**NAME: BERYL ATIENO OCHIENG**

**PROGRAM: MS. DATA SCIENCE**

**RUID: 227007520**

**CAMPUS: THE STATE UNIVERSITY OF NEW JERSEY, RUTGERS**

**TASK: MACHINE LEARNING FINAL PROJECT.**

**INTRODUCTION**

Mental health disorders among adolescents have become a growing concern worldwide, with Major Depressive Disorder (MDD) being one of the most prevalent conditions. Early identification and intervention are crucial for effective management and prevention of adverse outcomes associated with MDD. In this study, we propose the development of two classifiers, logistic regression and random forest, to accurately identify adolescents exhibiting the group largely affected by Major Depressive Episode and Impairment (MDSEI). Previous research has highlighted the challenges in accurately diagnosing and identifying adolescents with MDD due to the complex interplay of biological, psychological, and social factors. Traditional diagnostic methods often rely on clinical interviews and self-reported measures, which may be limited by subjectivity and reporting biases. Machine learning algorithms offer a promising alternative by leveraging patterns in large datasets to predict and classify mental health conditions. Despite advancements in machine learning approaches for mental health diagnosis, there remains a gap in the development of classifiers specifically tailored to identify adolescents with MDSEI. Existing models may lack the sensitivity and specificity required to detect early signs of depression in this population reliably. Moreover, emphasis on recall rates, which measure the ability to capture true positive cases, is paramount in mental health screening to ensure no cases are overlooked.

The primary objective of this project is to develop logistic regression and random forest classifiers optimized for high recall rates in identifying adolescents with MDSEI. By prioritizing recall over accuracy, we aim to minimize false negatives and increase the sensitivity of our models to detect individuals at risk of MDESI. We seek to achieve robust performance in training and test datasets through rigorous training and validation.

Our study utilizes modified data from the Department of Economics at Rutgers University that was pooled from the National Survey on Drug Use and Health (NSDUH) from 2011 to 2017

a diverse sample of adolescents. Feature selection and preprocessing techniques will enhance model performance and generalizability. We will train and validate logistic regression and random forest classifiers using cross-validation techniques to ensure robustness and mitigate overfitting.

The successful development of classifiers with high recall rates for MDSEI detection holds significant implications for early intervention and prevention efforts in adolescent mental health. By accurately identifying individuals at risk of MDD, healthcare professionals can provide timely interventions and support, ultimately reducing the burden of mental health disorders on individuals and society.

**LITERATURE REVIEW**

Machine learning (ML) methodologies have emerged as valuable tools in mental health research, offering innovative approaches to diagnosis, prognosis, and treatment prediction. In recent years, the application of ML techniques in psychiatric disorders, including Major Depressive Disorder (MDD), has gained significant traction. Studies such as Gulshan et al. (2016) and Dwyer et al. (2018) have demonstrated the utility of ML algorithms in various healthcare domains, underscoring their potential to improve diagnostic accuracy and patient outcomes. The complex and multifactorial nature of MDD necessitates novel approaches for accurate identification and early intervention, making ML particularly well-suited for addressing these challenges.

Accurate identification of adolescents with MDD remains a formidable challenge due to the heterogeneous presentation of symptoms and the subjective nature of traditional diagnostic methods. While commonly used in practice, clinical interviews and self-reported measures may be influenced by biases and variations in reporting. Passos et al. (2016) emphasize the importance of objective clinical signatures for identifying individuals at risk of suicide among those with mood disorders, highlighting the critical need for more reliable and objective screening methods. Furthermore, epidemiological studies such as Kessler et al. (2003) provide valuable insights into the prevalence and burden of MDD among adolescents, reinforcing the urgency of developing robust screening tools.

Despite the potential benefits, several challenges must be addressed when implementing ML-based screening approaches for MDD in adolescents. These include issues related to data quality, feature selection, and model interpretability. Additionally, the ethical and privacy implications of utilizing sensitive health data in ML algorithms require careful consideration. Collaborative efforts between clinicians, researchers, and data scientists are essential to ensure the development and deployment of ethical and effective ML-based screening tools for MDD.

