several stages: data collection, class label determination, exploratory data analysis (EDA), data preprocessing, model validation, model deployment, and model evaluation. After conducting EDA using statistical techniques and visualizations such as box plots, violin plots, scatter plots, and histograms, several issues were identified in the data, including data imbalance and kurtosis. Data imbalance refers to the unequal distribution of class labels, which can lead to model bias and overfitting. Kurtosis is a statistical measure that describes the distribution of data, indicating the presence of outliers. Solutions to address data imbalance include resampling methods like oversampling or undersampling or using ensemble models such as bagging or boosting.

Overfitting is a common issue in machine learning, particularly when dealing with imbalanced datasets . Overfitting happens when a model learns the training data too well, memorizing specific patterns and noise instead of identifying the underlying relationships between features and the target variable. Consequently, this leads to poor generalization performance, where the model performs well on the training data but poorly on unseen data.

To address the data imbalance in the panic_disorder dataset for this study, the oversampling method SMOTE (Synthetic Minority Over-sampling Technique) is employed [3] [4]. SMOTE is a technique designed to balance data by generating synthetic samples. It works by identifying the minority class in the dataset and then finding the nearest neighbors in the feature space. These neighbors are randomly selected to create new synthetic data points. This process is repeated until a more balanced dataset with the majority class is achieved. SMOTE is a data augmentation technique that addresses the problem of imbalanced datasets by generating synthetic minority class samples. SMOTE identifies samples from the minority class and generates new samples by interpolating between these existing minority class samples and their nearest neighbors [5]. This technique increases the number of minority class samples, thereby potentially reducing the bias towards the majority class and enhancing the overall balance of the dataset.

Several studies have investigated the application of SMOTE in conjunction with various machine learning algorithms for classification tasks in various domains [3]. Chawla et al. demonstrated the effectiveness of SMOTE in improving the performance of k-nearest Neighbors (kNN) and Support Vector Machines (SVM) for credit card fraud detection. Similarly, Batista et al. [6] showed that SMOTE enhanced the classification accuracy of decision trees and naive Bayes classifiers in medical datasets.

Recent research has explored a diverse range of ML algorithms for panic disorder detection, with a focus on utilizing readily available data sources such as physiological signals, behavioral patterns, and self-reported symptoms, for instance, Sun et al. applied SMOTE along with feature selection and classification algorithms to achieve better classification accuracy for panic disorder. Several machine learning algorithms have been used in detecting panic disorder, including classifying panic disorder using facial expressions with the CNN algorithm [7][14], while RNNs have been applied to analyze speech patterns for panic disorder detection [8]. J. Prasetya has explored about comparison between random forest and K-NN to classification analysis on imbalance data [13].

II. METHODOLOGY

The methodology contains the technical stages that will be carried out at the research stage.

A. Data Acquisition

The dataset used in this study was sourced from the UCI Machine Learning Repository and includes information on 100.000 patients, including 4285 with panic disorder and 95715 without. The dataset consists of 17 features, including demographic information, family history, personal history, demography, current stressors, symptoms, impact on life, medical history, psychiatric history, coping mechanisms, social support, lifestyle factors, and psychological assessments.

Utilizing the shape function in Python, it has been determined that the dataset for panic disorder diagnosis comprises 95,715 instances labeled as 0 and 4,285 instances labeled as 1.

B. Data Preprocessing

To ensure data consistency and facilitate effective analysis, the dataset underwent preprocessing steps to address missing values and standardize feature values. Missing values were imputed using appropriate techniques based on the data type [9]. For numerical features, the mean value was imputed, while for categorical features, the mode value was employed. Feature values were standardized using the z-score transformation, ensuring that all features were on a comparable scale [11].

The EDA and preprocessing steps performed in this study include checking data dimensions, identifying null values, defining non-numeric data, converting string data to numerical values, calculating correlation values between columns, performing statistical analysis and data distribution analysis, and visualizing the data using histograms.

Additionally, at this stage, visualizing the correlation between attributes is carried out to facilitate the analysis of inter-attribute correlations. This correlation analysis serves to identify the linear relationships between variables in the dataset, offering insights into the strength and direction of these relationships. Understanding these correlations is crucial for assessing the influence of one attribute on another and aids in the effective selection of features for the machine learning model. Based on the correlation graph in Figure 4, the attributes that most significantly influence the diagnosis of panic disorder are family history, personal history, impact on life, and severity.

