RUTGERS

APPLIED DATA MINING AND MACHINE LEARNING

FINAL PROJECT REPORT

Early Detection of Adolescent Depression: Applying Machine Learning Techniques to Enhance Diagnostic Accuracy and Intervention Strategies

NAME: Mohammed Nihil Puthiya Kottal
INSTRUCTOR: DR. I-MING CHIU

ID: 227001556

NetID: mnp126

ABSTRACT

We aim to use Machine learning techniques such as Random Forest and Logistic Regression to identify the MDESI cases for multiple age groups ranging between 12-17years. This data is a sample subset data from the pool of entries who participated in the National Survey on Drug Use and Health (NSDUH), 2011-2017. Using findings from the logistic and Random Forest model, the goal was to find to identify potential MDESI cases. The logistic predictive model successfully identified 72.07% of the MDESI cases (recall rate) and accurately identified 69.4% of the MDESI cases and Non MDESI cases (accuracy rate) in the test dataset. Whereas the random forest model successfully identified 72.88% of the MDESI cases (Recall rate) and accurately identified 69.4% of the MDESI cases and Non MDESI cases (accuracy rate) in the test dataset. Results from this study confirmed that both logistic and random forest model when used as a classifier, can identify potential MDESI cases with acceptable recall and reasonable accuracy rates.

INTRODUCTION

Depression ranks as the most prevalent mental health disorder worldwide, with annual global prevalence rates varying between 7% and 21% [1]. This condition can severely degrade quality of life, as it stands as the second leading cause of Disability-Adjusted Life Years (DALYs) and Years Lived with Disability (YLDs) [2, 3]. Depression also plays a critical role in contributing to suicide, impacting hundreds of thousands of cases annually [4, 5]. Beyond its profound personal and social effects, depression also imposes considerable economic costs.

Considering mental health care pathways, early diagnosis can offer significant benefits to patients by enabling timely interventions. For instance, Bohlmeijer et al. [6] found that patients who participated in acceptance and commitment therapy (ACT) as an early intervention experienced fewer symptoms of depression than those placed on a waitlist, with improvements noted both immediately and at a follow-up three months later. Additionally, a meta-analysis by Davey and McGorry [7] indicated a 20% decrease in depression incidence within 3 to 24 months after an early intervention. Conversely, delayed diagnoses of depression often lead to prolonged suffering in terms of symptom severity and disease progression, as well as increased use of healthcare resources [8, 9].

In recent years, the importance of using predictive models to identify individuals at risk of developing or currently having an undiagnosed mental health disorder has gained recognition among researchers and healthcare providers. Several recent studies have utilized machine learning to analyze predictors and levels of depressive symptoms.

However, none of these studies have developed a predictive model specifically aimed at identifying adolescents at a higher risk for major depression, particularly those with severe functional impairment related to depression who require more intensive, costly care. This study seeks to address this significant research gap by constructing a predictive model that employs machine learning techniques to detect severe functional impairment in adolescents with depression.

LITERATURE REVIEW

Over the years, significant efforts have been made to enhance the performance of machine learning classifiers across various domains. This review consolidates insights from several research papers focusing on the optimization and evaluation of classifier models.

Handoyo et al. [10] present a methodology for optimizing logistic regression and linear discriminant classifiers by exploring different threshold values. By leveraging performance metrics such as confusion matrix, precision-recall, F1 score, and ROC curve [16], the study identifies the importance of threshold tuning in achieving the best classifier performance. Notably, the implementation of principal component analysis (PCA) for dimensionality reduction underscores the significance of preprocessing steps in model development.

Probst et al. [11] emphasize the critical role of selecting an appropriate number of trees in random forest classifiers. Through analysis of the out-of-bag (OOB) error rate curve, the study highlights the significance of optimizing the number of trees to mitigate overfitting and improve classifier generalization. This approach contributes to enhancing the robustness and accuracy of random forest models.