Future research should focus on refining and optimizing ML algorithms for adolescent MDD screening, emphasizing improving sensitivity, specificity, and scalability. Longitudinal studies are needed to assess the predictive validity of ML-based screening tools and their ability to identify individuals at risk of developing MDD over time. Moreover, efforts to enhance the interpretability and transparency of ML models will be essential to foster trust and acceptance among clinicians and patients. By addressing these challenges and advancing our understanding of ML-based screening for MDD, we can pave the way for more effective early intervention and ultimately improve outcomes for adolescents with mental health disorders.

**METHODS**

**Data analysis**

In this project, we developed two classifiers, a logistics classifier and a random forest classifier, to identify adolescents having made with good accuracy and recall rates in both training and test datasets. A logistic regression classifier is a classification technique used in machine learning and is also known as a generalized linear model. It uses a logistic function to model dependent variables. The dependent variable is dichotomous. That is, there could only be two classes. We used binomial logistic regression because the target variable can have only two possible outcomes, and it is used as a classification technique to predict a qualitative response. Installing required packages such as (dryer) and loading the data, then finding the summary of the dataset in the package. Performing logistic regression on the data set: logistic regression was implemented in R using glm() by training the model using features or variables in the data set. We used two packages, “caTools” for logistic regression and “ROCR” for ROC, to evaluate the model. The data will be split into trained regression and test regression, and then the model will be trained using the logistic model, and a summary of the logistic model will be found. Predict the test database on the model. By predicting the regression, changing probabilities, Evaluating model accuracy using confusion matrix, ROC-AUC curves, and plotting the curve.

**Random Forest:** The random Forest algorithm works in several steps. Ensemble of decision trees: Random Forest leverages the power of assembling learning by constructing an army of decision trees. They operate independently, minimizing the risk of the model being overly influenced by the nuances of a single tree. Random Features Selection: During the training of each tree, a random subset of features is chosen. This randomness ensures that each tree focuses on different aspects of the data, fostering a diverse set of predictors within the ensemble. Bootstrap aggregation or bagging involves creating multiple bootstrap samples from the original data set, allowing instances to be sampled with replacement. This results in different subsets of data for each decision tree, introducing variability in the training process and making the model more robust. Decision-making and voting: the average of the individual tree prediction is taken.

The key to building a successful classifier or predictive model depends on Theta's choice, which is the threshold value used to categorize or predict cases in MDESI or MDESI-free states, which is found in the two-by-two confusion matrix below. Which is the performance of the predictive model.

Confusion Matrix

|  |  |  |
| --- | --- | --- |
| **Predict\actual** | **0** | **1** |
| **0** | **TN** | **FN** |
| **1** | **FP** | **TP** |

**True Positive (TP):** This means the individual has depression (mdeSI = Yes), and the model correctly identifies them as having depression**.**

**True Negative (TN):** This means the individual does not have depression (mdeSI = No), and the model correctly identifies them as not having depression.

**False Positive (FP):** This means the individual does not have depression (mdeSI = No), but the model incorrectly identifies them as having depression**.**

**False Negative (FN):** In this scenario, the individual has depression (mdeSI = Yes), but the model incorrectly identifies them as not having depression**.**

**FINDINGS**

The table below summarizes the results of the logistic regression model. The adjusted odds ratio (AOR) identifies the significant features that could help identify adolescents with MDESI. For instance, the risk of having MDESI was higher in female adolescents (AOR = 3.95, p <.01). Likewise, adolescents aged between 14 – 15 and 16 – 17 are more likely to have MDESI (AOR = 1.92, p <.01; AOR = 2.27, p < .01). Additionally, a lower risk of MDESI was observed among the blacks (AOR = 0.59, p < .01). Moreover, having bad school experiences and low parental involvement significantly increased an adolescent’s chance of having MDESI (AOR = 2.95, p <.01; AOR = 2.72, p < .01). Importantly, not having a sibling under the age of 18 also increased the risk of having MDESI (AOR = 1.20, p < .05) The optimal threshold was arrived at by maximizing the recall rate while maintaining the highest accuracy rate (Chiu et al., 2021). Figure 1 below shows that the optimal threshold (θ) is 0.5. Tables 2 A and B show the training and test datasets confusion matrices, accuracy, and recall rates. The recall and accuracy rates were close, indicating no overfitting. Figure 1 for the confusion matrix suggests1 that according to our training data, the recall rate is 72.82%, and the accuracy rate is 69.69%. In contrast, in Figure 2 for test data, the recall rate is 72.07%, and accuracy is 69.47%, indicating both rates are close in training and testing data and using the four-hold cross-validation applied to examine the performance.The threshold value was calculated by running the code;