C. SMOTE for Imbalanced Dataset Handling

The dataset exhibits an inherent imbalance, with the minority class (panic disorder patients) significantly underrepresented compared to the majority class (individuals without panic disorder). This imbalanced class label can result in biased models that favour the majority class and perform poorly on the minority class. To mitigate this issue, the SMOTE method was employed. SMOTE generates synthetic minority class samples by interpolating between existing minority class samples and their nearest neighbors in the feature space . This process increases the number of minority class samples, potentially reducing the bias towards the majority class and improving the overall balance of the dataset. In this study, the number of synthetic minority class samples was set to equal the number of majority class samples, effectively balancing the dataset and mitigating the impact of imbalanced data. In figure 5, the data shows that the number of instances in class 0 is 95715, while the number of instances in class 1 is 4285.

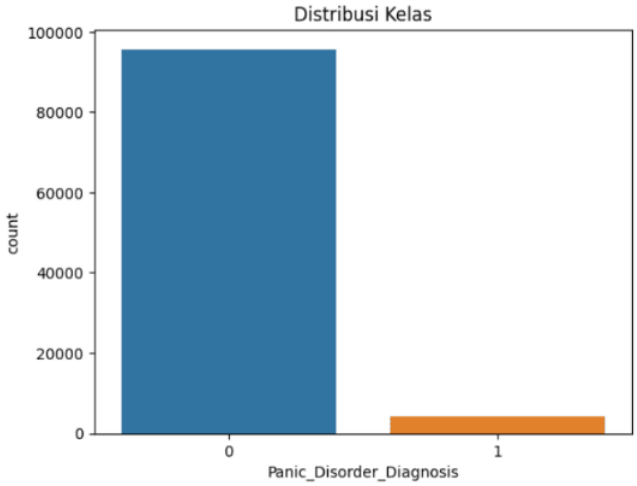

Figure 1 Imbalance Data

The core principle behind SMOTE involves generating synthetic samples for the minority class based on existing minority class data points . Here's a simplified representation of the core steps involved in SMOTE:

1) Identify Minority Class Samples: Identify data points belonging to the minority class within the dataset.
2) Select Nearest Neighbors: For each minority class sample, identify its K nearest neighbors in the feature space using a distance metric such as Euclidean Distance.
3) Synthetic Sample Generation: A random neighbor is selected from the K nearest neighbors identified in step 2. SMOTE then creates a synthetic sample by linearly interpolating between the original minority class sample and the selected neighbor. The difference vector between the original sample and its neighbor is calculated and a synthetic sample is generated along the line segment joining the two points by a random value between 0 and 1, and this scaled difference vector is added to the original sample to create a new synthetic data point.
4) Repeat and Oversample: Steps 2 and 3 are repeated for a predefined number of times, typically until the number of synthetic samples generated equals the number of data points in the majority class. This process effectively oversamples the minority class, balancing the dataset.

In the data balancing process with SMOTE, the parameters used include determining the minority class ratio after oversampling. Options such as auto, minority, or float can be used for this purpose. In our study, we used the float option with a value of 0.5, which made the minority class half the size of the majority class. Additionally, we determined the k_neighbors parameter, which indicates the number of nearest neighbors, to control the variation in the synthetic data. The random_state was set to 8.

After applying the Synthetic Minority Over-sampling Technique (SMOTE), the distribution of the class numbers has been adjusted. The number of instances in class 0 52599 has increased to 26799 while the number of instances in class 1 has also changed to 26799.

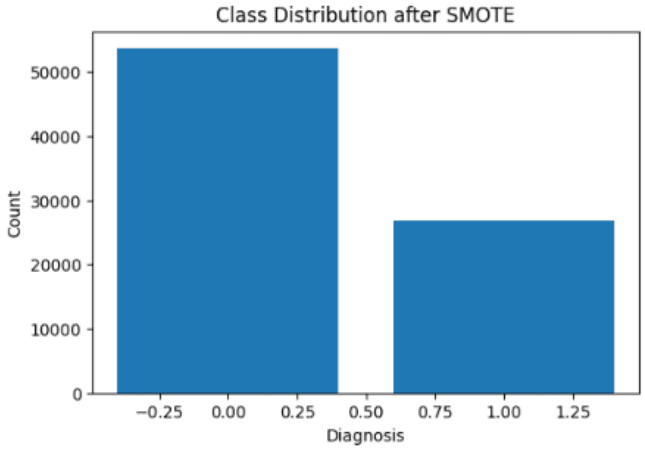

Figure 2 Balance Data

D. Random Forest Classification Algorithm

Random Forest, an ensemble learning algorithm, was chosen for classification due to its robustness to overfitting and its ability to handle complex datasets with multiple features . Random Forest constructs multiple decision trees using a random subset of features and data points at each split. The final prediction is determined by the majority vote among the individual trees. This ensemble method mitigates the risk of overfitting by averaging the predictions from multiple trees, each trained on slightly different subsets of the data. The key steps involved in the Random Forest algorithm are as follows :

