Logo

Description automatically generated

ALY6015 Group 2 Project

**Authored by: Sagar Khilavdiya, Mohan Samala & Prateek Singh**

Course: ALY6015 – Intermediate Analytics

Submission Date: 02/16/2022

**Instructor’s name: Jiyoung Yun**

# Introduction

Purpose: We brainstormed the modern society issues prevailing amongst today’s youth and decided to choose our topic for analysis to be Depression and Anxiety Problem specially amongst students. Machine Learning has been on the rise and healthcare is no exception to that. In healthcare, mental health is gaining more and more space. The diagnosis of mental disorders is based upon standardized patient interviews with defined set of questions and scales which is a time consuming and costly process. Our objective was to apply the machine learning model and to evaluate to see if there is predictive power of survey data to enhance the diagnosis of depression cases.

Many aspects of anxiety and depression overlap, including comorbidity rates, physical and emotional symptoms, diagnostic questionnaires, and risk factors such as unpleasant life events. The current self-reported symptoms-based diagnostic technique has limitations in detecting variations between disorders. Innovative diagnostic paradigms are required to further our understanding and ultimately to more precise treatment options. In this study, we provide a strategy that may be utilized to enhance the confidence in depression diagnosis.

For this objective, we conducted a survey on a small group of students from Northeastern University to collect the various parameters like sleep conditions, smoking & drinking habits, fitness routines, screen hours, etc.

Data Set EDA: We did some cleaning of the data collected by removing redundant and unwarranted variables and removed any observations which were embedded with null values. It was very necessary to recognize any numerical and categorical variables and the response and predictor variables to be used further in our analysis.

# Analysis

We conducted the survey on Qualtrics to get the student’s data about their lifestyle and if they are experiencing any depression or anxiety indications. Post cleaning the dataset we were remained with 20 observations and 18 variables. We decided to analyze upon depression parameter hence kept it as response variable. We decided initially upon several predictor variables and later in the report we will discuss the significant predictor variables used to optimally fit the model.

A picture containing table

Description automatically generated

Fig. 1: Data Set Narrative

Few variables had the null values present which were removed as the part of data cleaning process. Below chart shows the proportion of null values each variable was holding.

Table

Description automatically generated with medium confidence

Fig. 2: Null Values Proportion in the Data Set

From the below screenshot we can see that there is a diversity in each variable data collected. The group amongst which the survey was conducted ranges from 21 years to 28 years with an average of 24 years. Few students walk as less as 100 steps to some going as high as 20000 steps per day. Average steps walked by the students is roughly around 4600 per day. Sleep hours ranges from 6 hours per day to 10 hours per day for few students. There are no outliers in almost all the numerical variables however weight has a lower limit outlier as value 7 years and step has upper limit outlier as 20000 steps.

A picture containing chart

Description automatically generated

Fig. 3: Summary of Numerical Variables

From the below correlation table for the numerical variables, we can see that age and screentime, height and weight and steps and water intake have the strong positive correlation with each other. However, age and steps, screentime and weight and height and weight have strong negative correlation amongst each other.

Chart, scatter chart, bubble chart

Description automatically generated

Fig. 4: Correlation between continuous variables

Analysis of dependent variable, depression, suggests the number of students having the depression or anxiety troubles. In our data set it is recorded that total of 3 students out of 20 are experiencing definite depression symptoms, 9 were positive of not having any depressing thoughts and 8 students are not sure of having a depression or if it is just a temporary low feeling. In percentages, we can see 15% of the students have depression, 45% doesn’t and 40% may have or may not.

Chart, bar chart

Description automatically generated Chart, pie chart

Description automatically generated

Fig. 5: Analysis of Dependent Variable

We were curious to know that how smoking and drinking habits of a person is affecting the state of depression in an individual. Hence, we plotted a histogram of smoking habits against depression state and got some interesting results. Please note these findings can be subjective to biasness given the smaller number of observations getting analyzed and no historical data to support the claim. From the graph we can see that a smaller number of people who snokes have depression symptoms when compared to people who doesn’t. In other words, people who are smoking have fewer depressing thoughts.

