Executive Summary

This report aims to predict the chances of major depressive disorder (MDD) relapses in depressed patients using statistical models. The data used in this analysis was obtained from the largest healthcare network in Thailand, Bangkok Dusit Medical Services, and consists of 397 patient records after data filtering.

The report begins by providing an overview of the data, including the variables collected such as age, gender, family history, precipitating factors, co-morbidities, symptoms, duration of episode, time to first and last relapse, number of relapses, and treatment. Multicollinearity analysis is also conducted to identify any highly correlated variables in the dataset.

Three different models are utilized to predict MDD relapses: logistic regression, decision tree, and random forest. Initially, all features in the dataset were analyzed using these models. Subsequently, various combinations of five features were explored in order to enhance the performance of the models.

The best combination of predictors identified by logistic regression includes age, gender, family history, precipitating factors, and co-morbidities with the accuracy of 82.5%. The decision tree model achieved the highest accuracy of 92.5% using the combination of age, precipitating factors, decreased activities, hopelessness, and difficulty sleeping. The random forest model achieved an overall accuracy of 85% on the test data, with age being the most important predictor.

Based on the findings, the report suggests that age, precipitating factors, and difficulty sleeping are statistically significant predictors of MDD relapse. These insights can help clinicians identify patients at higher risk and tailor their treatment plans accordingly.

Introduction

MDD is a common mental disorder with a high rate of relapse, which can significantly impact an individual's quality of life and increase the risk of other mental and physical health problems. Developing reliable and valid models to predict MDD relapse can help clinicians identify patients who are at higher risk and may benefit from more intensive treatment or closer monitoring.

Furthermore, understanding the risk factors and predictors of MDD relapse can also inform the development of preventive interventions and personalized treatment plans. Therefore, researching the topic of predicting MDD relapses can have significant implications for both clinical practice and public health.

Relevant Work

1. *"Prediction of Probable Major Depressive Disorder in the Taiwan Biobank: An Integrated Machine Learning and Genome-Wide Analysis Approach" by Lin et al. (2021)*

   Data Used: The study utilized data from the Taiwan Biobank, a large-scale population-based cohort study. The dataset included genetic information, clinical data, and self-reported questionnaire responses. The sample consisted of individuals without a prior diagnosis of major depressive disorder (MDD).

   Data Preprocessing: Data preprocessing involves quality control measures for genetic data, including filtering for variants, imputation of missing genotypes, and adjustment for population structure. Clinical data underwent cleaning and standardization processes. Feature engineering was performed to extract relevant information from the questionnaire responses.

   Models Considered and Used: The models include logistic regression, random forest, gradient boosting, and deep neural networks.

   Metrics for Performance Evaluation: The performance of each model was evaluated using metrics such as accuracy, sensitivity, specificity, and area under the curve (AUC).

   Evaluation and Results: The deep neural networks model achieved the highest accuracy, sensitivity, specificity, and AUC-ROC compared to the other models. The AUC-ROC values for all models indicate good discriminative ability.

   Implications for the Project: This paper serves as a benchmark for evaluating the accuracy of my Logistic Regression and Random Forest models. Because the data input in this paper is drastically different than my data set, I decided to use different models to better explore the relationships between my data features. However, I decided to use the same performance metrics in my work both because of the thoroughness of the evaluation and for ease of comparison of Logistic Regression and Random Forest models.

2. *"Predictors of recurrence of major depressive disorder" by Lye et al. (2020)*

   Data Used: The study included 370 patients with a history of Major Depressive Disorder (MDD), and data were obtained from medical records and interviews.

   Data Preprocessing: Missing values and outliers were removed, and relevant features were selected for analysis.

   Models Considered and Used: Logistic regression, decision trees, and support vector machines were employed as learning models to predict MDD recurrence.

   Metrics for Performance Evaluation: The performance of each model was evaluated using metrics such as accuracy, sensitivity, specificity, and area under the curve (AUC).

   Evaluation and Results: The decision tree model achieved the highest accuracy (76.8%) and AUC (0.75), while the support vector machine model had the highest sensitivity (80.5%). The number of previous episodes, duration of the most recent episode, and comorbid anxiety disorders were identified as the most important predictors of MDD recurrence.