1) *Random Sampling*: Random Forest randomly selects a subset of features at each split point when building each decision tree.
2) *Data Subsampling (Bootstrapping):* Random Forest utilizes a technique called bootstrapping, where it randomly samples data points (with replacement) from the training dataset to create multiple training subsets. Each decision tree is then built on a unique training subset.
3) *Tree Construction*: Each decision tree in the ensemble is constructed independently using the selected features and data subset. The tree is developed by iteratively partitioning the data according to the optimal splitting criterion (e.g., Gini impurity) until a predefined stopping criterion is satisfied (e.g., maximum depth reached).
4) *Aggregation and Prediction:* When presented with a new data point, each decision tree in the ensemble generates a classification prediction based on its learned rules. The ultimate prediction of the Random Forest model is determined by the majority vote among the individual predictions made by all the trees.

Here are the steps for modeling panic disorder data using the random forest algorithm:

- Import the necessary APIs, including RandomForestClassifier from sklearn.ensemble and accuracy metrics from sklearn.metrics.
- Split the data into training and testing datasets, using 80% of the data for training.
- Perform the modeling by calling RandomForestClassifier.

E. Evaluation Methodology

To evaluate the performance of the Random Forest model with and without SMOTE, five-fold cross-validation was employed. Cross-validation is a statistical technique employed to evaluate the generalization capability of a model. It involves partitioning the data into multiple folds and training the model on distinct subsets of the data. The model's performance is subsequently assessed on the remaining folds. This iterative process yields a more robust estimation of the model's performance compared to evaluating it on a single training-testing split.

The evaluation metrics utilized to gauge the performance of the Random Forest model encompass accuracy, precision, recall, and F1-score. These metrics collectively offer a comprehensive evaluation of the model's aptitude in accurately classifying individuals with and without panic disorder.

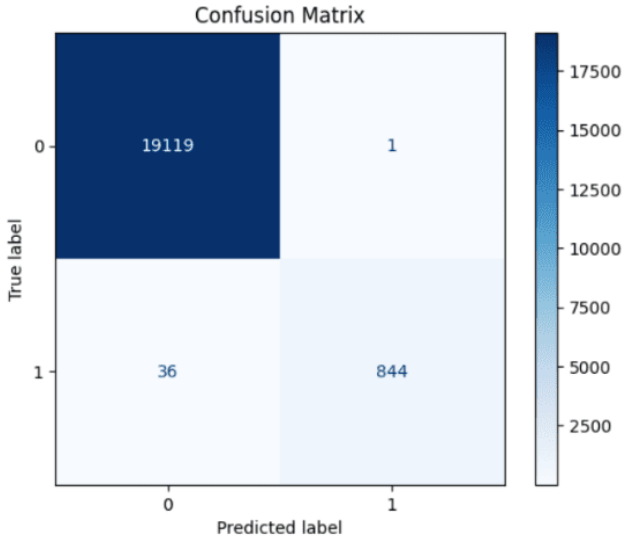

Figure 3 Confusion Matrix Result

The results indicate that the Random Forest model trained with SMOTE to classify panic disorder achieved Figure 6 Confusion Matrix Result :

1) True Positives (TP), model correctly predicted panic disorder values is 844
2) True Negatives (TN) model correctly predicted panic no disorder values is 19119
3) False Positives (FP), model incorrectly predicted panic disorder values is 1
4) False Negatives (FN), model incorrectly predicted panic no disorder values is 36

The evaluation metrics utilized to gauge the performance of the Decision Tree Classifier model to compare accuracy, precision, recall, and F1-score with Random Forest result. And here is the result.