Feature selection emerges as a pivotal strategy for improving classifier performance, particularly in scenarios involving imbalanced datasets. The weighted Gini index (WGI) feature selection method, proposed by [12], aims to identify a subset of features that optimize ROC AUC and F-measure results, particularly focusing on enhancing performance on minority classes. This underscores the importance of tailored feature selection techniques in addressing class imbalances and maximizing classifier effectiveness.

Receiver operating characteristic (ROC) curve analysis remains a cornerstone in evaluating the diagnostic ability of classifier models. As highlighted by [13], the ROC curve, coupled with the area under the curve (AUC), offers a robust measure of accuracy, facilitating the identification of optimal cutoff values and comparison between alternative diagnostic tasks. This underscores the pivotal role of ROC analysis in assessing classifier performance and informing decision-making processes in various applications.

Furthermore, the literature underscores the importance of considering application-specific objectives in classifier optimization. As noted by [14], the trade-off between

precision and recall should be carefully considered based on the specific requirements of the application. This highlights the need for a nuanced understanding of performance metrics and their implications in real-world contexts. [15] delves into the optimization of random forest and logistic regression models through appropriate tuning strategies. The study emphasizes the significance of expert-level expertise in leveraging tuning parameters effectively to enhance model performance. This underscores the importance of domain knowledge and expertise in maximizing the efficacy of machine learning algorithms. In summary, this literature review highlights the multifaceted strategies employed to optimize and evaluate classifier models. From threshold tuning and feature selection to OOB error rate analysis and application-driven performance considerations, the reviewed studies offer valuable insights into enhancing the effectiveness and applicability of machine learning classifiers across diverse domains.

METHODOLOGY

In our study, we employed two distinct machine learning algorithms, namely random forest, and logistic regression, to develop classification models for our analysis.

Random Forest is an ensemble learning method for classification and regression tasks. It constructs a multitude of decision trees during training and outputs the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. It improves accuracy and reduces overfitting by averaging the predictions of multiple decision trees.

Logistic Regression is a statistical model used for binary classification tasks. It predicts the probability of occurrence of an event by fitting data to a logistic function. Despite its name, it is primarily used for classification rather than regression. It estimates the probability that a given input belongs to a particular category based on its features, using a logistic (sigmoid) function to map input features to probabilities.

Data Collection

The dataset utilized in this study was sourced from a comprehensive database containing personal and socio-economic information of adolescents over several years. Special attention was given to variables such as parental influence, gender, age, and various socioeconomic factors.

Model Development

Two primary predictive models were developed:

1. **Random Forest Model**

- **Parameter Selection**: Initially, the model was set with default parameters, 500 ntree and square root of the count of variables as mtry.
- **Model Tuning:** Using Out-of-Bag (OOB) error rates, the model's parameters were fine-tuned. The number of trees was optimized based on the stabilization of OOB error at 700 trees, indicating diminishing returns on model performance

with additional trees. The optimal mtry value was determined to be 2, as it provided the lowest OOB error.

- **Feature Selection and Engineering:** Features were selected based on their importance, indicated by the Gini index and Chi-Square test values, and their practical relevance to the predictive goal. Feature engineering was employed to enhance the model's predictive accuracy.

2. **Logistic Regression Model**

- **Model Setup**: Logistic regression was applied using the same set of variables finalized for the Random Forest model to ensure comparability.
- **Threshold Optimization:** The default classification threshold of 0.5 was adjusted based on ROC curve and AUC analysis to fine-tune the balance between sensitivity and specificity.

Model Evaluation

- **Recall Optimization:** For both Random Forest and Logistic model the primary metric for model performance was recall, given the clinical priority to maximize identification of successful depression cases.
- **ROC and AUC Analysis**: The Logistic model and Random Forest models were compared using ROC curves to evaluate the trade-off between true positive and false positive rates under different threshold settings.
- **Threshold Adjustment**: Based on the desired recall rate (targeted at 75%), the threshold θ for logistic model was adjusted to maximize the identification of depressive cases while controlling for an acceptable increase in false positives.

