Class: AOS204

Assignment: Final Project

Instructor: Alexander Lozinski

Student's Name: Aosen Xu

Mental Health of University Students in Multiple Machine Learning Modeling

-Introduction

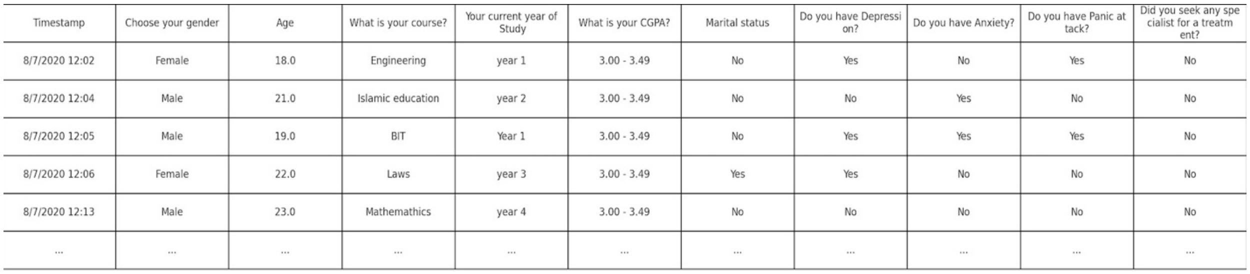
Mental wellbeing is a one of the most serious problems that young people are dealing with nowadays. Particularly, college students are vulnerable to mental health challenges, and this issue continues to grow on a global scale. Based on the Wikipedia statistics, approximately 35% of 14,000 college students from eight countries have an undiagnosed mental health condition. Also, take the United States for example, more than 60% of college students experienced at least one mental health issue during the 2020–2021 academic year [1]. Young people are overwhelmed with academic pressures such as their majors as well as GPA. Meanwhile, at this stage in their lives, many students may struggle to find the resilience needed to face challenges such as failing an exam or a course. As a result, they are more vulnerable to developing mental health issues of varying severity. The bad moods will, in turn, adversely affect their academic performance, which is a negative loop. More importantly, mental health issues can also lead to physical health problems in young people, including mechanical body issues. Therefore, discussing and conducting research on students' mental health is both valuable and necessary.

For the final project, I will explore the students' mental health as well as other important variables and provide some python plots to measure if there are any strong and weak correlations between these variables. Plus, I will use different modeling that I have learned from class through the quarter to examine the accuracy of the model predictions. I will also make a comparison between these multiple models.

-Dataset Analysis

The dataset I have been looking at is from the Kaggle website [1]. (https://www.kaggle.com/datasets/shariful07/student-mental-health?resource=download)

This dataset was collected by a survey conducted by Google forms from students at International Islamic University in Malaysia. The purpose of this dataset, emphasized by the creator, is to examine their current academic situation and mental health. Figure 1 displays all the variables that the dataset includes and the first five rows of the dataset, and rest of the dataset was denoted by the ellipsis. From Table 1, there are multiple variables including gender, age, course, year of study, GPA, and marital status and it is apparent that some of these variables are related to students' academic performance. On the other hand, except the last column "Did you seek any specialist for a treatment?", the rest of variables are all associated with specific symptoms of mental health including anxiety, depression, and panic attack. They are all different representatives of mental issues; they can also interact with each other; they have different meanings, respectively.



Timestamp	Choose your gender	Age	What is your course?	Your current year of Study	What is your CGPA?	Marital status	Do you have Depression?	Do you have Anxiety?	Do you have Panic attack?	Did you seek any specialist for a treatment?
8/7/2020 12:02	Female	18.0	Engineering	year 1	3.00 - 3.49	No	Yes	No	Yes	No
8/7/2020 12:04	Male	21.0	Islamic education	year 2	3.00 - 3.49	No	No	Yes	No	No
8/7/2020 12:05	Male	19.0	BIT	Year 1	3.00 - 3.49	No	Yes	Yes	Yes	No
8/7/2020 12:06	Female	22.0	Laws	year 3	3.00 - 3.49	Yes	Yes	No	No	No
8/7/2020 12:13	Male	23.0	Mathemathics	year 4	3.00 - 3.49	No	No	No	No	No
...	...	...	...	...	...	...	...	...	...	...

Table 1. Display of all variables from the dataset with the first five rows.

