PROJECT REPORT

Analysis Of Depression Among Adolescents using Random Forest and Logistic Regression model

Naman Singh

NS1588

Course – Applied Data Mining & Machine Learning

Instructor – Dr. I-Ming Chiu

**PROJECT REPORT**

“Analysis Of Depression Among Adolescents”

Course – Applied Data Mining & Machine Learning

Instructor – Dr. I-Ming Chiu

Submitted by – Naman Singh

**I. INTRODUCTION**

Depression is a mental health condition characterized by a prolonged sense of sadness and a lack of interest in daily activities. According to the DSM-5 published by the American Psychiatric Association, [1]depressive disorders are categorized into several types, including Disruptive mood dysregulation disorder, Major depressive disorder, Persistent depressive disorder (also known as dysthymia), Premenstrual dysphoric disorder, and Depressive disorder due to a medical condition. [2]These disorders share common symptoms such as feelings of sadness, emptiness, or irritability, along with physical and cognitive changes that severely impact an individual's ability to function. This overview discusses how depression is assessed and managed and highlights the importance of teamwork among healthcare professionals to ensure comprehensive care and improve outcomes for patients. [3]Depression among adolescents is a significant public health concern that affects millions worldwide, with major depressive episodes with severe impairment (mdeSI) being particularly acute. A lot of research is being done on this topic to predict depression among all age groups. In this project we utilized the capabilities of Logistic Regression and Random Forest machine learning techniques to predict mdeSI in adolescents based on the factors given in the datasets. Early detection and accurate prediction are crucial, as they enable timely interventions that can mitigate the long-term consequences of depression, such as academic decline, social isolation, and increased risk of more severe mental health issues. This report explores the methodology, results, and implications of the predictive models, highlighting their potential utility in clinical and educational settings to improve adolescent mental health outcomes.

**II. LITRATURE REVIEW**Adolescent depression is a complex condition influenced by a multitude of factors ranging from genetic to environmental. Recent studies highlight its rising prevalence and the significant impact it has on the long-term well-being of affected individuals[4]. The predictive analysis of depression, particularly using machine learning models, has gained traction as a method to anticipate the occurrence and severity of depressive episodes [6].

Genetic predispositions have been identified as significant predictors of depression; however, environmental factors such as family dynamics, academic stress, and peer interactions also play crucial roles[7]. Understanding these interactions is vital for developing effective preventive measures and treatment strategies. For instance, Zhang et al. (2018) demonstrated that adolescents with high parental involvement and positive school experiences show lower incidences of mdeSI, underscoring the importance of a supportive environment.

Furthermore, the application of Logistic Regression and Random Forest models in predicting adolescent depression has shown promising results, with recent studies reporting high accuracy and sensitivity in identifying at-risk youths [5]. These models not only help in early detection but also assist in customizing intervention strategies to the specific needs of individuals, potentially reducing the progression to more severe conditions [9].

This project builds on the existing literature by integrating these diverse factors into comprehensive predictive models aimed at enhancing our understanding and management of adolescent depression. Through this approach, it contributes to the ongoing efforts in mental health to develop targeted and timely interventions that can significantly alter the developmental trajectories of affected adolescents.

**III. METHODOLOGY**

The methodology of this project encompasses several stages, from data preprocessing and exploratory data analysis (EDA) to model development and evaluation.

**Exploratory Data Analysis (EDA)**

EDA was performed to understand the distributions of various features and the target variable (mdeSI), and to identify any patterns or anomalies that could influence the outcomes of the models. Techniques used included:

* **Data Exploration**: We started with understanding the dataset better by finding unique values and make the values as factor for the model to be implemented.
* **Dataset:**The dataset is a modified data set originally obtained from the NSDUH. The data is modified by approximately equalizing the two-subgroups (50% of depression vs. 50% of depress-free cases) to avoid the imbalanced data problem (i.e., there is about 10% of depression cases in the original data; when this is the case, most of the machine learning approaches may not work well).
* **FEATURES:   
  income info:** Unique Values : 4 [<20,000 | 20,000 - 49,999 | 50,000 - 74,999 | 75,000 or more ]

**gender info:** Unique Values : 2 **[**Male | Female ]

**age info: Unique Values : 3 [**14-15 | 16-17 | 12-13]

**race info:** Unique Values: 5 [White | Hispanic | Black | Asian/NHPIs | Other]

**insurance info:** Unique Values: 2 [Yes | No]

**fatherInHH\_info:** Unique Values: 2 [father in hh | no father in hh]

**motherInHH\_info:** Unique Values: 2 [mother in hh | no mother in hh ]

* **siblingU18\_info:** Unique Values: 2 **[**Yes | No]

**parentInv\_info:** Unique Values: 2 **[**good school experiences | bad school experiences]

* **schoolExp\_info:** Unique Values : 2 **[**good school experiences | bad school experiences]
* **Feature Selection**: In order to select the features, we worked each variable and tried to understand what each feature entails and what are the responses of mdeSI with respect to each feature.
* **Visualization**: Creation of histograms, correlation matrices, ROC curve plots to visualize distributions and relationships between variables.
* **Correlation Analysis**: Examination of the correlation coefficients between features to detect multicollinearity, which could affect the model performance.

