Quantifying the efficacy of ML models at predicting mental health illnesses

Jahnavi Thejo Prakash[1]

[1]*Student, Oakridge International School, Varthur Rd., Circle, Dommasandra, Bengaluru, Karnataka 562125, India*

---***---

Abstract - *Machine Learning has become increasingly pervasive in the field of medicine. Though the large majority of ML-based research focuses on detecting tumors, brain damage, and physical injuries, mental health has not received much attention. The current machine learning models typically fail to consider emotional variability and the extremes of data points when predicting the prevalence of depression. Furthermore, these models don't align with universally-accepted models like Beck's Depression Inventory. It is hypothesized that emotional variability, level of depressive symptoms, amount of labeled data and features correlate with improvements in the accuracy of an ML model. The preliminary results suggest that there is a positive correlation between the level of emotional variability and the amount of labeled data and features with a model's accuracy. In this study, we considered the ability to predict depression through self-reporting, where emotional variability was taken into account through a novel baseline model (which uses a participant's most frequently responded answer). Discussing the findings, we considered (i) an effective means for data collection through questionnaires was developed, (ii) a necessary quantitative improvement for each model was constructed, and (iii) a random forest classifier was found to be the best ML model to predict the incidence of depression. In brief, this research paper assesses the accuracy, reliability, and effectiveness of these ML algorithms, as well as the benefits and drawbacks of the implementation of these algorithms. Though further work and a larger-scale study are required, this paper takes a step in the right direction in self-reporting depression.*

Key Words: **Mental Health; Depression; Machine Learning; Supervised Learning; Artificial Intelligence; Detection; Models**

1. INTRODUCTION

The applications of machine learning to solve or contribute to complicated tasks have grown increasingly popular in the last two decades. Machine learning has assisted health care practitioners with accurately predicting, diagnosing, classifying, and assessing outcomes. While the status quo has been fixated on constructing models based on CT scans, X-rays, test results, and other heavily quantitative figures, emotional well-being has not been explored enough. Mental health illnesses are pernicious to those who are affected by them, "getting in the way of thinking, relating to others, and day-to-day function, according to Harvard Health [1]". Collectively, these illnesses, such as depression, have become highly pervasive affecting more than 264 million people in 2020 alone and they are now the "leading cause of disability in the world" making the treatment of these illnesses a worldwide priority. [2] Improving mental health, however, is extremely challenging as they can manifest in various symptoms and there is yet to be a standardized model for detecting the incidence of depression (and other mental health illnesses).

Therefore, this research aims to answer the question: "*How efficient are the four machine learning models (Logistic Regression & Random Forest Classifiers & Multi-Layer Perceptron) at quantitatively predicting the incidence of depression in young adolescents as compared to their BDI levels (Beck's Depression Inventory)?*"

This paper will explore the potential appropriateness of using machine learning models to detect, and ultimately predict, symptoms of depression based on the consistency of data collected. The research proposed encompasses a quantification of the efficacy of detecting depression using machine-learning models (Logistic Regression with L1 and L2 regularization, Random Forest Classifiers, Multi-Layer Perceptron, Extra Trees Classifier, and Decision Tree Classifier) on young adolescents aged 13-21. These machine learning models will be built upon the consistency of an individual's response to AADA's screening detection [3]. Until now, researchers, justified by eliciting self-response bias, have run machine learning models on alternative sources of data (like GPS tracking or the Screen Time of an individual [8]). As these measures are subject to the individual's personality and preference for electronics, these also carry significant biases. Hence, to eliminate this issue, I have turned to measure the data points several times to improve the precision of the research. I will also take into account these biases when constructing the models.

A limited number of research papers focused on comparing these machine learning models with universally accepted models used to diagnose individuals with mental health illnesses like BDI (Beck's Depression Inventory). Additionally, by surveying unconventional data points(i.e. young adolescents of ages 13-21) that are subject to emotional volatility rather than adults, this research will test the ends and means of depression-detecting models and perhaps tune themselves to fit these individuals. My research is predominantly based on surveys and empirical data collection. I will use a simple google form to collect participants' responses to the AADA's (Anxiety & Depression Association of America) screening detection survey. This survey is used globally to diagnose patients with depression or depressive symptoms. Every 4th day, I requested the participant to fill out the form over 2 weeks. After the 2nd week, I will go through the data and verify whether there are a sufficient number of data points to go with the study. Upon the preliminary selection of data, the 6 machine learning models will be constructed based on the consistency of the participant's responses. Using these models, I will develop correlations between their results to determine how accurate one model is as compared to another.

The dual-edge of my research is comparing these models to the participants' BDI levels, which will also be derived from a survey[4] and filled out only once on the first day of the study. I will again develop correlations between the results from the machine learning models and the respective participant's BDI levels. Rather than developing novel machine learning models that can detect depression symptoms, this research focuses on quantifying the efficacy of the pre-existing models at detecting the prevalence of depression and confirming the relationship between our models' prediction and their BDI levels, a universally-accepted way to detect depression and other mental health illnesses. While further work is needed in building new depression-detecting models, conducting this research is a step in the right direction in understanding whether or not machine learning models can detect depressive symptoms.