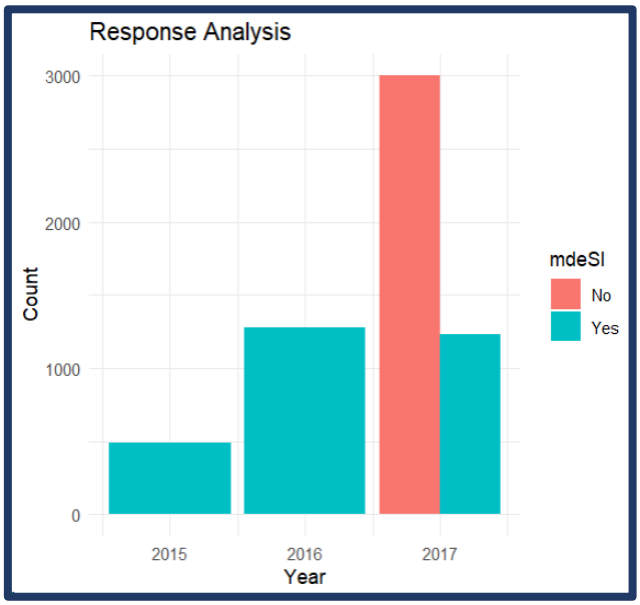
Implementation

- **Software and Tools**: The analysis was performed using R programming language, which is well-suited for statistical modelling and machine learning applications. Relevant packages such as randomForest for Random Forest model, pROC for ROC curve analysis were utilized and ggplot2 for the visualization.

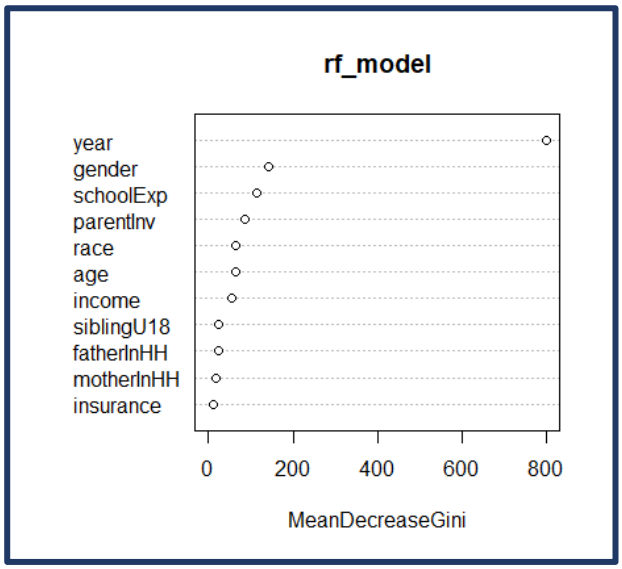
This methodology ensured a robust approach to developing and evaluating predictive models for identifying severe depressive symptoms in adolescents, with an emphasis on model recall and the practical applicability of predictions in a clinical setting.

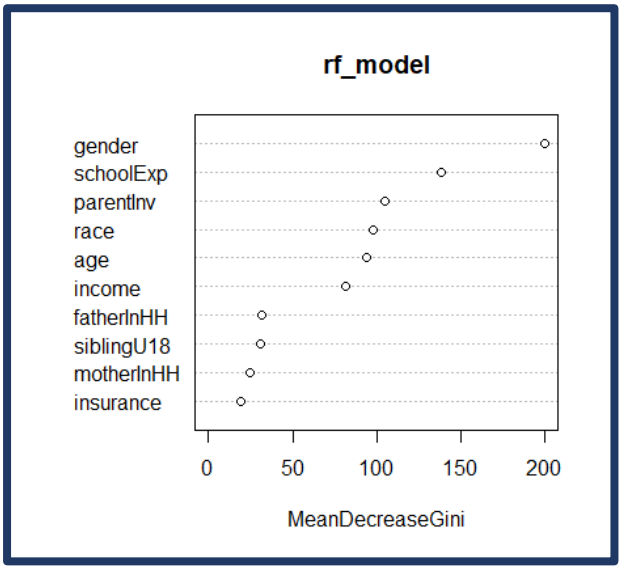
FINDINGS

I analysed the sample dataset and discovered that all recorded cases from the year 2015-2016 were marked as 'Yes' for depression (MEDSI), with no instances of 'No'. When we check the Gini index for the variable ‘year’ it clearly shows the significance the variable has in the model. ***This anomaly is further proven when we check model's accuracy, the accuracy increases to 82% from 69% when the 'year' variable is incorporated into the machine learning model (Random Forest).***



The Gini index value for the variable ‘year’ when compared to other variables is shown in the plot below.