   Implications for the Project: This paper also serves as a benchmark for the decision tree model. Note that the model predicts the MDD recurrence, not relapse. However, due to the limited research in this field, I deemed the model from this paper to be the best benchmark for the decision tree approach. In addition, this paper provides valuable insights into data preprocessing that I leveraged in my own work to improve the predictive accuracy of my models.

3. *"Risk factors for relapse and recurrence of depression in adults and how they operate: A four-phase systematic review and meta-synthesis" by Buckman et al. (2018)*

   Data Used: The paper conducted a systematic review and meta-synthesis, collecting data from multiple studies that investigated risk factors for relapse and recurrence of depression in adults.

   Data Preprocessing: The included studies were assessed for quality, and a meta-ethnographic approach was used to synthesize the data and identify themes and concepts related to risk factors.

   Models Considered and Used: As a systematic review, this paper did not involve the development or application of specific models.

   Metrics for Performance Evaluation: The test error of the model, which represents the accuracy of predictions on the test set, is 0.17500000000000004. This indicates that the model has an error rate of approximately 17.5% on the test data.

   As a systematic review, this paper focused on synthesizing existing research findings rather than evaluating model performance.

   Evaluation and Results: The paper identified several risk factors for relapse and recurrence of depression in adults, including a history of childhood maltreatment, previous depressive episodes, residual symptoms, poor social support, and a lack of psychotherapy. The study also highlighted the interaction between these risk factors and the increased likelihood of relapse or recurrence when multiple factors were present.

   Implications for the Project: I incorporated the identified prognostic risk factors into my statistical models to enhance their predictive accuracy. I decided to exclude any time-related factors (Duration of episode, Time to first relapse, Time to last relapse) from the data set as the paper concludes the lack of its effect on the chance of MDD relapse. I also simplified the models to predict the Boolean values instead of the number of relapses since the paper did not find evidence of the association.

Data

The raw data consists of 406 patient records obtained from Bangkok Dusit Medical Services, the largest healthcare network in Thailand. After dropping data that doesn't meet the criteria, the dataset is reduced to 397 records.

The variables collected include:

1. Age
2. Gender
3. Family history
4. Precipitating factors
5. Co-morbidities
6. Symptoms (points 6-14 in the preprocessed data)
7. Duration of episode
8. Time to first relapse
9. Time to last relapse
10. Number of relapses
11. Treatment

To conduct the analysis, I preprocessed the data and included the following factors:

1. Age
2. Gender
3. Family History
4. Precipitating Factors
5. Co-morbidities
6. Loss of Interest
7. Decreased Activities
8. Poor Concentration
9. Excessive Guilt
10. Hopelessness
11. Thoughts of Suicide
12. Difficulty Sleeping
13. Loss of Appetite
14. Fatigue
15. Severity
16. Treatment
17. Relapse

Regarding the multicollinearity result, Variance Inflation Factor (VIF) values were calculated for each feature. It appears that most of the features have low VIF values, indicating low multicollinearity. However, there are a few features with relatively higher VIF values:

- Loss of Appetite (VIF: 1.93)
- Fatigue (VIF: 1.71)
- Severity (VIF: 2.18)
- Poor Concentration (VIF: 1.66)
- Difficulty Sleeping (VIF: 1.65)

These VIF values suggest that there may be some correlation between these features, but the multicollinearity is not severe. Thus, no variables were dropped from the data set.

Data Analysis

Three different models are utilized to predict MDD relapses including logistic regression, decision tree, and random forest. The libraries used include pandas, numpy, statsmodels, sklearn and matplotlib.

At a baseline where all features from the preprocessed data set are included, the three models perform as the followings:

1. **Logistic Regression**

   The model achieved convergence successfully, indicating that the optimization process reached a stable solution.

   The pseudo-R-squared value of 0.07716 suggests that the model explains approximately 7.7% of the variation in the dependent variable. The overall significance of the model is determined by the LLR (Likelihood Ratio) p-value, which is 0.02736. This suggests that the model is statistically significant at a significance level of 0.05. The test error of the model, which represents the accuracy of predictions on the test set, is 0.17500000000000004. This indicates that the model has an error rate of approximately 17.5% on the test data.