1.1 Models

This study heavily weighs on two universally-accepted models for predicting the incidence of depression, namely the Anxiety & Depression Association of America's (AADA) screening for depression [9] and Beck's Depression Inventory [10]. The AADA's screening is "based on (the) Patient Health Questionnaire-9 (PHQ-9) developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke, and colleagues." The screening consists of 10 multiple choice questions which target the emotional stability and depressive symptoms in its participants. There are 4 options for every question in the screening: "not at all", "several days", "more than half the days", and "nearly every day". Based on the responses to the screening, it is possible to get the respondent's level of depression, after consulting a healthcare professional. Since this form highlights key depressive symptoms, it has been used to determine the impact of these factors on an individual's mental health. In other words, the 10 factors that affect an individual's mental health are as follows: levels of (1) interest/pleasure in the conduct of activities, (2) feeling down/depressed/hopeless, (3) ableness/disableness to sleep, (4) energy, (5) ableness/disableness to eat, (6) self-esteem, (7) concentration in the conduct of activities, (8) slowness/restfulness in the conduct of activities, (9) suicidal thoughts, and (10) difficultness in the ability to conduct one's life.

Beck's Depression Inventory is a "21-item, self-rating inventory that measures characteristic attitudes and symptoms of depression [11]." Both of these models have been widely accredited for accurately predicting depression amongst their participants. The BDI model categorizes respondents' level of depressive symptoms under 6 categories, based on adding up the points assigned for every option chosen in the questionnaire (also known as the BDI index/value). Respondents with BDI values between 1-10, 11-16, 17-20, 21-30, 31-40, and over 40 are diagnosed as having "ups and downs that are considered normal", "mild mood disturbance", "borderline clinical depression", "moderate depression", "severe depression", and "extreme depression." In this paper, the BDI model was used to determine whether or not a person was "diagnosed" with depression, where a BDI count greater than 17 or "Borderline Clinical Depression" was counted as depressed. Similarly, the AADA's screening was used to get consistent data on the participant's emotional state and build our models.

2. RELATED WORKS

The necessity of this research paper is highlighted by the shortage of papers that look at detecting depression in young adolescents, one of the largest populations that are adversely affected by mental health illnesses [13], through the lens of machine learning. The large majority of papers look at predicting the incidence of depression amongst an older population or the working population, as seen in papers [6] and papers [7].

Past research has predominantly focused on using surveys to identify factors that can affect one's mental health, but the application of machine learning tools has not been touched upon to such an extent. For instance, Melissa

Deziel et. al from the University of Waterloo, Canada [5] employed survey results that rate five essential factors (i.e. ability to enjoy life, resilience, balance, emotional flexibility, and self-actualization) of a student's mental health as defined by the Canadian Mental Health Association. Though machine learning models like classification and regression have been used, the researchers do not compare and contrast the efficiency of these two models or shed light on whether or not these models can even detect depressive symptoms. Previous research papers also overlook providing improvements to models that can be measured quantitatively. Similarly, M. Srividya et. al from the Journal of Medical Systems [6] does use various machine learning models such as cross Validation, decision trees, naïve Bayes classifier, K-nearest neighbour classifier, and logistic regression to identify how depressed their participants were. Though the study primarily intends to focus on various ML models to detect depression, the paper does not provide insight into the efficiency of these models and how we can improve our current models. The paper's omission of our progression as a field was not substantiated or rationalized. Cognizant of these limitations, paper [8] not only quantitatively compares each machine learning model to each other but also Beck's Depression Inventory.

However, unlike paper [8], this research paper will target data collection through self-report rather than GPS tracking, as this facet weighs extensively on the upbringing of an individual, whether or not they use their mobile devices frequently or in limited usage, etc. Contrary to popular belief [12], this paper proves that it is possible to use self-report to predict the incidence of depression in an individual using a few guidelines that are mentioned below. Additionally, this paper takes into consideration an adolescent's emotional variability [14], a large factor in the improvement in the accuracy of the ML models; this facet is not taken into account in the papers above, a missed opportunity.

As a field, we still lack adequate knowledge of improving our current models, anticipating environmental variability, and a focused set of data points. This research paper will build upon the current applications of machine learning in the depression domain, with a special emphasis on quantitatively critiquing the efficiencies of each model amongst each other and the BDI index based on the consistencies of the data responses.

