Introduction

The decline of mental health has been, and continues to be, a major issue- especially with college students. With the transition from home life to college life, and increased workload, many students find themselves scared and feeling lost. This transition occurs during a period of physical and mental growth that is tumultuous on its own. Young adults between the ages of 15-24 are more likely to endorse symptoms of depression, and most lifelong mental health disorders have their first onset during this time as well (Costello et al, 2022). The pressure of this new life can further exacerbate an already existing condition or lead to anxiety, panic attacks, and depression. According to one study, 48.6% of students reported moderate stress, while 27.9% experienced high stress within the past year. Additionally, if a student is part of a marginalized group, their risk of developing depression increases. Being part of a marginalized group can impact resiliency and sense of control. This stress over time leads to negative health outcomes (Reohr et al, 2023). Universities need to examine factors that could be impacting the decline of mental health. Doing so will positively impact students' performance in academics, their college experience, and their overall health.

While mental health struggles can impact anyone, there are certain factors that increase the risk of developing them. Age, year of education, declared major, GPA, and gender identity can all play a factor in the decline of mental well-being. Having to work a part time job may also contribute to increased stress. It is important to explore potential causes of increased stress among college students, so that institutions can take steps to mitigate it. Possible interventions include self care workshops, networking activities, extracurriculars for students, and easy access to mental health counseling. Special focus should also be placed on students who are in high

stress majors such as medicine and engineering. By identifying root causes of stress, specialized interventions can be used to ensure students are set up for academic and personal success.

The datasets we are utilizing are from Kaggle. The first one is titled Student Mental Health (Student Mental health | Kaggle), and contains anonymous responses from numerous individuals across the internet via a google form. The responses were compiled and used to create the said dataset.

The second is titled Medical Student Mental Health (Medical Student Mental Health-Kaggle), and contains self reported data and results from psychological tests collected from medical students. The dataset includes demographic information, internal measures (presence of a job or partner, year of study, self-reported health status, and use of psychological services). Also included are the scores from various psychological inventories such as: Jefferson Scale of Empathy (JSPE), Center for Epidemiologic Studies Depression Scale (CES-D), Questionnaire of Cognitive and Affective Empathy (QCAE), State-Trait Anxiety Inventory (STAI), and Maslach Burnout Inventory (MBI) .

Methods

Different independent variables were evaluated to determine how they impacted students' level of stress and their overall mental health. These independent factors included gpa, major, hours spent studying, and whether they work a job. The dependent variables included if students reported having a mental health condition, and their scores from the previously discussed inventories and scales. Identifying the relationships between these variables will help pinpoint causes of stress and allow administration and faculty to implement solutions to help students.

Medical Student Data

For the Medical Student Dataset, SPSS was used to conduct feature selection, data preprocessing, and data visualizations. The data had no missing values, and the data was numerical or categorical. Independent variables for this particular dataset were age, year of study, whether the student had a partner or job, hours of studying, perception of health and whether the student was receiving psychotherapy services. Dependent variables were students' scores on various psychological tests, such as job satisfaction, empathy, depression, anxiety and burnout. A line graph of the test results was made, as well has bar graphs of each independent variable with associated test scores.

Next, in Orange, four supervised machine learning models; kNN, random forest, linear regression and SVM were built for each independent variable. Each model's performance was compared on their fit with the data. Then, linear regression models for each dependent variable were built to determine which independent variables had the strongest correlation with each outcome. This was done using Stata.

Student Mental Health Dataset

The binary output of if students suffered from depression (1 = yes, 0 = no) was utilized as the dependent variable, which was used with several independent variables, such as cGPA (cumulative GPA), gender, age, major/course, and year in college. The most important independent variables, at least for this project's purposes, were cGPA, course/major, and year in college.

Preprocessing and exploratory analyses were performed in Orange and Stata, where the best models and variables were identified to be used, and where binary/dummy variables were created. The data was also reconfigured to be easier to read within Orange, Stata, and SPSS. In Orange, Naive Bayes, Random Forest, Neural Network, Logistic Regression, SVM, and kNN

were done. In Stata, logistic regression, Spearman's Correlation, t-tests, and general descriptive summaries were performed, in order to examine what independent variables correlate to the dependent variable (having depression). Spearman's Correlation test was chosen, due to the fact that the relationship is not necessarily linear, and that there are two kinds of variable types in use- nominal/scale and continuous numerics. In SPSS, Naive Bayes was performed, in order to see how well the model predicts if a student has a high possibility of developing depression.

Results and Discussion

Medical Student Data

The data visualizations for the Medical Student Data pointed towards a few interesting trends. Overall, students that reported receiving psychotherapy services reported higher scores on the psychological tests. Students who were in their early years of study scored higher on the tests for depression, anxiety, and emotional burnout. Students who were further along reported higher levels of cynicism and lower sense of achievement. Students who were dissatisfied or neutral about their overall health reported higher scores of depression, anxiety, and burnout. Scores on these tests also increased the more time students spent studying outside of classes. Having a job or partner made little impact on psychological test scores. Job satisfaction and academic motivation was unchanged across independent variables.

With the supervised machine learning models, linear regression was the best performing model for predicting depression, anxiety, and burnout. With these variables, the p values ranged from 0.00 to 0.2. For empathy and job satisfaction none of the four models performed well, with negative R-squared values. Thus, in order to determine the relationships between the test scores and the independent variables, linear regression analysis models were built.