   Based on the significance of the predictors, the following predictors are statistically significant at a significance level of 0.05: 'precipitating factors', 'difficulty sleeping', and 'treatment'.

   Performance Evaluation:

   Compared to the performance of the benchmark model from *Lin et al. (2021)* as provided,

Metrics	My Model	Lin et al
Accuracy	0.85	0.72
Sensitivity	1.0	0.68
Specificity	0.14	0.75
AUC-ROC	0.57	0.78

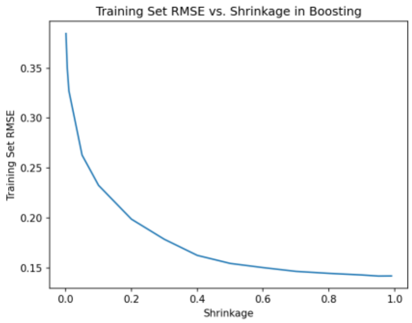
   My model performs better in terms of overall accuracy and sensitivity. However, the benchmark model outperforms in terms of specificity and AUC-ROC, indicating better performance in capturing negative cases and overall predictive discrimination.

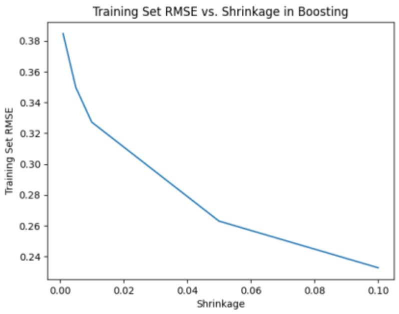
2. Decision Tree

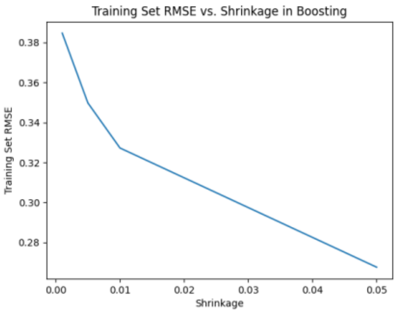
After exploring various shrinkage values and sizes, the most accurate decision tree model is achieved through the shrinkage size of five values including 0.001, 0.005, 0.01, 0.05, 0.1.

The performances of decision tree models with different shrinkage values and sizes are as follows:

Size	15	5	4
Shrinkage Values	0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99	0.001, 0.005, 0.01, 0.05, 0.1	0.001, 0.005, 0.01, 0.05
Accuracy	0.75	0.8	0.725
Sensitivity	0.82	0.88	0.82
Specificity	0.43	0.43	0.29
AUC-ROC	0.62	0.65	0.55







From the results, smaller shrinkage values seem to lead to better accuracy, sensitivity, and AUC-ROC compared to larger shrinkage value, indicating finer adjustments to the model’s predictions based on the gradient information from each individual tree.

Performance Evaluation:

Comparing my most accurate model with the shrinkage size of five values to the performance of the benchmark model from *Lye et al. (2018)* as provided

Metrics	My Model	Lye et al
Accuracy	0.8	0.77
Sensitivity	0.88	0.67
Specificity	0.43	0.74
AUC-ROC	0.65	0.75

My model, again, performs better in terms of accuracy and sensitivity, while Lye et al.'s model performs better in terms of specificity and AUC-ROC.

3. Random Forest

The feature importance analysis highlights the importance of features such as age, poor concentration, gender, treatment, and difficulty sleeping in predicting the target variable.

Precision:

- For the '0' class (no relapse), the precision is 50%, indicating that out of the instances predicted as '0', only half of them is actually '0'.
- For the '1' class (relapse), the precision is 86%, indicating that out of the instances predicted as '1', 86% of them are actually '1'.

Recall:

- For the '0' class, the recall is 29%, indicating that out of all the actual instances of '0', only 29% of them are correctly predicted as '0'.
- For the '1' class, the recall is 94%, indicating that out of all the actual instances of '1', 94% of them are correctly predicted as '1'.

F1-Score:

- The F1-score is a harmonic means of precision and recall and provides a balanced measure of the model's performance. The F1-score for the '0' class is 0.36, and for the '1' class, it is 0.90.