3. METHODOLOGIES

3.1 Summary of Methodology used in Study

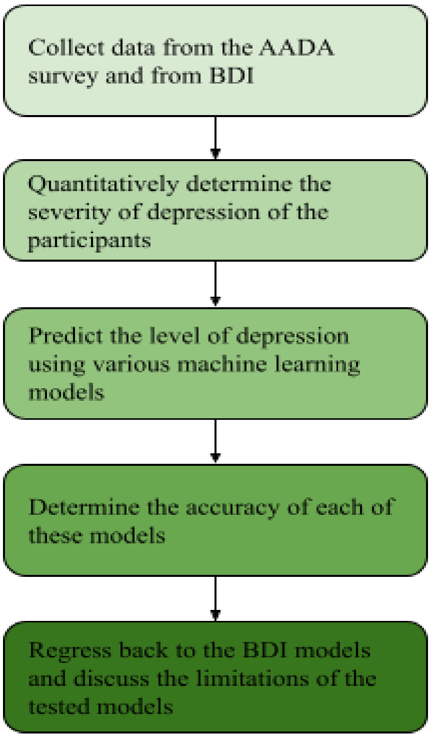
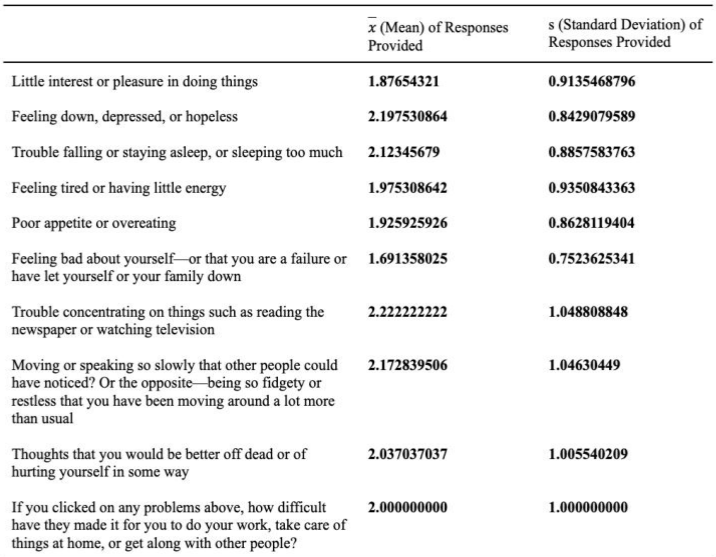

Collect data from the AADA survey and from BDI
Quantitatively determine the severity of depression of the participants
Predict the level of depression using various machine learning models
Determine the accuracy of each of these models
Regress back to the BDI models and discuss the limitations of the tested models

Figure 1.1: representing the five planned flow of data processing from the collection of data till the evaluation of the data.

This research paper entails 5 key steps as represented by Figure 1.1 to the left. All data were collected through a survey with the question set from the AADA's screening of depression [9] and Beck's Depression Inventory [10]. The lack of copious data points attrited the scale of the study. The study initially reached out to 45 individuals, of which 3 did not feel comfortable with the given questions and declined to answer. Out of the 42 responses provided, 11 participants wished to be anonymous and did not share their demographic information. After collecting the necessary data, a pre-analysis was conducted wherein the BDI value was calculated to check the severity of depression for each participant. Every 4 days for the next two weeks, participants would be sent a link to a form with the AADA's screening and a reminder if the participants did not fill out the form within 2 days. If the participant's response was not received within 4 days, the participant's data for the entire duration of the study was not taken into consideration. Without an incentive (ex. compensation sent to participants who took part in the study throughout its duration), there was a drastic fall in the number of data points available to construct the ML models. As mentioned before, every question on the AADA screening consists of 4 options: not at all, several days, more than half the days, and nearly every day To quantize the data collected from the screening, an integer value was assigned to each of these options with the smallest being the least severe and highest most severe. "Not at all" was assigned a value of 1; "Several days", "More than half the days", and "Nearly every day" were assigned a value of 2, 3, and 4, respectively. Out of the 42 participants, only 27 consistently provided their responses throughout the study.

In this study, high emotional variance is defined as having a large standard deviation in the data collected from the AADA questionnaire. In brief, the mean of the standard deviation of a participant's 10 responses throughout the 3 datasets was considered. If this mean has larger than or equal to 3, the participant was defined as having high emotional variance; whereas, if the mean is smaller than 3, the participant was defined as having low emotional variance.