**Feature Selection**

Feature selection was conducted to identify the most relevant predictors for the models. This involved:

* **Chi-square tests** for categorical variables to evaluate their independence from the target variable.
* **Feature importance metrics** from preliminary model fits, using methods inherent to Random Forest and logistic regression (e.g., Gini importance and regression coefficients).

**Model Development**

Two machine learning models were developed to predict the occurrence of mdeSI among adolescents:

* **Random Forest**: An ensemble learning method known for its high accuracy and robustness against overfitting. The model was tuned by adjusting parameters such as the number of trees and the depth of the trees.
* **Logistic Regression**: A statistical model used for binary classification that estimates probabilities using a logistic function. The model was optimized by selecting features and tuning the regularization parameter to enhance performance.

**Model Evaluation**

The models were evaluated based on several performance metrics:

* **Accuracy**: The proportion of total correct predictions (both true positives and true negatives).
* **Area Under the ROC Curve (AUC)**: A measure of the model’s ability to distinguish between the classes.
* **Recall (Sensitivity)**: The ability of the model to identify all relevant instances (true positive rate).
* **Precision**: The proportion of positive identifications that were actually correct.

Validation of model performance was conducted using a split of the data into training (75%) and testing (25%) sets to ensure that the models could generalize well to new data.

**IV KEY FINDINGS**

**YEAR:**

**A graph of a number of years

Description automatically generated with medium confidence**

**AGE:**

**A graph with red and blue squares

Description automatically generated**

**RACE:**

**A graph with red and blue squares

Description automatically generated**

**INCOME:**

**A graph of income and income

Description automatically generated with medium confidence**

**FATHER IN HOUSEHOLD:**

**A graph with a red and blue rectangle

Description automatically generated**

**SIBILING UNDER 18:**

**A graph with red and blue squares

Description automatically generated**

**PARENTAL INVOLVEMENT:**

**A graph with red and blue squares

Description automatically generated**

**SCHOOL EXPERIENCE:**

**A graph with red and blue squares

Description automatically generated**

* **Yearly Trends:** A significant increase in mdeSI from 2015 to 2017 suggests worsening mental health conditions among adolescents.
* **Gender Impact:** Females are disproportionately affected by mdeSI, indicating a need for gender-specific mental health strategies.
* **Ethnic Differences:** White and Hispanic adolescents show higher instances of mdeSI than other ethnic groups.
* **Insurance Factor:** Those with insurance report mdeSI more frequently, possibly due to better access to diagnostic services.
* **Income Influence:** Higher income levels correlate with increased mdeSI reports, highlighting possible disparities in healthcare access.
* **Parental Presence:** The presence of parents, particularly mothers, is linked to higher reports of mdeSI, suggesting the importance of family structure in mental health.
* **Siblings' Effect:** Adolescents with siblings under 18 report more mdeSI cases, possibly reflecting the dynamics of larger families.
* **Parental Involvement:** High parental involvement is associated with increased reporting of mdeSI, emphasizing the role of parental engagement in adolescent mental health.
* **School Experience:** Positive school experiences generally decrease mdeSI rates, whereas negative experiences increase them, pointing to the significant impact of the educational environment on adolescent well-being.

**CHI SQUARE TEST:**

These points collectively underscore the complexity of factors influencing adolescent mental health and the necessity of multifaceted approaches in prevention and treatment strategies.

A screenshot of a computer screen

Description automatically generated

Almost all the factors are significant at 0.01 alpha level. However, if consider 0.05 we have two factors that which are showing as insignificant. That are insurance and mother in households.  
  