Anxiety, which can also be interpreted as Generalized anxiety disorder (GAD), is often characterized by a constant sense of worry or unease that can disrupt everyday activities [2]. Depression, on the other hand, is a prevalent mental health disorder marked by ongoing feelings of sadness, despair, and a diminished interest or enjoyment in daily activities [3]. Lastly, a panic attack refers to a sudden surge of overwhelming fear that causes intense physical and emotional reactions, even in the absence of any real threat or obvious reason [4]. Anxiety and depression can contribute to the onset of panic attacks, which, in turn, can impact mental health, further exacerbating anxiety and depression. Therefore, it is also a loop that negatively impacts each other, and the mental health can get much worse eventually. For the rest of the study, I am picking up depression as a representative of mental health problems as well as a target variable since from the pie chart (Figure1), almost one third of students suffered depression based on the dataset. I will be also applying depression to the modeling section.

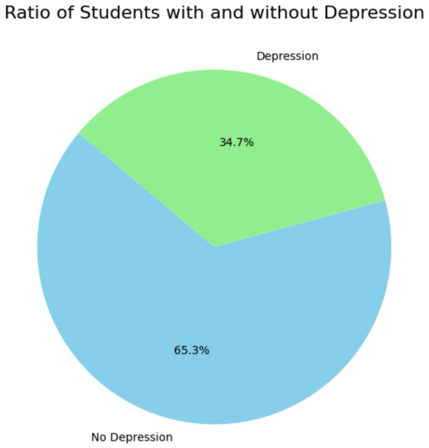


Figure 1. A pie chart showing the ratio of students with depression to students without depression

Based on the entire dataset, I am also wondering if the mental health issues have a correlation with our features and I would like to discuss if these features significantly affect the psychological state of university students. For instance, how is depression linked to age, gender, etc., and whether there is a strong connection between the target variable and these features. Henceforth, in attempt to explore their relationships, I made several plots that exhibit depression with respect to each feature.

First, figure 2 illustrates the proportion of the male and female students who experienced depression relative to the total number of female and male students. It is noticeable that the female students have a higher ratio of suffering depression compared to the male students. It may suggest that female students are more susceptible to academic pressures by pursing a good GPA and they are more vulnerable compared to male students.

Likewise, figure 3 shows the ratio of students with depression at ages 18 to 24. However, due to the lack of the number of people who took the survey at certain age, such as the age of 21 and 22, the statistics may not be very subjective. Still, we can see from the right plot, it is evident that the average proportion of students aged 18 to 24 with mental health issues is approaching 40%, highlighting that young college students are more likely to experience poor mental health.

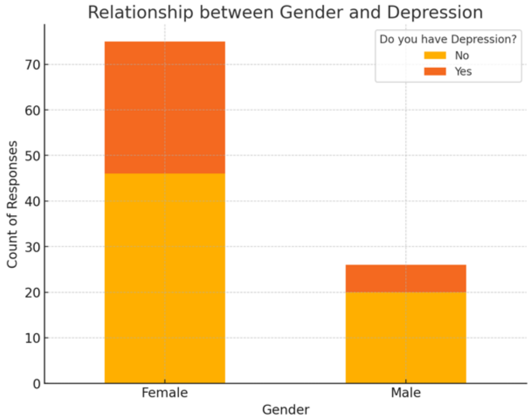


Figure 2. A stacked bar chart illustrates the relationship between gender and depression among respondents.