	$\bar{x}$ (Mean) of Responses Provided	s (Standard Deviation) of Responses Provided
Little interest or pleasure in doing things	**1.87654321**	**0.9135468796**
Feeling down, depressed, or hopeless	**2.197530864**	**0.8429079589**
Trouble falling or staying asleep, or sleeping too much	**2.12345679**	**0.8857583763**
Feeling tired or having little energy	**1.975308642**	**0.9350843363**
Poor appetite or overeating	**1.925925926**	**0.8628119404**
Feeling bad about yourself—or that you are a failure or have let yourself or your family down	**1.691358025**	**0.7523625341**
Trouble concentrating on things such as reading the newspaper or watching television	**2.222222222**	**1.048808848**
Moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual	**2.172839506**	**1.04630449**
Thoughts that you would be better off dead or of hurting yourself in some way	**2.037037037**	**1.005540209**
If you clicked on any problems above, how difficult have they made it for you to do your work, take care of things at home, or get along with other people?	**2.000000000**	**1.000000000**

Figure 2.0: Summary Statistics of the data collected from the AADA screening per question

Cognizant of the drastic effect emotional variability plays in predicting depressive symptoms [14], this model made use of a novel method called the "baseline model." The baseline model was constructed by taking the participant's most frequently responded answer; in other words, this model predicts that the participants remain at their most common emotional state. In orTotize a model's improvement in predicting the level of depression amongst its participants as compared to the base model, a new item called user lift (the absolute value of the difference between the model accuracy and baseline accuracy) was created. In other words, a high user lift indicates that the ML model is much better than guessing (i.e. the baseline model); while a low user lift indicates that the ML model is worse than guessing.

Upon exploring several applications to develop the ML models, it was decided that the ML models would be built from scratch using Python. Most notably, Orange, "an open-source machine learning visualization software tool for both novices and experts [15]", was considered to construct the ML models. However, Orange was not considered as the source code for the models was not available to the user, and many research papers have not acknowledged Orange as an official tool for constructing these models.

Due to the lack of copious data points, feature selection was used to automatically select those features which contribute the most to the level of depression in the participants and use the data points efficiently The models created were (1) Logistic Regression with L1 regularization, (2) Logistic Regression with L2 regularization, (3) Random Forest Classifier, and (4) Multi-Layer Perceptron. Using a combined dataset, 4 of the models above were created. And all 4 models are compared with baseline models to determine if ML model prediction is better than baseline. All the steps described here have also been diagrammatically illustrated in figure 1.1.

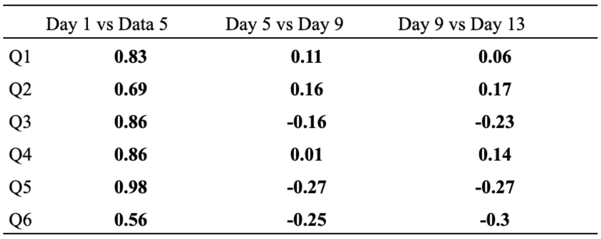
3.2 An Overview and Evaluation of Logistic Regression with L1 and L2 Regularization

According to Yale University, Logistic Regression "attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered an explanatory variable, and the other is considered a dependent variable." A common method for fitting a linear equation to a dataset is by minimizing the least squares, the sum of the squares of the residuals (i.e. the distance from the actual datapoint and predicted data point). Though simple in practice and easy for interpretation, the model performs poorly when there are lurking variables, non-linear relationships, and extrapolation of data. A model which uses L1 regularization is called Lasso Regression, which penalizes the sum of the absolute value of the weights (i.e. coefficients of the variables); While, whereas a model that uses L2 regularization is called Ridge Regression, which penalizes the sum of the squares of the weights of the variables. The model is penalized through the addition of an independent constant to the regression line.

3.3 An Overview and Evaluation of Random Forest Classifiers

Random forest classifiers attempt to train multiple decision trees. Each decision tree is built using a randomly selected subset of data. Decision trees are built in such a way that the leaves represent labels and nodes represent decision points based on a particular feature or subset of features. The final result is obtained by taking the average of the outputs of all decision trees. Among all the ML models, Random Forest Classifier is known to have the highest accuracy as it efficiently utilizes large datasets and automatically balances data sets. However, the most notable limitation of the tree is its large runtime and ineffectiveness for real-life predictions.

3.3 An Overview and Evaluation of Multi-Layer Perceptron



	Day 1 vs Data 5	Day 5 vs Day 9	Day 9 vs Day 13
Q1	**0.83**	**0.11**	**0.06**
Q2	**0.69**	**0.16**	**0.17**
Q3	**0.86**	**-0.16**	**-0.23**
Q4	**0.86**	**0.01**	**0.14**
Q5	**0.98**	**-0.27**	**-0.27**
Q6	**0.56**	**-0.25**	**-0.3**

Figure 2.2 Correlation of data between 3 sample days

Multi-layer perceptron models are based on feedforward neural networks. They comprise at least three layers of neurons: the input layer, the hidden layer, and the output layer. Each of the neurons has activation functions that can be trained using supervised learning. The learning is carried out using the backpropagation of errors. The connection weights are tuned over multiple iterations of activation function evaluation and error backpropagation. Multi-Layer Perceptrons are renowned for their ability to solve "complex nonlinear problems", unlike their counterparts (e.g. logistic regression). Unfortunately, multi-layered perceptrons are also inefficient functions as they do not take into account spatial information and are redundant.