The plots above shows the value of gini index of each variable after adding and removing the ‘year’ variable. This clearly shows that the ‘year’ variable plays a significant role in the model and the decision to include the ‘year’ variable should be decided purely on how the data is collected. ***Since it is mentioned in our problem statement that the variable ‘year’ signifies in which year the data was collected we shall not include it in our analysis.***

Also, the response of each gender also requires some attention as the ratio of males to females having depression is skewed and this can lead to some bias in our model.

	No	Yes
Male	1702	751
Female	1298	2249

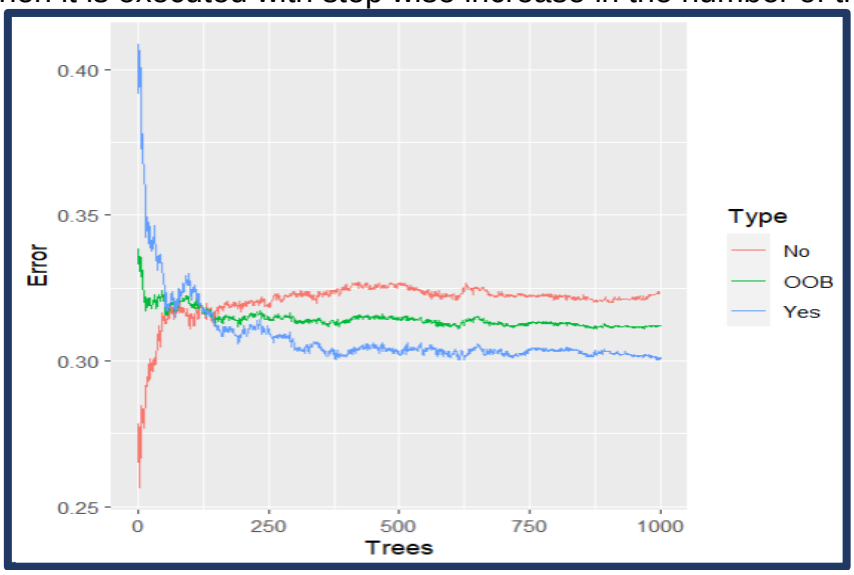
Model 1: Random Forest

We are running the base model with all the variables (**Except year**) and with default parameter values. (number of trees = 500, mtry = $\sqrt{No.of\ variables}$)

```
Accuracy, recall, and precision for the base model is: 0.689333, 0.708447, 0.673575
```

Since we need to select the model which identifies depressive cases more successfully, we shall concentrate on the recall value. ***The recall value for the base model is 70.8%.***

Now to decide on the optimum number of trees, we shall find the OOB error rate for the model when it is executed with step wise increase in the number of trees.



The OOB error rate is initially high but as the number of trees increases the OOB error rate reduces and is stable after 700 signifying that adding more trees doesn't necessarily increase the performance. Hence, we shall take the number of trees as 700.

Now to find the optimum mtry value, lets apply the same method and check for what number of variables we have the least OOB error. From the R code we can see that the model with parameter mtry = 2 has the least OOB error.

Now let's run the model again with the updated parameters and check if there is any increase in performance:

```
Accuracy, recall, and precision for the base model is: 0.688000, 0.712534, 0.670513
```

There is an increase in the model performance as recall has increased to **71.25%.**

Now after performing feature selection and feature engineering, we were able to increase the performance of the model and its mentioned below.

```
Accuracy, recall, and precision for the base model is: 0.694000, 0.728883, 0.672956
```

recall for the model with feature selection done has increased to **72.88%.**

Model 2: Logistic Regression

We have applied the logistic regression to the same set of variables as used in the final model of Random Forest, the performance metric is shown below.

```
Accuracy, recall, and precision for the logistic model is: 0.694667, 0.720708, 0.676471
```

Threshold Adjustment

We will now analyse the threshold values and try to increase our sensitivity (True positive rate) cases.