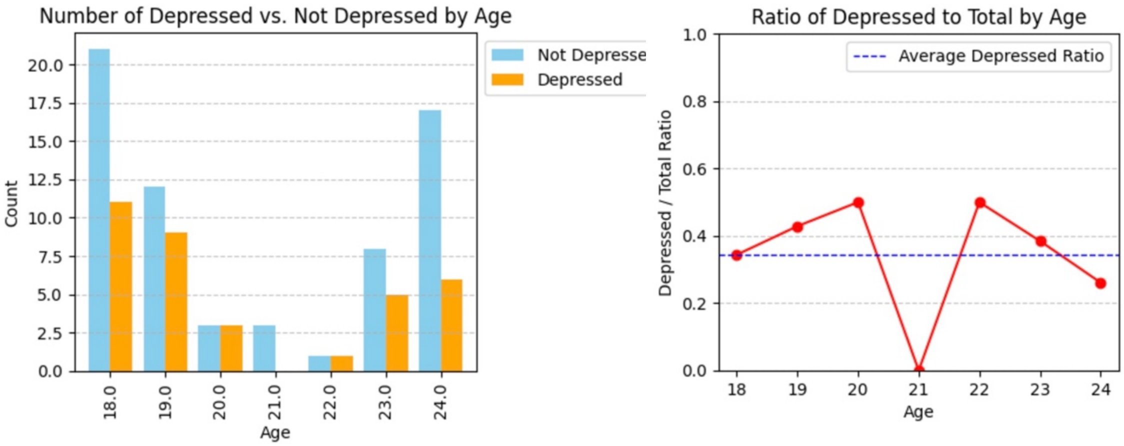


Figure 3. (left) Number of students with and without depression by age. (right) Ratio of students with depression categorized by ages from 18 to 24.

Moreover, another feature that seems intriguing is marital status. It may seem unusual for students within the specified age range to report their marital status as “yes.” However, a glance at the marital status column does reveal several instances of “yes,” suggesting that marrying at a young age is both socially acceptable and legally permitted in Malaysia. Figure 4 displays the distribution of students with and without depression by marital status; interestingly, all students with marital status as “yes” have mental health problems. This is reasonable because students are too young to handle the trivial things from marriage, which will impact their academic performance and eventually leads to mental breakdown.

The last feature I am looking at is students’ GPA since GPA is the most direct representation of their performance on school and thus it may have a larger correlation with mental state. Figure 5 shows the feature importance for GPA with students who experienced depression by implementing the logistic regression model. First of all, GPA ranges in 3.00 – 3.49 and 2.50 – 2.99 has significantly positive coefficient values, indicating that a student with GPA in these ranges, especially within 3.00-3.49, is more likely to be associated with depression than students in the baseline category. On the other hand, GPA ranges in 3.50 – 4.00 and 2.00– 2.49 has negative coefficient values, indicating that a student with GPA in these ranges, especially within 3.50 – 3.99, is less likely to report depression compared to the baseline GPA category. The results can also be justified through careful consideration. For example, students with mid-range GPAs (3.00–3.49 and 2.50–2.99) may experience increased anxiety about their academic performance since they strive to achieve higher grades, but the reality of lower GPAs will discourage them and eventually contribute to depression. On the contrary, students either with highest GPAs (i.e. 3.50- 4.00) or lowest GPAs (i.e. 2.00 – 2.49) are two extremes. The first group of students may be satisfied with what they already achieved, and the latter group of students might have given up and also be happy with where they are currently at. Therefore, the feature importance of GPAs linked to depression provides a reasonable and meaningful outcome.

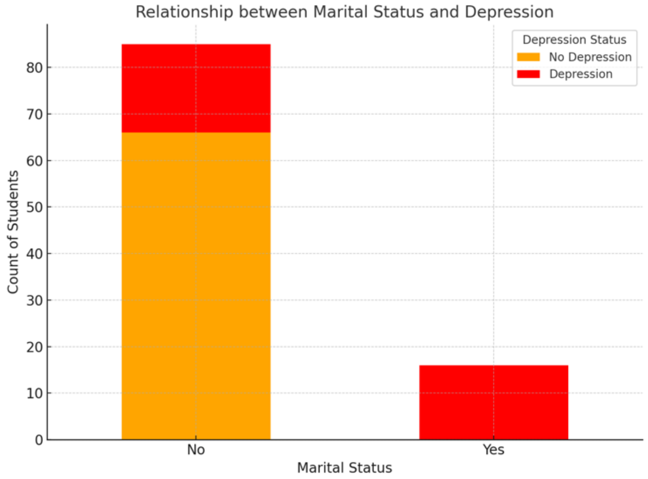


Figure 4. Number of students with and without depression based on marital status.