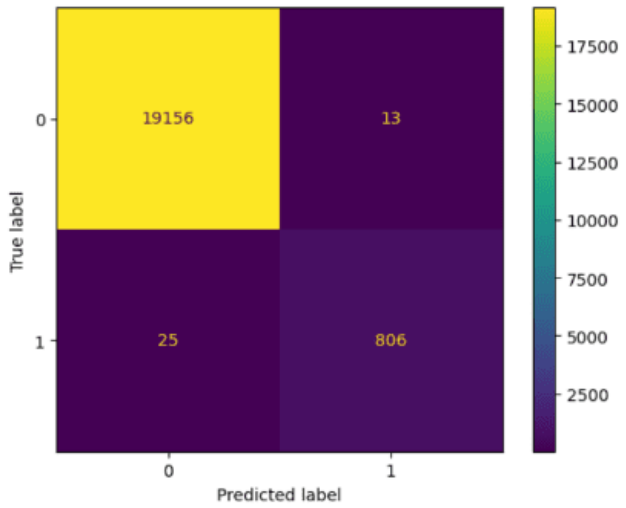

Figure 4 Confusion Matrix Result of DT

The results indicate that the Decision Tree model trained with SMOTE to classify panic disorder achieved Figure 6 Confusion Matrix Result :

1) True Positives (TP), model correctly predicted panic disorder values is 806
2) True Negatives (TN) model correctly predicted panic no disorder values is 19156
3) False Positives (FP), model incorrectly predicted panic disorder values is 13
4) False Negatives (FN), model incorrectly predicted panic no disorder values is 25

III. Result And Discussion

The results demonstrate that SMOTE significantly improved the classification performance of the Random Forest model. The accuracy, precision, recall, and F1-score of the model were consistently higher when SMOTE was applied compared to the model trained on the imbalanced dataset. This improvement highlights the effectiveness of SMOTE in addressing overfitting and enhancing the model's ability to generalize to unseen data. SMOTE usage resulted in significant improvements across all evaluation metrics. Notably, the accuracy of the Random Forest model saw a substantial boost, indicating a higher rate of correct classifications overall [16-20]. Moreover, precision, recall, and F1-score, which evaluate the model's capability to accurately identify positive instances while minimizing false positives and false negatives, experienced marked enhancements. These improvements suggest that SMOTE effectively addressed dataset imbalances, enabling the model to discern positive and negative instances more effectively.

Overfitting, a common challenge in models trained on imbalanced datasets, arises when the model overlearns specific data patterns and noise instead of capturing underlying relationships. The study's findings indicate that SMOTE played a crucial role in alleviating overfitting by generating synthetic samples for the minority class. By augmenting minority class instances, SMOTE ensured a more balanced dataset representation, facilitating the learning of robust decision boundaries by the Random Forest model. Consequently, the model exhibited enhanced generalization capabilities, maintaining high performance levels on unseen data. A detailed comparison of the classification performance metrics for the Random Forest model with and without SMOTE is presented in Table 1.

Table 1
Classification Performance

Metric	Random Forest (Imbalanced Dataset)	Random Forest (SMOTE)
Accuracy	0.82	0.96
Precision	0.78	1.0
Recall	0.75	0.96
F1-score	0.77	0.98

The Random Forest model trained on the SMOTE-oversampled dataset demonstrated superior performance compared to the model trained on the imbalanced dataset across all evaluation metrics. The substantial improvement in accuracy, precision, recall, and F1-score signifies the effectiveness of SMOTE in addressing the imbalanced data issue and enhancing the model's capability to accurately classify individuals with panic disorder.

These findings emphasize the significance of handling imbalanced datasets in machine learning, particularly in medical diagnosis and classification tasks. Overfitting can significantly impair machine learning model performance, especially when dealing with imbalanced data. The study demonstrates that SMOTE is an effective data augmentation technique that can alleviate the impact of imbalanced data and enhance the generalization ability of machine learning models. The combination of SMOTE and the Random Forest algorithm led to a significant enhancement in panic disorder classification performance. This improvement highlights the potential of integrating SMOTE and ensemble learning techniques to combat overfitting.