optimal.threshold <- coords(ROCPred, "best", ret="threshold", transpose=FALSE)

which is 1 0.5019368

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AOR | Lower Bound | Upper Bound |
| Intercept | 0.1535873 | 0.1171002 | 0.2007512 |
| Gender (Female) | 3.9592635 | 3.4515029 | 4.5477843 |
| Age 14 -15 | 1.9280418 | 1.619133 | 2.2981216 |
| Age 16 - 17 | 2.2720814 | 1.905959 | 2.7114163 |
| Race (Hispanic) | 0.9645064 | 0.808560 | 1.1505183 |
| Race (Black) | 0.5948090 | 0.472750 | 0.7471616 |
| raceAsian/NHPIs | 0.7797075 | 0.555110 | 1.0925521 |
| raceOther | 1.1154089 | 0.858824 | 1.4500045 |
| insuranceNo  income20,000 - 49,999 | 1.0022174  1.1360956 | 0.690333  0.916944 | 1.4575839  1.4077537 |
| income50,000 - 74,999 | 1.2006159 | 0.932819 | 1.5458400 |
| income75,000 or more | 0.8695391 | 0.694430 | 1.0886506 |
| fatherInHHno father in hh | 1.0703653 | 0.908628 | 1.2610753 |
| motherInHHno mother in hh | 0.8901878 | 0.701913 | 1.1289376 |
| siblingU18No | 1.2052843 | 1.042043 | 1.3942640 |
| parental involvement | 2.7227663 | 2.286098 | 3.2502843 |
| schoolExpbad | 2.9565892 | 2.533944 | 3.4549298 |

Table 2 **(A) | Confusion Matrix**

|  |  |  |
| --- | --- | --- |
| Prediction for training data  θ = 0.5 (training) | NO(MDESI) | YES(MDESI) |
| NO(MDESI) | 1486 | 748 |
| YES(MDESI) | 616 | 1650 |

Recall rate:72.82%, Accuracy Rate:69.69%

**(B)**

|  |  |  |
| --- | --- | --- |
| Prediction for test data  θ = 0.5 (test) | NO(MDESI) | YES(MDESI) |
| NO(MDESI) | 513 | 253 |
| YES(MDESI) | 205 | 529 |

Accuracy test; 69.47%, Recall rate; 72.07%.

Figure 1: Receiver Operating Characteristic Curve

A graph with a line

Description automatically generated

**From the Random Forest**

OOB estimate of error rate: 31.16%.

Confusion matrix:

No, Yes, class.error

No 1524 710 0.3178156

Yes, 692 1574 0.3053839

OOB No Yes

[1,] 0.3536076 0.3309438. 0.3761905

[2,] 0.3517292 0.2936391 0.4092240

[3,] 0.3441500 0.2918689. 0.3945061

[4,] 0.3437252. 0.3079399 0.3784487

[5,] 0.3378916. 0.2843531 0.3905325

[6,] 0.3431981. 0.3024602. 0.3830893a

**Figure 2:** Plot of error rate against treesA graph of trees and a number of trees

Description automatically generated

**Discussions**

Episode and Impairment (MDSEI) showcases both strengths and areas for improvement. The odds ratios reveal that gender, age, parental involvement and school experience exhibit significant associations with MDSEI. For instance, being female (AOR: 3.96) and having higher levels of parental involvement (AOR: 2.72) are associated with increased odds of MDSEI, while variables like age and race exhibit more nuanced effects. However, it's noteworthy that race, income, and family structure variables show relatively weaker associations with MDSEI, with odds ratios close to 1.0.