4. RESULTS

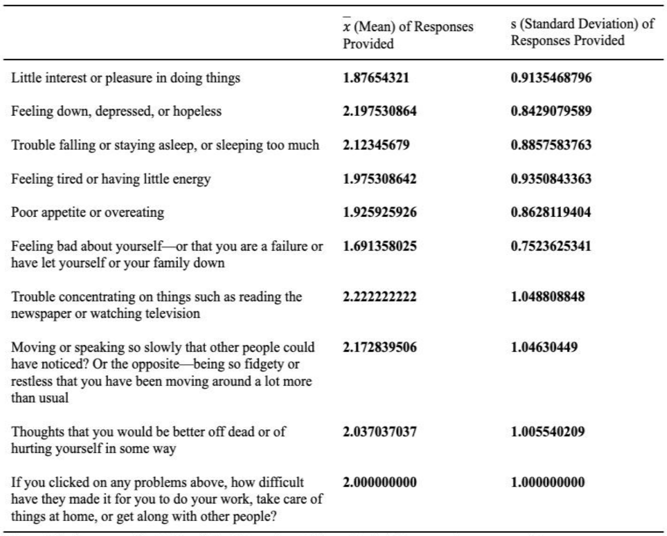
4.1 Data Statistics

Of the individuals who participated in the study, 27 individuals provided sufficient data to analyze and construct the Machine Learning models. The limited number of data points was due to a lack of consistent responses from participants throughout the study. The level of depression for each participant was determined by their BDI value. The average BDI value reported was 16.59, with a standard deviation of 10.55. Of the 27 participants, 33.33% were categorized as having "ups and downs that are considered normal"; 25.93% were categorized as having "mild mood disturbances"; 11.11% were categorized as having "borderline clinical depression"; 18.52% were categorized as having moderate depression; 11.11% were categorized as having "severe depression". None of the participants studied had a BDI index greater than 40, which would have categorized the participant as having "extreme depression." Similarly, the average AADA value and standard deviation computed were approximately 2.02 and 0.33, respectively. The summary statistics for the AADA screening have been depicted in Figure 2.0.

4.2 Feature Importance

From each ML model built, the coefficients of each feature (e.g. "Little interest or pleasure in doing things") was used to calculate the average weightage a factor, overall, had on the prediction of depression in a respondent. The weights calculated for each feature are presented in Figure 2.1. Of the questions asked, "Poor appetite or overeating" came out as the most differentiating feature, having a weightage of 20.14%. While, "Feeling down, depressed, or hopeless" had the least weightage of 3.77%. This relatively low percentage may be due to self-reporting bias in the respondent's answers to this question. This implies that to prevent the underreporting of depression amongst the respondents, questionnaires must be designed with "indirect" questions to assess depression. [12]. Other important features are "Feeling tired or having little energy" and "Trouble falling or staying asleep, or sleeping too much".

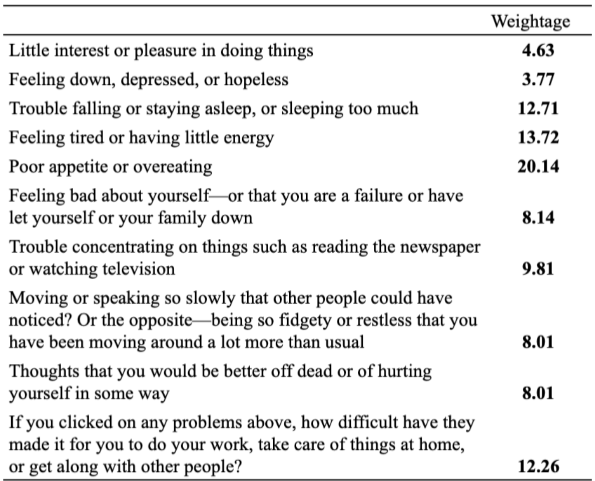
4.3 Baseline Model



	$\bar{x}$ (Mean) of Responses Provided	s (Standard Deviation) of Responses Provided
Little interest or pleasure in doing things	**1.87654321**	**0.9135468796**
Feeling down, depressed, or hopeless	**2.197530864**	**0.8429079589**
Trouble falling or staying asleep, or sleeping too much	**2.12345679**	**0.8857583763**
Feeling tired or having little energy	**1.975308642**	**0.9350843363**
Poor appetite or overeating	**1.925925926**	**0.8628119404**
Feeling bad about yourself—or that you are a failure or have let yourself or your family down	**1.691358025**	**0.7523625341**
Trouble concentrating on things such as reading the newspaper or watching television	**2.222222222**	**1.048808848**
Moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual	**2.172839506**	**1.04630449**
Thoughts that you would be better off dead or of hurting yourself in some way	**2.037037037**	**1.005540209**
If you clicked on any problems above, how difficult have they made it for you to do your work, take care of things at home, or get along with other people?	**2.000000000**	**1.000000000**

Figure 2.0: Summary Statistics of the data collected from the AADA screening per question

As mentioned above, three samples of data were collected every 4 days. The correlation between the data points collected as shown in figure 2.2 is weak. This highlights the presence of high variance (i.e. emotional/mental instability) in the data set and implies that the ML models might struggle to predict the level of depression in the participants.