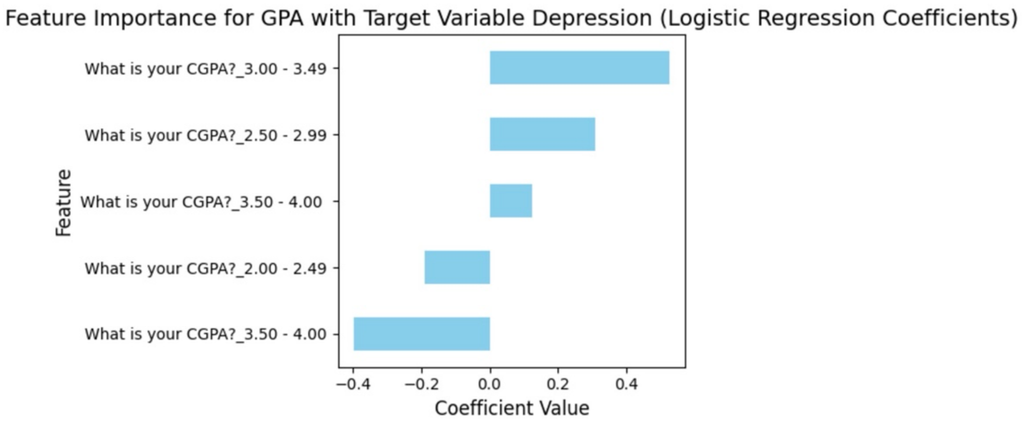


Figure 5. Feature importance for GPA with logistic regression coefficients with respect to students with depression.

- Modeling Applications & Comparison

For the modeling section of the final project, I will be applying multiple machine learning algorithms that I have learned from the class through the quarter as well as examining their performance on the chosen dataset, and there are three models that are employed: random forest, logistic regression and decision trees. Before applying the models into the dataset, I made some changes of the input variables based on the professor's comments. For example, I deleted "What is your course?" column. The column is not a simple binary categorical column with "yes" and "no"; instead, it contains numerous course categories such as psychology, economy, engineering, etc., which applying a numerical order (i.e. 1, 2, 3…) implies a relationship or hierarchy. In other words, if we set psychology as 1, economy as 2, and engineering as 3, then economy is closer to psychology than to engineering by implication, which is not true. Not to mention that this feature column has nearly 30 courses. Therefore, dropping the column will avoid any confusions and biases to the modeling results. What is more, besides dropping the course feature, I also converted the rest of categorical features (e.g. gender and marital status) into numerical features with "1" and "0" by executing "pd.get_dummies()" on python.

The first machine learning algorithm that is being applied is random forest. There are multiple reasons of making use of random forest. First of all, random forest is not only robust against noisy or any irrelevant features, but it is also capable of handling well mixed data types. For example, the dataset includes both categorical features such as marital status and gender as well as numerical features like GPA. Second, random forest generates a feature importance ranking, highlighting which variables (e.g., anxiety, marital status, GPA) are most influential in predicting depression. Moreover, depression is a complicated mental health condition affected by complex relationships between different variables; for instance, age might interact with GPA or marital status. And random forest has the ability to capture these nonlinear interactions effectively. Thus, it is crucial to implement random forest to the dataset.

Figure 6 shows the top ten feature importance in predicting depression according to random forest regression model as well as the ROC curve. One noticeable feature among different variables is that "marital status" ranks the highest feature importance with respect to depression, which is high up to 0.30 and the value is also outstanding from those of other variables. This is illustrating that whether a student is married or not is the most influential factor in predicting depression, which corresponds to the fact that students report their marital status as “yes” are all experiencing depression as indicated in Figure 4. Besides, following “marital status” are age, anxiety, and panic attack. Especially, students with depression also suffer anxiety and panic attack, meaning that these different symptoms of mental issues are close to and interact with each other. The right plot of figure 6 shows the ROC curve, and an AUC of 0.73 suggests that random forest has a 73% likelihood of accurately differentiating between a randomly selected student with depression and one without depression, which is a moderate performance but has space to improve. Plus, the curve looks not smooth because of multiple threshold values shown on the ROC curve. It may be also due to the small dataset I am using. The implementation of random forest finally leads to test accuracy of 67%. Meanwhile, the results indicate that the model predicts students without depression more accurately than those with depression, demonstrating the need for improvement in its performance for students with depression.