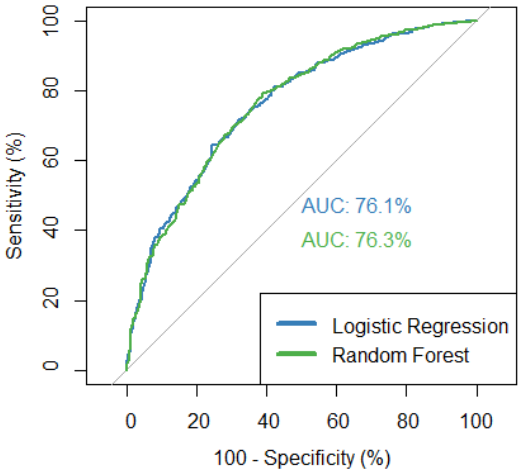
From the R output (roc_curve_lr.df), we have the true positive rate (TPP) and false positive rate (FPP) for each threshold sequence. Given our goal of more effectively identifying depressive cases, we aim to lower the threshold. This adjustment is expected to increase the true positive rate while allowing for an acceptable rise in the false positive rate.

In our logistic regression model, we have achieved a recall rate of 72.07%. To increase the recall rate to 75%, we reviewed the R output to identify threshold values corresponding to TPPs in the vicinity of 75%. We will apply this threshold in our calculations shown below and recompute the accuracy and recall values using the confusion matrix.

```
> (accuracy.test = (CM.test[1,1] + CM.test[2,2])/1500)
[1] 0.694
> (recall.test = CM.test[2,2]/(CM.test[1,2] + CM.test[2,2]))
[1] 0.7506812
> (precision.test = CM.test[2,2]/(CM.test[2,1] + CM.test[2,2]))
[1] 0.6662636
```

TPP	*FPP*	*Threshold*
75.61308	62.14099	***0.4662***
75.61308	62.27154	***0.4673***
75.61308	62.40209	***0.4675***
75.34060	63.31593	***0.4678***
75.34060	63.57702	***0.4692***
75.20436	63.70757	***0.4717***
75.06812	63.70757	***0.4746***
75.06812	63.96867	***0.4771***
74.93188	63.96867	***0.4793***
74.93188	64.09922	***0.4808***

Table 1:TPP-FPP values for different threshold setting



Combined ROC curves of the Logistic Regression and Random Forest models

In the above plot, the AUC value for both Logistic regression and Random Forest is almost the same. Also, the TPP and FPP spread is similar.

DISCUSSION

How you decide the number of trees and the subset of features when generating decision trees?

There are multiple ways to decide the number of trees and subset of features like Grid Search, Cross Validation, Visualization of performance metrics against different 'nTree' and 'mtry' values and analysing the 'OOB' error values.

For our analysis I have opted to check OOB error values for different 'nTree' and 'mtry' values.

The Out-of-Bag (OOB) error is an important measure used primarily with Random Forest and other ensemble learning methods that involve bootstrapping (i.e., randomly sampling with replacement). It serves as an internal method of validating the model during training, and it is particularly useful for tuning model parameters.

How it's generated:

- In a Random Forest, each decision tree is trained on a different bootstrap sample from the original dataset. About one-third of the cases are left out of the bootstrap sample and not used to train a given tree.
- These left-out cases, known as the "Out-of-Bag" samples for that tree, are used to estimate the model's performance as if they were a validation or test set.

Calculation:

- After a tree is trained, it is used to predict the responses for its OOB samples. This process is repeated for each tree in the forest.
- The OOB error is then calculated by comparing the actual and predicted values for the OOB samples across all trees, and typically, it is expressed as an error rate for classification or mean squared error for regression.