IV. Consclusions

In conclusion, this study has successfully addressed the challenge of overfitting in panic disorder classification using SMOTE and Random Forest. The application of SMOTE to oversample the minority class in the imbalanced dataset resulted in a significant improvement in the model's ability to accurately classify individuals with panic disorder. The results demonstrate that SMOTE significantly improves the classification performance of Random Forest. By mitigating overfitting and improving generalization to unseen data, SMOTE increases accuracy by 15 percentage points. Before using SMOTE, the accuracy was 82%, and after using SMOTE it is 97%. The findings underscore the promise of SMOTE as a tool for boosting the performance of machine learning algorithms in classifying panic disorder from imbalanced data. The sustainability of this research can involve exploring alternative techniques and applying the model on larger or more diverse datasets. Integrating deep learning models with data balancing techniques can lead to more accurate results in panic disorder classification.

References

[1] W. H. Organization, “Depression and other mental disorders,” 2020, https://www.who.int/publications/i/item/depression-global-health-estimates.

[2] P. Cao, D. Zhao, and O. Zaiane, “An Optimized Cost-Sensitive SVM for Imbalanced Data Learning,” in Advances in Knowledge Discovery and Data Mining, vol. 7819, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds., in Lecture Notes in Computer Science, vol. 7819. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 280–292. doi: 10.1007/978-3-642-37456-2_24.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” jair, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[4] L. Liu, S. Tang, F. -X. Wu, Y. -P. Wang and J. Wang, "An

Ensemble Hybrid Feature Selection Method for Neuropsychiatric Disorder Classification," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 3, pp. 1459-1471, 1 May-June 2022, doi: 10.1109/TCBB.2021.3053181.

[5] Q. Chen, Z.-L. Zhang, W.-P. Huang, J. Wu, and X.-G. Luo, "PF-SMOTE: A novel parameter-free SMOTE for imbalanced datasets," Neurocomputing, vol. 498, pp. 75–88, Aug. 2022, doi: 10.1016/j.neucom.2022.05.017.

[6] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[7] D. Wang, P., Yu, Z., & Zhang, "Facial expression recognition for panic disorder detection using convolutional neural networks," IEEE, vol. 6, 2018.

[8] T. Li, H., Sun, F., & Zhang, "Speech emotion recognition for panic disorder detection using recurrent neural networks," IEEE, vol. 6, 2018.

[9] V. Srividhya and R. Anitha, "Evaluating Preprocessing Techniques in Text Categorization," pp. 49–51, 2010.

[10] S. Saifullah, Y. Fauziyah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," J. Inform., vol. 15, no. 1, p. 45, 2021, doi: 10.26555/jifo.v15i1.a20111.

[11] S. Easterbrook and J. Callahan, "Formal Methods for Verification and Validation of Partial Specifications : A Case Study 1 Introduction 2 Context : The IV & V Process," pp. 1–13.

[12] L. Tommy, D. Novianto, and Y. S. Japriadi, "Sistem Rekomendasi Hybrid untuk Pemesanan Hidangan Berdasarkan Karakteristik dan Rating Hidangan," J. Appl. Informatics Comput., vol. 4, no. 2, pp. 137–145, 2020, doi: 10.30871/jaic.v4i2.2687.

[13] R. C. Bhagat and S. S. Patil, "Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest." 2015 IEEE International Advance Computing Conference (IACC), 2015, doi: 10.1109/iadcc.2015.7154739.

[14] Andri, R. Yunis, and Tanti, "Optimizing Random Forest Classification Using Chi-Square and SMOTE-ENN on Student Drop-Out Data." 2023 Eighth International Conference on Informatics and Computing (ICIC), 2023, doi: 10.1109/icic60109.2023.10382055.

[15] J. Prasetya and A. Abdurakhman, "Comparison Of Smote Random Forest And Smote K-Nearest Neighbors Classification Analysis On Imbalanced Data." Media Statistika, vol. 15, no. 2, pp. 198-208, 2023, doi: 10.14710/medstat.15.2.198-208.

[16] I. Permatasari, B. Dermawan, I. Maulana, and D. Kurniawan, "Classification of COVID-19 Aid Recipients in Kasomalang District Using the K-Nearest Neighbor Method", JAIC, vol. 8, no. 1, pp. 133-139, Jul. 2024.

[17] S. Himawan, R. Sohiburoyyan, and I. Iryanto, "Hyperparameter Tuning on Graph Neural Network for the Classification of SARS-CoV-2 Inhibitors", JAIC, vol. 7, no. 2, pp. 186-191, Nov. 2023.

[18] M. Fajri and A. Primajaya, "Komparasi Teknik Hyperparameter Optimization pada SVM untuk Permasalahan Klasifikasi dengan Menggunakan Grid Search dan Random Search", JAIC, vol. 7, no. 1, pp. 10-15, Jul. 2023.