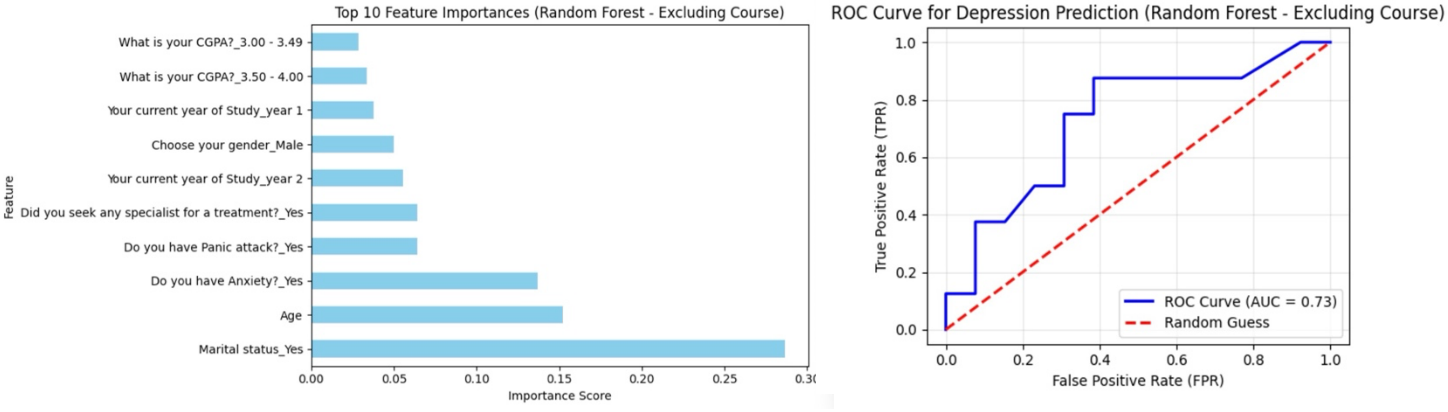


Figure 6. Top 10 feature importances (left) and ROC curve (right) for depression prediction based on random forest.

The second machine learning algorithm that is being employed is logistic regression model. Logistic regression model is designed for binary classification problems, helping to estimate the likelihood of an outcome falling into one of two categories. For instance, the target variable depression is a binary variable with only “yes” and “no”. Similar to random forest, logistic regression is also good at coping with both categorical as well as numerical features. Nevertheless, the model assumes linear relationships between the feature variables and the target variable – depression, which will lead to overfitting and mislead of performance metrics. Therefore, it is important to move depression out of the feature variables while coding. In order to address the issue, I used “X = data.drop(columns=['Do you have Depression?'…] to isolate the target variable. Plus, I split the data into training and test sets, for which 20% of dataset will be allocated to the test set and the remaining 80% of the dataset will be used for training the model. Finally, the refined codes generate additional figures that are worth looking at.

Figure 7 describes the confusion matrix as well as the ROC curve for logistic regression model. First of all, the True Negatives on top left of the matrix indicates that the model correctly predicted Without Depression for 12 individuals who actually did not have depression. Second, the False Positives on top right of the matrix illustrates that the model incorrectly predicted With Depression for only one individual who did not have depression. Third, the False Negatives on bottom left of the matrix demonstrates that the model incorrectly predicted Without depression for five individuals who actually had depression. In the end, the True Positives on bottom right of the matrix implies that the model correctly predicted With Depression for three individuals who actually had depression. Generally speaking, the model performs well in identifying individuals without depression, with the accuracy over 90%; on the other hand, the model performs poorly in identifying individuals with depression, achieving an accuracy of only around 40%. Besides the confusion matrix, the ROC curve plot (right) of figure 7 shows an accuracy of 81%, suggesting that logistic regression model is good at discriminating between students with and without depression, which is significantly better than random guessing. Furthermore, the overall test accuracy of the model is 71%, with a better performance on predicting students without depression, corresponding to what the confusion matrix reflects.