Support:

- The support indicates the number of instances for each class in the test set. In this case, there are 7 instances of class '0' and 33 instances of class '1'.

Performance Evaluation:

Compared to the performance of the benchmark model from *Lin et al. (2021)* as provided,

Metrics	My Model	Lin et al
Accuracy	0.85	0.73
Sensitivity	0.97	0.65
Specificity	0.29	0.77
AUC-ROC	0.63	0.80

My model, again, performs better in terms of accuracy and sensitivity, while Lin et al.'s model performs better in terms of specificity and AUC-ROC.

There seems to be a clear pattern across the three models. My model consistently demonstrates superior performance in accuracy and sensitivity compared to the benchmark models. These results indicate that my model excels at correctly classifying samples and accurately identifying positive instances. However, it is worth noting that the benchmark models exhibit better performance in terms of specificity and AUC-ROC. This suggests that the benchmark models are more effective at correctly identifying negative instances and have better overall discriminative power. Nonetheless, the consistent outperformance of my model in accuracy and sensitivity highlights its strength in accurately capturing positive cases.

After evaluating the baseline approach, I explored the best combination of features for each model. The feature size is fixed at five variables. The best combinations are determined through the highest accuracy score for each model.

1. **Logistic Regression**

   The best combination of features for my logistic regression model is:

   - Age
   - Gender
   - Family history
   - Precipitating factors
   - Co-morbidities

   Although the specific features identified as statistically significant differ from the baseline implementation, the overall performance of the model has been enhanced. While the model excels in accuracy and sensitivity, it acknowledges the need for further improvement in specificity and AUC-ROC to surpass the benchmark model.

Metrics	Baseline	Improved Model	Lye et al
Accuracy	0.85	0.83	0.77
Sensitivity	1.0	1.0	0.67
Specificity	0.14	0.0	0.74
AUC-ROC	0.57	0.5	0.75

2. **Decision Tree**

   The best combination of predictors includes:

   - Age
   - Precipitating factors
   - Decreased activities
   - Hopelessness
   - Difficulty sleeping.

Metrics	Baseline	Improved Model	Lye et al
Accuracy	0.8	0.93	0.77
Sensitivity	0.88	1.0	0.67
Specificity	0.43	0.57	0.74
AUC-ROC	0.65	0.76	0.75

The model incorporating feature selection demonstrates remarkable improvements across all performance metrics compared to the baseline. Notably, it surpasses the benchmark model in all aspects except for specificity, where it remains competitive. Most notably, the accuracy achieves a remarkable increase from 80 to 93 percent, showcasing the model's exceptional capability for accurate predictions.

3. Random Forest

The best combination for the random forest model includes:

- Age
- Poor concentration
- Gender
- Treatment
- Difficulty sleeping.

Metrics	Baseline	Improved Model	Lye et al
Accuracy	0.85	0.85	0.77
Sensitivity	0.97	0.97	0.67
Specificity	0.29	0.29	0.74
AUC-ROC	0.63	0.63	0.75

The random forest model demonstrated the importance of the same set of features as the baseline model, resulting in comparable performance. While the model achieved same accuracy and sensitivity which are better than the benchmark, it is important to note that the specificity and AUC-ROC metrics did not surpass the benchmark model. This finding highlights the need for further enhancements in order to achieve superior specificity and AUC-ROC performance.

Overall, the performance of the improved models surpassed the benchmark models in accuracy and sensitivity. However, most of the models still do not surpass the benchmarked specificity and AUC-ROC. This indicates that although the models showed strengths in correctly classifying positive instances and overall accuracy, they struggled with accurately identifying negative instances and achieving a high discrimination capability. This indicates that there is room for further refinement and optimization of the models to improve their performance in terms of specificity and AUC-ROC.

Findings

Among the evaluated parameters, namely accuracy, sensitivity, specificity, and area under the ROC curve, the decision tree model consistently outperformed the other models. Notably, the random forest model yielded identical results in both approaches, further validating its performance.

One intriguing finding is the recurring presence of age across all three models. Additionally, gender, precipitating factors, and difficulty sleeping emerged as significant factors in two out of the three models. It is important to note that previous research studies have not strongly supported the association of age and gender with MDD relapses, although some studies suggest that females may be more susceptible to MDD than males.