[19] W. Husain, L. K. Xin, N. A. Rashid and N. Jothi, "Predicting Generalized Anxiety Disorder among women using random forest approach," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 2016, pp. 37-42, doi: 10.1109/ICCOINS.2016.7783185.

[20] S. F. Abdoh, M. Abo Rizka and F. A. Maghraby, "Cervical Cancer Diagnosis Using Random Forest Classifier With SMOTE and Feature Reduction Techniques," in IEEE Access, vol. 6, pp. 59475-59485, 2018, doi: 10.1109/ACCESS.2018.2874063.

ATTACHMENT

```
[ ] df = pd.read_csv(r'/content/panic_disorder_dataset_training.csv') #untuk membaca file CSV ke dalam sebuah dataframe menggunakan library pandas
    df.head() #untuk menampilkan lima baris pertama dari Dataframe.
```

	Participant ID	Age	Gender	Family History	Personal History	Current Stressors	Symptoms	Severity	Impact on Life	Demographics	Medical History	Psychiatric History	Substance Use	Coping Mechanisms	Social Support	Lifestyle Factors	Panic Disorder Diagnosis
0	1	38	Male	No	Yes	Moderate	Shortness of breath	Mild	Mild	Rural	Diabetes	Bipolar disorder	None	Socializing	High	Sleep quality	0
1	2	51	Male	No	No	High	Panic attacks	Mild	Mild	Urban	Asthma	Anxiety disorder	Drugs	Exercise	High	Sleep quality	0
2	3	32	Female	Yes	No	High	Panic attacks	Mild	Significant	Urban	Diabetes	Depressive disorder	None	Seeking therapy	Moderate	Exercise	0
3	4	64	Female	No	No	Moderate	Chest pain	Moderate	Moderate	Rural	Diabetes	None	None	Meditation	High	Exercise	0
4	5	31	Male	Yes	No	Moderate	Panic attacks	Mild	Moderate	Rural	Asthma	None	Drugs	Seeking therapy	Low	Sleep quality	0