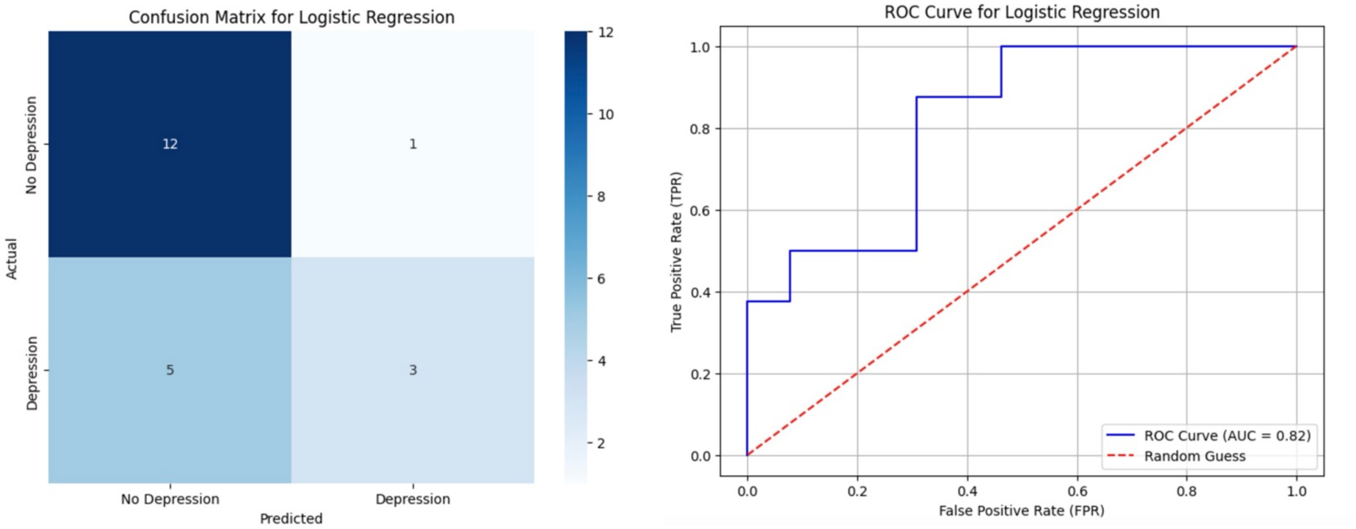


Figure 7. Confusion matrix for logistic regression model (left) and ROC curve (right) for depression prediction based on logistic regression model.

The last modeling algorithm that is being utilized is decision tree model. Contrary to logistic regression model, decision tree does not assume linear relationships between feature variables and the target variable. Instead, it is good at handling nonlinear relationships that could happen between the input features in my dataset. Also, it has the most intuitive representation of the importance of the features.

Figure 8 shows the confusion matrix for decision (top) tree as well as a decision tree with the max depth of five (bottom) for the dataset. And examining the true negatives, true positives, false negatives, as well as false positives located on the four quadrantes of the confusion matrix draws to a conclusion that decision tree is better at predicting "Without Depression" with ten correct numbers compared to "With Depression" but only with three correct numbers, which might shed light on a fact that there might be a bias toward the majority class if "Without Depression" is more common in the dataset. The decision tree with a max depth of five represents the relationships between various features and the target variable, and the reason I choose a max depth of five is that it is the most optimal number to visualize and interpret. Either too large or too small max depth will induce overfitting or lack important information. For the decision tree, the top split shown in decision tree is marital status, which is considered the most key determinant of depression; after marital status is anxiety and panic attack and so on. The outcomes are aligning with those in the data analysis section as well as of random forest. The overall test accuracy for decision tree is 62%, which is quite low. This could be either caused by the shallow depth of my dataset or a noisy dataset or both. Apparently, there is still room for improvement.