**A) RANDOM FOREST:**

The graph, "Accuracy vs. Number of Trees," was generated by training multiple Random Forest models with varying numbers of decision trees (100, 200, 300, 400, and 500). For each model, accuracy was calculated by comparing the predicted outcomes against the actual mdeSI status in the test dataset. The plot shows the relationship between the number of trees in the model and the accuracy achieved, indicating how model performance changes as the complexity of the model (in terms of number of trees) increases. **A.1) Model Performance Analysis**

**A.1.1) Accuracy and AUC:  
A.1.1.1)** Train Accuracy and AUC: **The model achieved a training accuracy of 86.22% and an AUC of 0.929.** These metrics suggest that the model is highly effective on the training set, demonstrating strong capability in accurately identifying both true positives and true negatives.

**A.1.1.2)** Test Accuracy and AUC**:** On the test set, the model recorded an accuracy of 82.2% and an AUC of 0.886. While these figures are slightly lower than those of the training set, they still indicate a robust performance, affirming the model's ability to generalize to unseen data effectively.  
 **A.1.2) Sensitivity (Recall):**  
**A.1.2.1)** Train Sensitivity: **A train sensitivity of 95.54%** indicates that the model is exceptionally proficient at identifying true positives (adolescents experiencing mdeSI) within the training data.

**A.1.2.2)** Test Sensitivity: **The test sensitivity of 91.55%** underscores the model's effectiveness in detecting most true positive cases in the test set. This metric is critical for clinical settings to ensure that adolescents with mdeSI are not overlooked.

**A.2) DISCUSSION:**

The Random Forest model leverages an ensemble method to manage the complexities and variability within the dataset effectively. By building multiple decision trees and combining their outcomes, the model reduces the risk of overfitting while maintaining excellent sensitivity and accuracy. This robustness makes the Random Forest an invaluable predictive tool for identifying major depressive episodes with severe impairment among adolescents.

**A.3) ROC Curve Analysis**

**A graph of a curve

Description automatically generated**

The ROC curves, representing both the training and testing phases, demonstrate the trade-off between sensitivity and specificity across different thresholds. The high AUC values indicate that the model is capable of distinguishing between adolescents with and without mdeSI effectively, across a broad range of decision thresholds.

**A.4)Future Recommendations**

**A.4.1)** Cross-Validation: Implement k-fold cross-validation to verify the model's stability and enhance its generalizability.

**A.4.2)**Feature Engineering: Investigate additional features or interactions among existing features to capture more complex associations within the data.

**A.4.3)**Alternative Models: Evaluate other predictive models like Support Vector Machines (SVM) or Logistic Regression to determine if they provide superior or complementary performance.

**B) LOGISTIC REGRESSION MODEL**

**B.1) Model Performance and Metrics  
B.1.1)AUC (Area Under the Curve)  
B.1.1.1)**Training AUC: 0.909, demonstrating excellent model discrimination capability during training. **B.1.1.2)**Testing AUC: 0.901, indicating a high ability to discriminate between cases and controls in unseen data. **B.1.2) Accuracy  
B.1.2.1)** Training Accuracy: 86.22%, reflecting strong overall prediction accuracy in the training dataset. **B.1.2.2)**Testing Accuracy: 82.13%, showing a slight drop but still maintaining high predictive accuracy on new data.  
**B.1.3)Sensitivity (Recall)  
B.1.3.1)**Training Sensitivity: 95.76%, signifying that the model is extremely effective at identifying true positive cases in the training set.

**B.1.3.2)**Testing Sensitivity: 92.23%, slightly lower than training but still very high, ensuring most true cases are correctly identified in the test set.

**B.2)Model Coefficients and Significance**

The logistic regression output indicates several significant predictors of mdeSI:

- Gender (Female), Age (14-15 and 16-17), Race (Black), Income (50,000 - 74,999), Parental Involvement (low), and School Experience (bad) show significant effects on mdeSI, each with a p-value below 0.05, indicating strong evidence against the null hypothesis of no effect**.**

**B.3) Graphical Analysis of ROC Curves**

**A graph of a curve

Description automatically generated**

**-** The ROC curves confirm the model’s strong discriminatory power, with curves for both training and testing phases lying close to the top-left corner, denoting excellent sensitivity and specificity trade-offs.

**B.4)Conclusions and Clinical Implications**

The logistic regression model has demonstrated high accuracy, exceptional sensitivity, and excellent AUC, making it a robust tool for predicting severe depressive impairments in adolescents. However, while accuracy and AUC are high, the moderate sensitivity in the testing phase suggests there might still be room for improvement, especially in clinical settings where high sensitivity is critical.