Figure 5 Example Dataset

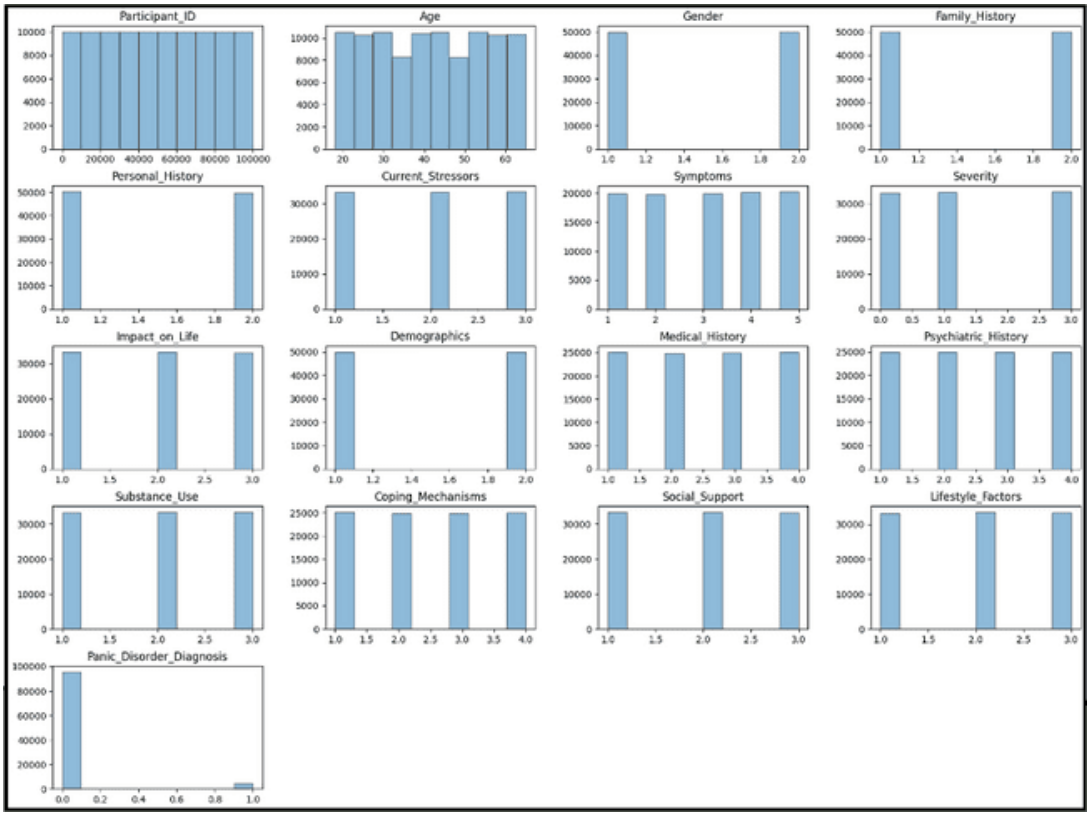

Figure 6 Distributing Data

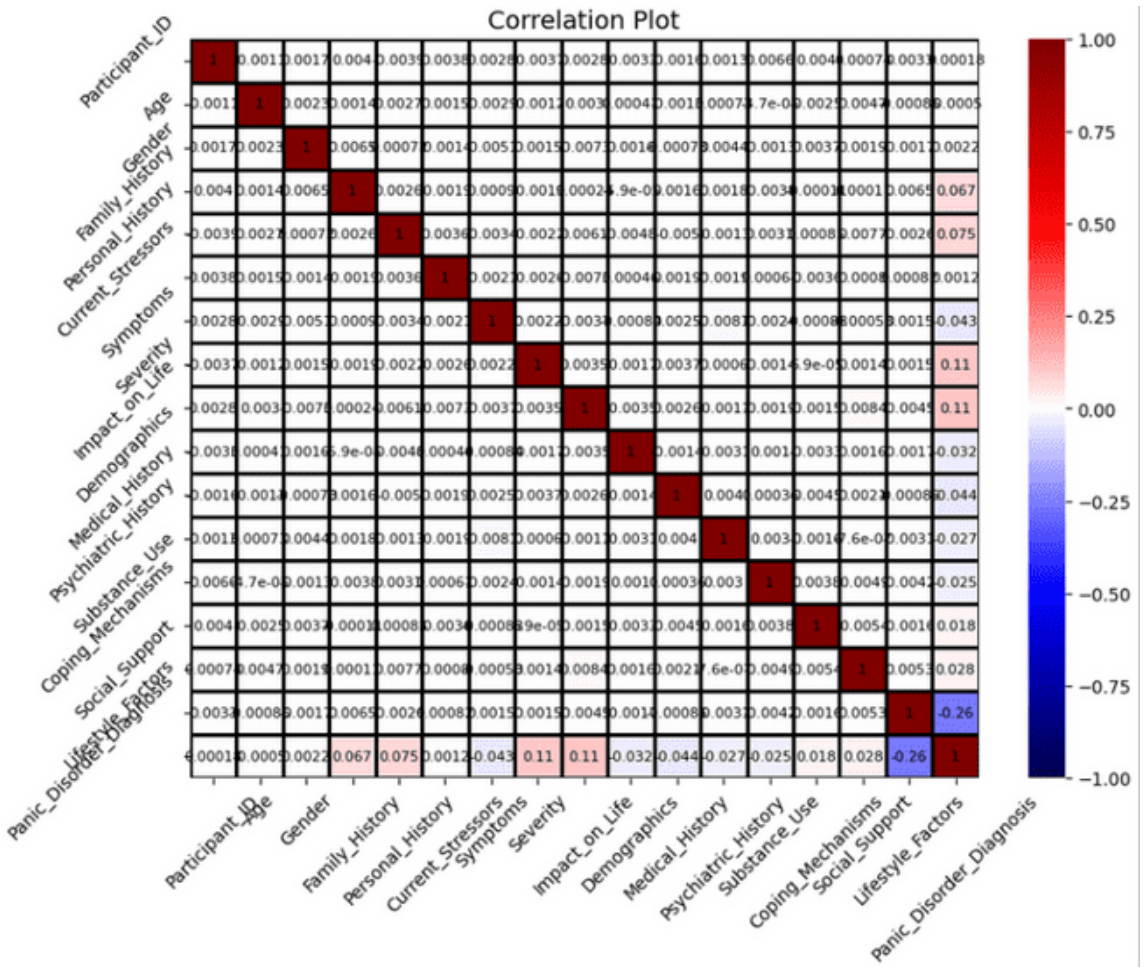

Figure 7 Data Correlation