Use Case in Tuning Parameters:

- Number of trees (nTree): the number of trees in the forest can decrease the OOB error, up to a point. Beyond a certain number, improvements will plateau, providing a way to choose an optimal number of trees without wasting computational resources.
- Number of features considered at each Split (mtry): The OOB error can help determine the best number of features to consider at each split. Trying different values and selecting the one with the lowest OOB error can lead to a more accurate and generalized model.

How you choose an optimal threshold value "θ" to classify individuals to the depression and depression-free cases?

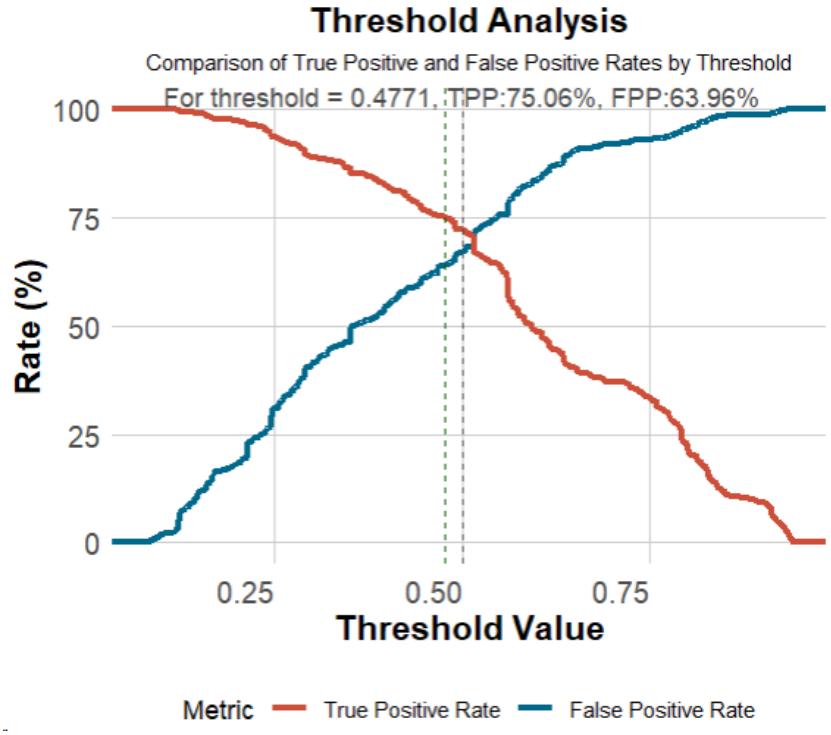
The threshold θ in logistic regression is the probability cut-off used to decide between the two classes. By default, this threshold is typically set at 0.5, meaning that if the predicted probability of depression is greater than or equal to 0.5, the individual is classified as having depression; otherwise, they are classified as depression-free.

- Lowering the threshold increases sensitivity, capturing more true positives but potentially increasing the false positives.

- Raising the threshold increases specificity, reducing false positives but potentially missing true positive cases.

The above-mentioned scenarios can be done with the help of ROC curve and AUC.

- Sensitivity (True Positive Rate) is plotted on the y-axis, and 100-Specificity (False Positive Rate) is on the x-axis, typically shown as increasing from 0 to 100. (*In the plot x & y axes are in %age*)
- The curve shows how TPP and FPP trade off as the decision threshold changes. A perfect classifier would reach the top right corner, with Sensitivity of 100% and a specificity of 100%.
- The grey diagonal line represents a random classifier (AUC = 0.5), where the classifier has no ability to distinguish between positive and negative classes better than random guessing



From the above graph, we can clearly see that for a threshold value of 0.4771 the TPP rate is 75% (as calculated in the threshold adjustment section for logistic regression).

Also, in this analysis we have identified three areas of interest: generalizability (can the model be reused with, e.g., different populations), interpretability (is the model's information readily understandable to its users), and performance (does the model meet the needs e.g. in AUC-ROC, for the purpose for which it is intended) as key components to consider for predictive models of depression built on the use of ML. All three would need careful evaluation before moving from research to a clinical application environment.