	Weightage
Little interest or pleasure in doing things	**4.63**
Feeling down, depressed, or hopeless	**3.77**
Trouble falling or staying asleep, or sleeping too much	**12.71**
Feeling tired or having little energy	**13.72**
Poor appetite or overeating	**20.14**
Feeling bad about yourself—or that you are a failure or have let yourself or your family down	**8.14**
Trouble concentrating on things such as reading the newspaper or watching television	**9.81**
Moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual	**8.01**
Thoughts that you would be better off dead or of hurting yourself in some way	**8.01**
If you clicked on any problems above, how difficult have they made it for you to do your work, take care of things at home, or get along with other people?	**12.26**

Figure 2.1 Coefficients/Weightage of each 10 variables on the level of depressive symptoms calculated by taking the average of the coefficients/weightage from ML model.

As stated above, to improve the efficacy of the ML model, a baseline model (constructed based on a participant's most frequent response to the AADA questionnaire). In the presence of high variance (i.e. the average of the range of the responses to the AADA questionnaire across the 3 datasets) in the responses collected, ML models might struggle to predict Depression or No Depression (DND). To decide on the efficacy of the ML model, we need a baseline model to compare. We designed a baseline model that will do predictions based on the mode of the data from three samples. The Baseline model represents the user in their most common state.

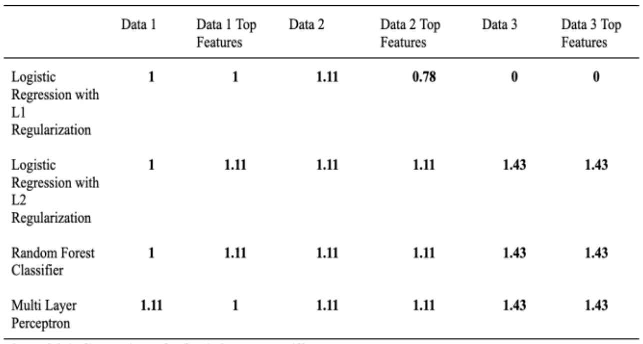
4.4 Prediction Data

Four ML algorithms to train/test with the data set. These four models are then compared with the baseline model as shown in figure 2.3. When ML model prediction accuracy vs. base model accuracy is positive indicates how well is ML model in predicting user state than guessing.

As represented in the figure, all models had a positive lift in predicting the level of depressive symptoms in the participants. The highest user life of 0.26 occurs when comparing simpler models built with random forest classifiers and the baseline model. For more complex ML algorithms such as the Random Forest classifier, the gains are smaller which indicates that complex ML algorithms can learn effectively even on the baseline (or most common state). Comparing prediction power within 4 ML algorithms, Random Forest Classifier performed better than all other models. This strongly indicates the relationship between the questionnaire and DND is non-linear and modeled better with the Random Forest classifier.

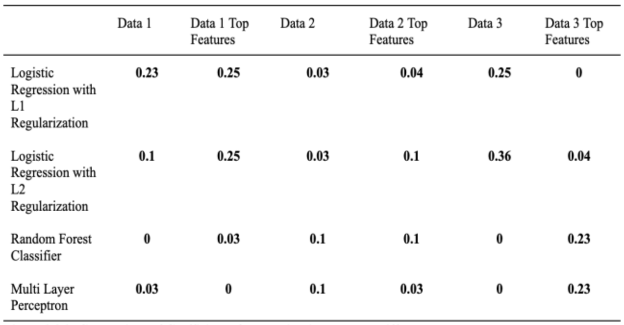
4.2 Model Accuracy

The three figures below (2.0.1, 2.0.2, and 2.0.3) contain the F-Statistic, Coefficient of Determination, and Accuracy of each of the ML models. As mentioned before, 3 datasets were taken from the participant; hence, the names "data 1", "data 2", and "data 3." The "Data 1 / 2 / 3 Top Features" represents the data that has been pre-analysed based on the feature selection algorithm.



	Data 1	Data 1 Top Features	Data 2	Data 2 Top Features	Data 3	Data 3 Top Features
Logistic Regression with L1 Regularization	**1**	**1**	**1.11**	**0.78**	**0**	**0**
Logistic Regression with L2 Regularization	**1**	**1.11**	**1.11**	**1.11**	**1.43**	**1.43**
Random Forest Classifier	**1**	**1.11**	**1.11**	**1.11**	**1.43**	**1.43**
Multi Layer Perceptron	**1.11**	**1**	**1.11**	**1.11**	**1.43**	**1.43**