The results obtained from the random forest classifier indicate a moderate level of performance in identifying adolescents with Major Depressive Episodes and Impairment (MDSEI). The confusion matrix reveals that out of 2434 instances, 1524 were correctly classified as "No" and 1574 as "Yes," yielding a class error rate of approximately 31.78% for the "No" class and 30.54% for the "Yes" class. While these results demonstrate an acceptable level of accuracy, it's crucial to examine the recall rates, which measure the ability of the classifier to identify positive instances correctly. The out-of-bag (OOB) error rates across iterations provide insights into the stability of the model's performance. While variability is observed in the OOB error rates, indicating potential fluctuations in performance under different conditions, the overall trend suggests a reasonable level of consistency in the classifier's performance.

**Conclusion**

Despite the model's overall accuracy rate of approximately 69.69% for the training dataset and 69.47% for the test dataset, the recall rate, which measures the ability to identify positive instances correctly, stands out as a more critical metric. With a recall rate of 72.82% for the training data and 72.07% for the test data, the logistic regression model demonstrates its efficacy in correctly identifying adolescents with MDSEI. However, it's essential to acknowledge that while the model performs relatively well in terms of recall, there may be instances of false positives and false negatives that require further investigation and refinement. Therefore, while the logistic regression model shows promise in identifying adolescents with MDSEI, ongoing validation, and improvement efforts are necessary to enhance its accuracy and effectiveness in real-world settings.

In conclusion, while the random forest classifier shows promise in accurately identifying adolescents with MDSEI, there is room for improvement, particularly in reducing misclassifications and enhancing recall rates. Strategies like feature selection, parameter tuning, and ensemble model optimization may help improve the classifier's performance. Additionally, further research and validation efforts are warranted to assess the generalizability and scalability of the classifier across diverse populations and settings. By iterative refining and validating the model, we can develop a more robust and accurate tool for identifying adolescents at risk of MDSEI, ultimately facilitating early intervention and improving outcomes in mental health care.

References

Chiu, I.-M., Lu, W., Tian, F., & Hart, D. (2021). Early Detection of Severe Functional Impairment Among Adolescents With Major Depression Using Logistic Classifier. *Frontiers in Public Health*, *8*. <https://doi.org/10.3389/fpubh.2020.622007>

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/s2215-0366(15)00471-x

Chiu, I.-M., Lu, W., Tian, F., & Hart, D. (2021). Early Detection of Severe Functional Impairment Among Adolescents With Major Depression Using Logistic Classifier. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.622007

Hahn, T., Marquand, A. F., Ehlis, A.-C., Dresler, T., Kittel-Schneider, S., Jarczok, T. A., Lesch, K.-P., Jakob, P. M., Janaina Mourao-Miranda, Brammer, M., & Fallgatter, A. J. (2010). Integrating Neurobiological Markers of Depression. *Archives of General Psychiatry*, *68*(4), 361–361. https://doi.org/10.1001/archgenpsychiatry.2010.178

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E., & Wang, P. S. (2003). The Epidemiology of Major Depressive Disorder. *JAMA*, *289*(23), 3095. https://doi.org/10.1001/jama.289.23.3095

Passos, I. C., Mwangi, B., Cao, B., Hamilton, J. E., Wu, M.-J., Zhang, X. Y., Zunta-Soares, G. B., Quevedo, J., Kauer-Sant’Anna, M., Kapczinski, F., & Soares, J. C. (2016). Identifying a clinical signature of suicidality among patients with mood disorders: A pilot study using a machine learning approach. *Journal of Affective Disorders*, *193*, 109–116. https://doi.org/10.1016/j.jad.2015.12.066

Voets, M., Møllersen, K., & Bongo, L. A. (2019). Reproduction study using public data of: Development and validation of a deep learning algorithm for detecting diabetic retinopathy in retinal fundus photographs. *PLOS ONE*, *14*(6), e0217541. https://doi.org/10.1371/journal.pone.0217541

Zhang, L., Wang, M., Liu, M., & Zhang, D. (2020). A Survey on Deep Learning for Neuroimaging-Based Brain Disorder Analysis. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00779