While further analysis could be pursued to enhance the accuracy and reliability of the work, it is noteworthy that the current study provides valuable insights to the MDD community. The findings can serve as a foundation for future adaptations and applications, potentially contributing to the advancement of knowledge in this field.

Reflections

Throughout the project, significant emphasis was placed on meticulous ideation and model selection processes, which proved to be time-consuming. Extensive research was conducted to gain a comprehensive understanding of the subject matter and existing literature. In addition to the cited relevant work, five additional studies were carefully examined to establish a strong foundation for the project.

One of the challenges encountered was the need for medical domain knowledge to fully comprehend the raw data, establish the project's fundamentals, and conduct data preprocessing for optimal feature selection. Discussions with medical professionals were undertaken to ensure the inclusion of important issues and factors, addressing them appropriately in the project.

One trade-off was the selection of feature sets. While efforts were made to identify the most relevant features, there were instances where the features selected did not align with those of the benchmark model. However, I believe that this trade-off was deemed necessary to explore alternative combinations that could potentially improve accuracy and sensitivity.

To arrive at the final solution, a rigorous model selection process was conducted. Extensive research and analysis were undertaken to understand the subject matter and existing work in the field. The three models including logistic regression, random forest, and decision tree, were chosen based on their potential to uncover relationships between the features and the target variable.

The implementation of the selected models and subsequent performance analysis proved to be straightforward, as the primary objective was to compare their performance against the benchmark. This streamlined approach allowed for a focused evaluation and insightful comparisons. Nonetheless, one main challenge was the limited improvement in specificity and AUC-ROC metrics across the models, highlighting the difficulty in achieving a well-rounded performance across all evaluation metrics.

Data Limitations

Data size was one of the main concerns in this project. While the smaller size eases the data preprocessing as it is easier to check for edge cases and error, the insufficient sample size can be one of the factors that inhibit the model performances.

In addition, the features from the obtained data set does not match with what were considered in the work by *Buckman et al. (2018).* Some of the most supportive factors determined by Buckman such as neuroticism, psychosocial impairment and coping style, childhood maltreatment and residual symptoms are not included in the data set. In theory, having these features will increase the accuracy of the models which is something worth exploring.

Next Steps

1. **Fine-tuning the Models**
   I would conduct further analysis and evaluation identify areas for improvement. I would consider parameter tuning, feature engineering and trying out various combinations of feature sets, rather than limiting the selection to just five features per set.

2. **Considering feature selections based on other parameters**

   I would want to analyze the variations in feature selection based on parameters other than model accuracy to explore the relationship between different parameters and feature sets and identify potential patterns or correlations.

3. **Exploring other machine learning algorithms**

   It may also be worthwhile to explore other machine learning algorithms such as gradient boosting and deep neural networks, as demonstrated by Lye et al. (2021). The expansion will allow better assessment of the model suitability for the given dataset and problem.

4. **Increasing Sample Size**

   I would contact BDMS to obtain a larger sample size for future analysis. With a larger dataset, I would be able to assess whether the models' performance improves and whether the specificity and AUC-ROC metrics show enhancements

References

Buckman, J. E. J., Underwood, A., Clarke, K., Saunders, R., Hollon, S. D., Fearon, P., & Pilling, S. (2018). Risk factors for relapse and recurrence of depression in adults and how they operate: A Four-phase systematic review and meta-synthesis. Clinical Psychology

Lin, E., Kuo, P.-H., Lin, W.-Y., Liu, Y.-L., Yang, A. C., & Tsai, S.-J. (2021). Prediction of probable major depressive disorder in the taiwan biobank: An integrated machine learning and genome-wide analysis approach. Journal of Personalized Medicine, 11(7), 597.

Lye, M.-S., Tey, Y.-Y., Tor, Y.-S., Shahabudin, A. F., Ibrahim, N., Ling, K.-H., Stanslas, J., Loh, S.-P., Rosli, R., Lokman, K. A., Badamasi, I. M., Faris-Aldoghachi, A., & Abdul Razak, N. A. (2020). Predictors of recurrence of major depressive disorder. PLOS ONE, 15(3). https://doi.org/10.1371/journal.pone.0230363