Figure 2.0.1: Comparison of F-Statistic between different models and datasets



	Data 1	Data 1 Top Features	Data 2	Data 2 Top Features	Data 3	Data 3 Top Features
Logistic Regression with L1 Regularization	**0.23**	**0.25**	**0.03**	**0.04**	**0.25**	**0**
Logistic Regression with L2 Regularization	**0.1**	**0.25**	**0.03**	**0.1**	**0.36**	**0.04**
Random Forest Classifier	**0**	**0.03**	**0.1**	**0.1**	**0**	**0.23**
Multi Layer Perceptron	**0.03**	**0**	**0.1**	**0.03**	**0**	**0.23**

Figure 2.0.2: Comparison of Coefficient of Determination between different models and datasets

According to Figure 2.0.1, Logistic Regression with L2 Regularization, Random Forest Classifier, and Multi-Layer Perceptron with data 3 and data 3 top features are the "best" models as they have a higher F-Statistic value. On the other hand, Logistic Regression with L1 Regularization with data 3 and data 3 top features produces the "worst model", having an F-Statistic value of 0. Concerning figure 2.0.2, Logistic Regression with L2 Regularization with Data 3 creates the "best" model with the highest coefficient of determination; while, the "worst" model is "Random Forest with Data 1, Random Forest Classifier with Data 3, Multi-Layer Perceptron with Data 1 Top Features, and

Multi-Layer Perceptron with Data 3. Similarly, Logistic Regression with L1 Regularization and Data 3, Logistic Regression with L1 Regularization and Data 3 Top Features, and Logistic Regression with L2 Regularization and Data 3 are the most accurate, with an accuracy level of 0.78 or 78%.

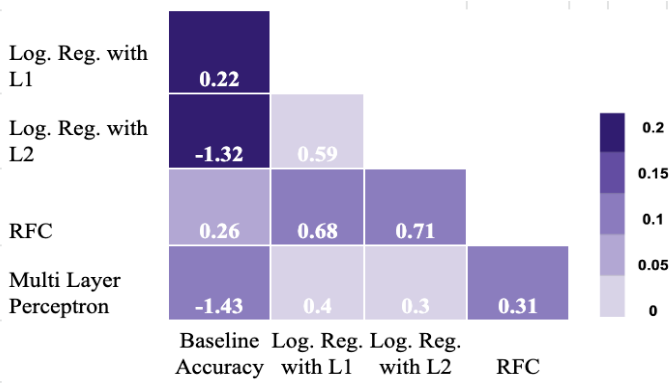


Figure 2.3: User Lift comparision across models and the base line model predicted the level of depressive symptoms in an individual.

4. DISCUSSION

In this research paper, we presented a case study of how effective machine learning models are at predicting the incidence of depression amongst young adolescents, a new data point. We applied Logistic Regression with L1 and L2 regularization and two classifier algorithms, a random forest classifier and a multi-layer perceptron classifier. The results of the study suggest several recommendations for future self-report studies: (i) As seen in the feature importance section of the paper, indirect questions (i.e. questions that focus on secondary effects of depressive symptoms) prevent the underreporting of depression amongst the respondents. For instance, rather than framing a question like "Are you feeling down, depressed, or hopeless?", a question focused on sleep deprivation or poor appetite carries more weightage in saying whether a respondent is depressed or not. (ii) The naïve baseline model (i.e. equivalent to guessing the level of depression of an individual) helps to validity whether or not the responses were biased or carry significant data to create ML models. For instance, if a participant answers randomly for every question (an inadequate data point to build a model from), the user lift of the model would be negative, indicative of a bad model and a potential inaccurate self-report. Regarding the results and the high model accuracy, we can deduce that it is possible to predict the incidence of depression using self-reporting methods. Retracting back to the research question, "*How efficient are the four machine learning models (Logistic Regression & Random Forest Classifiers & Multi-Layer Perceptron) at quantitatively predicting the incidence of depression in young adolescents as compared to their BDI levels (Beck's Depression Inventory)?*", it was seen that Random Forest Classifier was the most accurate model to predict the incidence of depression as it has the largest user lift, as seen in figure 2.3. In the figure, we see that the Random Forest classifier consistently outperforms the other models, indicating that the relationship between the questionnaire's response and the level of depressive symptoms is non-linear and best modeled with RFC. This may also be due to the, comparatively, small number of features used and the limited number of participants in the study.

When model improvement and self-reporting were related, a negative relationship between the consistency of a participant's responses and the ability to predict was deduced. This indicates that people with less consistent responses have a higher ability to predict the incidence of depression. There are limitations in this paper, most notably using a smaller sample size. A larger sample size would allow us to see whether or not the relationships and accuracies of these models remain consistent. A larger cohort would also enable us to take into account more variabilities and thus, improve the models used in the study.

This study acts as a case study for understanding the disparity between our machine learning models and world-reputed models, like BDI. Though successful, it is highly emphasized that larger organizations conduct these experiments on a larger scale across all ages, races, creeds, gender, occupation, etc. This would account for a larger set of variability and a more in-depth understanding of the accuracy of these models. In the future, it is recommended to diversify the participants to generalize this study to a larger population. This could include the diversification of other sources (e.g. GPS tracker, sleep deprivation, accelerometer activity, etc.) With more descriptive data, the relationships between each facet (described above) and the incidence of depression can be deduced.