**B.5) Future Recommendations**

- Advanced Feature Engineering: Investigate more complex interactions between variables to capture deeper insights.  
- Model Tuning: Adjust logistic regression parameters or explore regularization techniques to enhance model performance and prevent overfitting.  
- Comparative Analysis: Contrast logistic regression findings with other models like Random Forest to find the best approach for different scenarios.  
This detailed analysis solidifies the utility of logistic regression in medical and psychological research, particularly for conditions like mdeSI where early and accurate detection is crucial. The high performance metrics suggest that with further tuning and validation, this model could be highly effective in clinical applications.

**V DISCUSSIONS**

In the discussion of the findings from this project, it is evident that both the Random Forest and Logistic Regression models provide valuable insights into the factors influencing major depressive episodes with severe impairment (mdeSI) among adolescents. The Random Forest model demonstrated superior performance, likely due to its ability to handle complex interactions and non-linear relationships among the variables effectively. This highlights the potential of advanced machine learning techniques in mental health applications, where nuanced understanding and prediction can significantly impact interventions and outcomes. Meanwhile, the Logistic Regression model, while not as powerful in terms of sensitivity, offered important insights due to its interpretability, which is crucial for clinical decision-making. These results underscore the importance of choosing the right model based on the specific requirements of sensitivity and interpretability in mental health studies. Further research should consider integrating these models into real-world clinical settings to validate their effectiveness and refine their predictive capabilities, ensuring they can be reliably used to support adolescent mental health initiatives.

**VI. CONCLUSION:**

Throughout our analysis, we explored and compared the performances of Random Forest and Logistic Regression models for predicting major depressive episodes with severe impairment (mdeSI) among adolescents using various metrics such as AUC, accuracy, sensitivity, and precision. Random Forest demonstrated a slight edge in handling complex data interactions and provided high sensitivity, which is crucial for clinical applications to avoid missing true cases. Logistic Regression, valued for its interpretability and efficiency, showed robust generalization capabilities, particularly reflected in its test AUC performance. Both models maintained high accuracy levels, demonstrating their potential utility in clinical settings. Additionally, we addressed computational considerations and threshold optimization for Logistic Regression, enhancing model precision and tailoring decision thresholds to balance sensitivity and specificity effectively. These insights underscore the importance of selecting the right model and parameters based on the specific needs of the healthcare application, balancing the trade-offs between model complexity, interpretability, and performance.

**VII REFERENCES**

[1]The American Psychiatric Association’s Diagnostic Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)

[2] Ropper, A. H., & Samuels, M. A. (2017). *Depression*. In StatPearls [Internet]. StatPearls Publishing. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK430847/>

[3] I.-Ming Chiu1, Wenhua Lu2\*, Fangming Tian1 and Daniel Hart3. Early Detection of Severe Functional Impairment Among Adolescents With Major Depression Using Logistic Classifier.*<https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2020.622007/full>*

[4] Malinauskiene V, Malinauskas R. Predictors of Adolescent Depressive Symptoms. Int J Environ Res Public Health. 2021 Apr 23;18(9):4508. doi: 10.3390/ijerph18094508. PMID: 33922778; PMCID: PMC8122983.

[5] Lin S, Wang C, Jiang X, Zhang Q, Luo D, Li J, Li J, Xu J. Using machine learning to develop a five-item short form of the children's depression inventory. BMC Public Health. 2024 Apr 23;24(1):1118. doi: 10.1186/s12889-024-18657-w. PMID: 38654267; PMCID: PMC11041003.

[6] Kirkbride JB, Anglin DM, Colman I, Dykxhoorn J, Jones PB, Patalay P, Pitman A, Soneson E, Steare T, Wright T, Griffiths SL. The social determinants of mental health and disorder: evidence, prevention and recommendations. World Psychiatry. 2024 Feb;23(1):58-90. doi: 10.1002/wps.21160. PMID: 38214615; PMCID: PMC10786006.

[7]Lopez, S., et al. (2019). "Environmental and Genetic Influences on Adolescent Depression: A Systematic Review." *American Journal of Psychiatry*, 176(10), 785-794.

[8] Ren Y, Wu X, Zou S, Wang X. The integral contributions of parental involvement and parenting style to adolescent adjustments: a regression mixture analysis. Curr Psychol. 2023 Feb 22:1-12. doi: 10.1007/s12144-023-04364-z. Epub ahead of print. PMID: 36845203; PMCID: PMC9944778.

[9] Glassgow AE, Wilder J, Caskey R, Munoz G, Van Voorhees B, Kim S. Mental Health Diagnoses among Children and Adolescents with Chronic Medical Conditions in a Large Urban Cohort. J Behav Health. 2020;9(4):1-8. Epub 2020 Oct 30. PMID: 34413989; PMCID: PMC8373015.