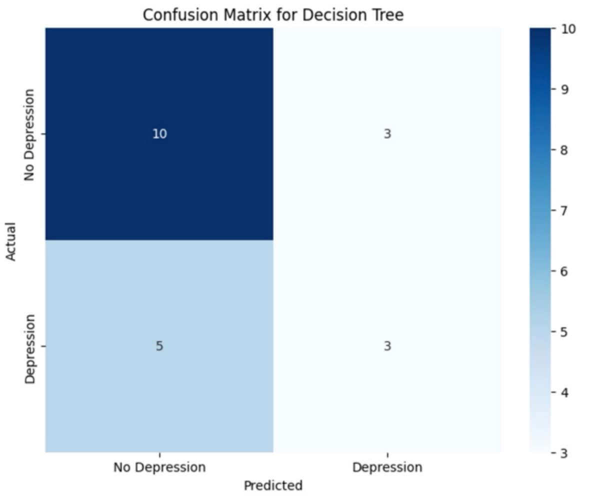


Figure 8 (top). Confusion matrix for decision tree

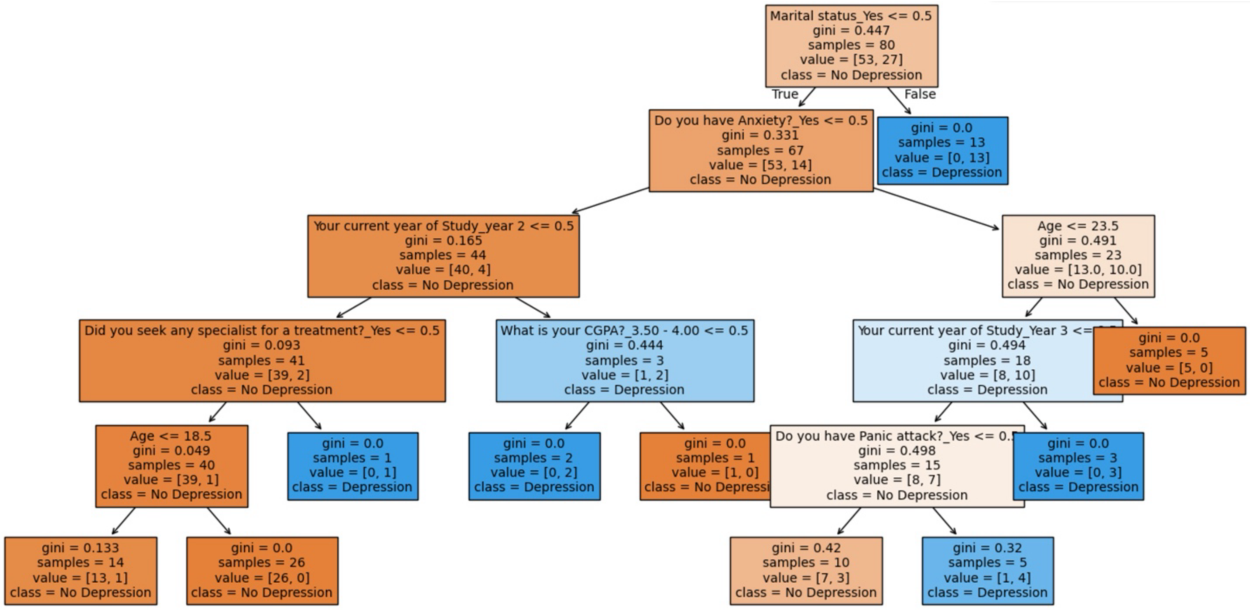


Figure 8 (bottom). Decision tree with a max depth 5 for the dataset

Finally, the test accuracies of random forest, logistic regression, and decision tree are 67%, 71%, and 62% respectively. The comparison between the models demonstrates that logistic regression model has the best performance on the dataset, while decision tree exhibits the poorest and random forest is in between. This is implying that the chosen dataset has primarily linear relationships due to the characteristics of logistic regression, while decision tree and random forest are not the best-fit models compared to logistic regression.

- Conclusion

The chosen dataset regarding mental health of college students has multiple variables, and I set depression as the target variable and the rest of variables excluding courses as the feature variables. Exploring the target variables and different feature variables reveals that depression is highly related to gender and students' marital status, and it's also linked to their GPA. The examination of different modeling algorithms indicates that logistic regression shows a better performance on the dataset than random forest and decision tree, with a higher accuracy and ROC score. Meanwhile, logistic regression and decision tree all illustrate a better prediction on students without depression. In summary, mental health remains a critical concern for college students worldwide, underscoring the need for further research to address this pressing issue.

Reference Page

[1] Shariful, M. (2020). *Student Mental Health*. Kaggle. Retrieved November 25, 2024, from https://www.kaggle.com/datasets/shariful07/student-mental-health

[2] American Psychological Association. (2022, October). *Mental health on campus: Examining the concerns and care options for college students*. Retrieved from https://www.apa.org/monitor/2022/10/mental-health-campus-care

[3] National Institute of Mental Health. (n.d.). *Anxiety disorders*. U.S. Department of Health and Human Services. Retrieved November 25, 2024, from https://www.nimh.nih.gov/health/topics/anxiety-disorders

[4] American Psychiatric Association. (n.d.). *What is depression?* Retrieved November 25, 2024, from https://www.psychiatry.org/patients-families/depression/what-is-depression

[5] Saprea. (n.d.). *Panic attacks: Symptoms, causes, and treatments.* Saprea. Retrieved November 25, 2024, from https://saprea.org/heal/panic-attacks/?gad_source=1&gclid=Cj0KCQiAgJa6BhCOARIsAMiL7V8WrI7HEuwwZxGc4rM3YIAqEHLHqmBZ63id3LuHadyTWQtgpqMw_yQaApMjEALw_wcB