CONCLUSION

In this analysis, we investigated the potential of a Random Forest model and a Logistic Regression model for identifying adolescents at risk of severe depressive symptoms, emphasizing the role of the variable 'year' in classification accuracy, with its inclusion improving model accuracy from 69% to 82%. The Gini index values underscored the 'year' variable's impact, suggesting careful consideration of how data is collected and utilized in model training.

The application of machine learning techniques, particularly the tuning of the Random Forest model parameters, such as the number of trees and the number of variables tried at each split (mtry), was guided by Out-of-Bag (OOB) error analysis. This approach helped optimize the model settings, stabilizing the OOB error rate and improving recall from 70.8% to 72.88% after adjustments and feature engineering.

Furthermore, the study explored optimal threshold settings in Logistic Regression to classify individuals accurately into depression and depression-free groups. Adjustments to the threshold were made based on ROC curve and AUC analysis, aiming to enhance the model's sensitivity without compromising specificity excessively. This fine-tuning is crucial for ensuring that the models not only predict depressive cases accurately but also minimize the risk of misclassifying non-depressive cases, thereby optimizing both the clinical and practical utility of the predictive models.

From the combined ROC curve plot that we generated, the performance of both the Random Forest and Logistic Regression models is quite similar, with no significant differences observed.

Overall, the findings suggest that while machine learning models hold promise for early detection of severe depressive symptoms in adolescents, the selection of features, model parameters, and threshold values requires careful consideration to balance accuracy and applicability. The insights gained from this analysis pave the way for further research into refining these models, potentially leading to more effective and timely interventions in mental health care.

REFERENCES

1. Lim GY, Tam WW, Lu Y, Ho CS, Zhang MW, Ho RC. Prevalence of depression in the community from 30 countries between 1994 and 2014. Sci Rep. 2018;8(1):2861
2. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry. 2016;3(2):171–178.
3. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJL, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLOS Med. 2013;10(11):e1001547.

4. Chesney E, Goodwin GM, Fazel S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. World Psychiatry. 2014;13(2):153–160.
5. Organization WH. Depression and other common mental disorders: global health estimates. 2017
6. Bohlmeijer ET, Fledderus M, Rokx TAJJ, Pieterse ME. Efficacy of an early intervention based on acceptance and commitment therapy for adults with depressive symptomatology: evaluation in a randomized controlled trial. Behav Res Ther. 2011;49(1):62–67.
7. Davey CG, McGorry PD. Early intervention for depression in young people: a blind spot in mental health care. Lancet Psychiatry. 2019;6(3):267–272.
8. McGorry PD, Hickie IB, Yung AR, Pantelis C, Jackson HJ. Clinical staging of psychiatric disorders: a heuristic framework for choosing earlier, safer and more effective interventions. Aust N Z J Psychiatry. 2006;40(8):616–622.
9. McGorry P, van Os J. Redeeming diagnosis in psychiatry: timing versus specificity. The Lancet. 2013;381(9863):343–345.
10. Handoyo, Samingun & Chen, Ying-ping & Irianto, Gugus & Widodo, Agus. (2021). The Varying Threshold Values of Logistic Regression and Linear Discriminant for Classifying Fraudulent Firm. Mathematics and Statistics. 9. 135-143. 10.13189/ms.2021.090207.
11. Probst, P., & Boulesteix, A.-L. (2018). To Tune or Not to Tune the Number of Trees in Random Forest. Journal of Machine Learning Research, 18(181), 1–18. http://jmlr.org/papers/v18/17-269.html
12. H. Liu, M. Zhou, X. S. Lu and C. Yao, "Weighted Gini index feature selection method for imbalanced data," 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, China, 2018, pp. 1-6, doi: 10.1109/ICNSC.2018.8361371
13. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013 Spring;4(2):627-35. PMID: 24009950; PMCID: PMC3755824.
14. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, Parasa S. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022 Apr 8;12(1):5979. doi: 10.1038/s41598-022-09954-8. PMID: 35395867; PMCID: PMC8993826.
15. Couronné, R., Probst, P. & Boulesteix, AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* **19**, 270 (2018). https://doi.org/10.1186/s12859-018-2264-5
16. Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/1143844.1143874