5. CONCLUSIONS

Based on the findings of the research paper, Random Forest Classifier was concluded as the most effective ML model to predict the incidence of depression amongst young adolescents. This paper stands as a case study for larger-scale research projects to refer to whether or not self-reporting can be used to predict the incidence of depression. Upon completing this research paper, a potential solution is to conduct the same experiment on a larger scale and increase diversity amongst the participants.

ACKNOWLEDGEMENT

Thank you for the guidance of my mentor, Adinawa Adjagbodjou, PhD candidate from Carnegie Mellon University in the development of this research paper.

REFERENCES

[1] Harvard Health. (n.d.). Retrieved August 5, 2022, from https://www.health.harvard.edu/topics/mental-health#:~:text=A%20mental%20illness%20is%20a,day%2Dto%2Dday%20function

[2] SingleCare Team | Updated on Feb. 15, Team, S. C., & Team, S. C. (2022, February 15). *Statistics about depression in the U.S.* The Checkup. Retrieved August 5, 2022, from https://www.singlecare.com/blog/news/depression-statistics/

[3] *Screening for depression*. Screening for Depression | Anxiety and Depression Association of America, ADAA. (n.d.). Retrieved August 5, 2022, from https://adaa.org/living-with-anxiety/ask-and-learn/screenings/screening-depression

[4] *Beck's depression inventory - ismanet.org*. (n.d.). Retrieved August 5, 2022, from https://www.ismanet.org/doctoryourspirit/pdfs/Beck-Depression-Inventory-BDI.pdf

[5] *Analyzing the mental health of engineering students using ...* (n.d.). Retrieved August 5, 2022, from https://www.educationaldatamining.org/EDM2013/papers/rn_paper_34.pdf

[6] Srividya, M., Mohanavalli, S., & Bhalaji, N. (2018, April 3). *Behavioural modelling for Mental Health Using Machine Learning Algorithms - Journal of Medical Systems*. SpringerLink. Retrieved August 5, 2022, from https://link.springer.com/article/10.1007/s10916-018-0934-5

[7] School, H. Y. I., Yang, H., School, I., School, P. A. B. I., Bath, P. A., & Metrics, O. M. V. A. (2019, May 1). *Automatic prediction of depression in older age: Proceedings of the third international conference on medical and health informatics 2019*. ACM Other conferences. Retrieved August 5, 2022, from https://dl.acm.org/doi/10.1145/3340037.3340042

[8] *Algorithm Demasi - GitHub Pages*. (n.d.). Retrieved August 5, 2022, from https://ubicomp-mental-health.github.io/papers/2017/algorithm-demasi.pdf

[9] "Screening for Depression." *Screening for Depression | Anxiety and Depression Association of America, ADAA*, https://adaa.org/living-with-anxiety/ask-and-learn/screenings/screening-depression.

[10] *Beck's Depression Inventory - Ismanet.org*. https://www.ismanet.org/doctoryourspirit/pdfs/Beck-Depression-Inventory-BDI.pdf.

[11] "Beck Depression Inventory (BDI)." *American Psychological Association*, American Psychological Association, https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/beck-depression.

[12] AC;, Hunt M;Auriemma J;Cashaw. "Self-Report Bias and Underreporting of Depression on the BDI-II." *Journal of Personality Assessment*, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/12584064/.

[13] "Products - Data Briefs - Number 379 - September 2020." *Centres Disease Control and Prevention*, Centers for Disease Control and Prevention, 23 Sept. 2020, https://www.cdc.gov/nchs/products/databriefs/db379.htm.

[14] Neumann, Anna, et al. "Emotional Dynamics in the Development of Early Adolescent Psychopathology: A One-Year Longitudinal Study." *Journal of Abnormal Child Psychology*, Springer US, July 2011, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101359/.

[15] "Orange - Machine Learning Training for Health Sciences." *Orange - Machine Learning Training for Health Sciences | The Center for Biomedical Informatics and Biostatistics*, https://cb2.uahs.arizona.edu/orange-machine-learning-training-health-sciences.

[16] *Linear Regression*, http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

[17] "CS 188: Introduction to Artificial Intelligence." *CS 188: Introduction to Artificial Intelligence, Spring 2021*, https://inst.eecs.berkeley.edu/~cs188/sp21/.

BIOGRAPHIES:

Jahnavi Thejo Prakash is a senior at Oakridge International School Bangalore interested in computer science, computer engineering, robotics, and bringing about a positive change in society. I am deeply interested in Artificial Intelligence and Human-Computer Interaction. By promoting digital literacy, empowering women in rural villages, and helping my peers through an educational platform, I am eager to make a difference.