For the regression analysis models, each dependent variable (depression, anxiety, empathy or burnout scores) had different statistically significant predictors. For example, the statistically significant predictors for the State-Trait Anxiety scores and their p-values were hours spent studying (0.018), year of study (0.022), overall perception of health (0.00), and use of therapy services (0.00). This is similar for the depression scale scores (p values were 0.00) except for hours spent studying, which was not statistically significant. For the burnout scores, hours spent studying, overall health perception and use of therapy services were significant predictors, with p values of 0.00. However for the empathy scores, the model itself was not statistically significant.

From the data visualizations, regression analysis and performance of the machine learning models, it is important to note that not one individual independent variable, such as hours spent studying or year of studying, were the ultimate predictors for the dependent variables. It is important to note that individuals who seek out therapy services are more likely to report mental health symptoms, which is reflected in the increase in test scores for those specific cases in the data. These analyses show that students' mental health is impacted by various different factors in their personal, professional, and academic lives. Colleges and universities thus must provide holistic resources to aid students, as well as encourage work-life balance.

Student Mental Health Dataset

In orange, some of the test results had promising outcomes, while some others did not. For each test performed, performance matrices of precision, recall, F1, MCC, AUC, and CA were all calculated. Many models had scores in each category above 0.5, and were close to 0.7 or

0.8, indicating a good model fit. Naive Bayes, logistic regression, random forest, and SVM had good models, as mentioned previously.

Naive Bayes was performed in SPSS. The independent variables chosen were: students' cGPA, current year in their degree, and what course/major they are in. The dependent variable was if they had depression or not. The subset summary showed that the best predictor variables to use were students' majors and their cGPA. Classification results showed that the model accurately predicts students not having depression by 92.4%, and predicts if students do have depression with 77.1% correctness. Overall, the Naive Bayes model accurately predicted if a student would have depression or not at 87.1%.

In Stata, a table for easy visualization was created that shows how many students had depression, based upon what their major was and what their cGPA was. It showed that students in the bachelor's of computer science program with a 3.5 cGPA have higher rates of depression than any other 3.5 cGPA category. The program with the highest rate of depression, overall, was engineering. Students with cGPA averages of 3.0 and 3.5+ had higher rates of depression than those with cGPA averages of less than 2.5. It is important to note that these results could also be due to there being more students in certain programs/majors than others.

Spearman's Correlation test showed a positive relationship between having depression and marital status and gender, meaning that more women have depression than men, and that married students have a higher rate of depression. There are negative relationships between depression and age, year in school, and cGPA. This shows that as age increases, depression rates decrease. Also, it shows that as students get further in their degree, the rate of depression decreases. For cGPA, it shows that the higher the cGPA is, the lower the likelihood of depression is, which is interesting, considering other tests/charts showed the opposite. The test also showed

that if a student has depression, then they are also more likely to have anxiety, panic attacks, and seek help for mental health.

The t-test showed that there is evidence to state that having depression is related to cGPA, with the null hypothesis being not having depression, and the T>t being 0.542, which is much greater than the standard significance level of 0.05. This matches with the results of the chart discussed previously, and further disproves the Spearman's Correlation test about cGPA.

Conclusion

Mental health issues are an ongoing crisis across universities, world-wide. Students oftentimes struggle in silence, while their grade point averages, social habits, course work, and overall public appearances speak for them. With these two datasets, it shows how impactful different stressors in life impact mental health. GPA/cGPA, gender, age, chosen major / program, year into the program, job satisfaction, studying habits, test scores, and overall outlook on life all impact, or are impacted by, mental health.

Multiple tests and Machine Learning Models were performed, such as T-tests, Correlation tests, kNN, logistic and linear regressions, and Naive Bayes, to name a few. The outputs showed how much mental health struggles, like depression and anxiety, impact a student's pre-disposed hardships in school. On the flip side, striving for perfection and academic achievement may lead to higher risk of developing depression and/or other mental health struggles.

Students who have depression, and mental health struggles, in general, are more likely to identify as female, be in a younger age group, be in a major / program along the lines of engineering or computer science, and suffer from burnout (or are already suffering from burnout. Similarly, students who strive for high GPAs and high test scores, are more likely to report

having struggles with mental health. Universities should strive to help their students as much as possible, in order to assist them in achieving success. With mental health crises, there should be programs or discussions in place that help to destigmatize mental health and bring more awareness to it, as well as find ways to help improve vulnerable groups' mental health statuses.

Link to OneDrive Folder: Group 10 Screenshots

References

Costello, M. A., Nagel, A. G., Hunt, G. L., Rivens, A. J., Hazelwood, O. A., Pettit, C., & Allen, J. P. (2022). Facilitating connection to enhance college student well‐being: Evaluation of an experiential group program. American Journal of Community Psychology, 70(3/4), 314–326. https://doi-org.pitt.idm.oclc.org/10.1002/ajcp.12601

Reohr, P., Irrgang, M., Loskot, T., Siegel, L., Vik, P., & Downs, A. (2023). Stress and Mental Health Among Racial Historically Marginalized and Advantaged Undergraduate Students. Psi Chi Journal of Psychological Research, 28(3), 180–190. https://doi-org.pitt.idm.oclc.org/10.24839/2325-7342.JN28.3.180