Chart, bar chart

Description automatically generated

Fig. 6: Smoking effects on depression

However, the data for drinking when considered the depression state is alarming as it suggests people in depression tend to lower themselves towards drinking habits more.

Chart, bar chart

Description automatically generated

Fig. 7: Drinking effects on depression

Further in our investigation we saw that how depression affects an individual in their day-to-day activity. Hence, we selected steps and screen time variables for our study. We can see the median is different for steps during different stages of depression. The steps median goes decreasing when we move from no depression to may be and finally yes. Students under depression tends to go out less and want to remain inside.

Chart, box and whisker chart

Description automatically generated

Fig. 8: Depression effects on steps

Similarly, screen time is also affected at different depression states. Students screen time goes low when in depression, increases when they are not sure about their state and is almost constant when they are not in depression.

Chart

Description automatically generated

Fig. 9: Depression effects on screen time

Because we are having a multinomial response variable, we releveled our categorical variables. Post that we created a train and test model to determine the significant coefficients by fitting the model. We created three models to verify the best fitting model, one with all the variables and other two with selective variables. Below is the tabulated description of the same. We can see the model three is the best fit model with significant predictor variables as it has the lowest AIC, and we will consider this same model for our further investigation purposes. The residual deviance is also less for the model 3, compared to other models, thus supporting our claim to use this model.

Table

Description automatically generated with medium confidence

Fig. 10: Model comparison

We further created a confusion matrix to see the accuracy of our model. As anticipated, our values for accuracy, sensitivity, specificity, etc. is very less because of the fewer observations and missing historical data.

A screenshot of a computer

Description automatically generated with medium confidence

Fig. 11 Confusion Matrix

Post the modelling and confusion matrix analysis of our model test data set is done we are going to plot receiver operator characteristic curve. We can see the vertical line is going straight up until 0.5 value and then becoming horizontal. It then overlaps with the linear line from 0.5 specificity. This shows that our model is not adept of performing flawless predictions under the specified threshold of 0.5.

Chart, line chart

Description automatically generated

Fig. 12: Test Model ROC Curve

And finally, Area under the ROC Curve is calculated to be 62.5%.

Graphical user interface, text, application, email

Description automatically generated

Fig. 13: Area under the Curve for Test Model

# Conclusion

* Challenges while cleaning the data.
* Fewer observation is affecting the analysis and prediction.
* Drinking habits are affected due to depression state.
* Screen time gets reduced while in depression.
* Depression affects the steps and healthy habits of an individual.
* If we apply advanced algorithms like SVM, Random Forest & Bays theorem we can get better predictions.

# Reference

1. Bevans, R. (2021, June 18). An introduction to the Akaike Information Criterion. Scribbr. Retrieved February 16, 2022, from <https://www.scribbr.com/statistics/akaike-information-criterion/>
2. Datasciencebeginners. (2020, May 27). Multinomial logistic regression with R. R-bloggers. Retrieved February 16, 2022, from <https://www.r-bloggers.com/2020/05/multinomial-logistic-regression-with-r/>
3. Le, J. (2018, April 10). Logistic regression in R. DataCamp Community. Retrieved February 16, 2022, from <https://www.datacamp.com/community/tutorials/logistic-regression-R>
4. Rodriguez, C. I. (2022, February 2). Multinomial logistic regression. Carlos I. Rodriguez. Retrieved February 16, 2022, from <https://www.carlosivanrodriguez.com/guides/regression/multinomial-regression/>
5. Siddiqui, N. (2021, March 6). How to remove a character in an R data frame column? tutorialspoint. Retrieved February 16, 2022, from <https://www.tutorialspoint.com/how-to-remove-a-character-in-an-r-data-frame-column>