Peter Gansallo
Professor Basheer
Final Report
Capstone Project

Predicting Students at High Risk of Depression

Abstract

This project examines depression risk factors among college students using data from the Healthy Minds Study, a long-running annual mental health survey conducted across multiple U.S. universities. The study focuses on variables such as financial stress, exercise habits, GPA, loneliness, and gender/athlete status, alongside self-reported depression scores. Previous research has suggested that exercise can reduce mental health issues, but college athletes also face unique pressures that may put them at higher risk (Stansbury, 2023). This project aims to address gaps in existing research by using machine learning models to predict students at high risk of depression and identifying which factors contribute most strongly to mental health outcomes. To explore these relationships, I used logistic regression and other machine learning models to predict depression risk based on socio-economic, psychological, and lifestyle variables. I also created a Tableau story to visually show how key predictors, such as financial stress and exercise, influence mental health outcomes. The results show that financial stress was the strongest risk factor for depression, while regular exercise and higher GPA were among the biggest protective factors. This project aims to provide insights that can help colleges develop better early intervention systems to support students who may be most vulnerable to depression.

Introduction

Mental health concerns among college students have become a growing crisis in the United States. Depression and anxiety are now the two most common mental health challenges on college campuses, with rates rising steadily over the past decade (Zhai et al., 2024). Student-athletes, though often perceived as resilient, face unique pressures balancing academics, athletics, and social expectations. These demands can increase psychological stress, and stigma surrounding mental health, especially in athletic communities, may discourage students from seeking support (Stansbury, 2023). Financial stress, loneliness, and academic pressure further contribute to mental health struggles across student populations. Exercise is often cited as a protective factor (Sharma, 2006), yet the relationship between athletic participation and mental health outcomes is complex. Few studies have used machine learning to predict which students are most at risk for depression by analyzing multiple factors together. This project addresses that gap by applying machine learning models to data from the Healthy Minds Study. By focusing on key factors like financial stress, exercise, GPA, gender, athlete status, and loneliness, the study aims to predict depression risk and support early intervention strategies.

Literature Review

Previous research has consistently documented rising rates of depression and anxiety among college students. The Healthy Minds Study highlights financial stress as one of the strongest predictors of poor mental health outcomes (Zhai et al., 2024). Exercise is well-established as a protective factor, improving mental health and reducing depression (Sharma, 2006). However, findings related to student-athletes are mixed. While Edwards et al. (2023) reported that athletes had slightly lower anxiety rates than non-athletes, athletes also faced distinct pressures that could worsen mental health. Stigma is another key barrier to mental health support, particularly among student-athletes who may fear being seen as weak (Stansbury, 2023). Loneliness has also been shown to significantly increase depression risk, while strong social support can serve as a protective factor. While these factors have been studied individually,

fewer projects have combined them to predict depression risk using machine learning. Traditional research often analyzes single relationships, but machine learning allows for the exploration of complex interactions across multiple variables. This project builds on existing findings by using predictive models to identify students most at risk including factors like stigma and major which aren't accounted for in many studies, offering a practical tool for early intervention and support.

Methodology

Data source:
The Healthy Minds Study, a national mental health survey given yearly since 2007 across 500 colleges in the US, including PHQ-9 scores and 26 variables such as GPA, therapy history, exercise habits, financial stress, and more.

Preprocessing:

- Data cleaning involved removing missing values, normalizing numerical variables, and encoding categorical features.
- Removed Columns that didn't really have an effect on depression risk, and further removed columns that had p-values much greater than .05
- Reduced the 1550 column available to 26 columns I deemed has an decent effect on depression also having statistical significance (p values less than .05)
- Created a new column called depression risk if depression score was 10 or higher which aligns with clinical standards of screening depression determining those who need intervention

Modeling:

Multiple machine learning models were tested:

Logistic Regression, Support Vector Machine(SVM), KNN Classification, Decision Tree, Random Forest Tree, ADA Boosting, XG Boosting, Gradient

Boosting

- Logistic Regression was ultimately selected as the best-performing model based on AUC-ROC evaluation.

- Exploratory data analysis (EDA) was conducted in Tableau to identify patterns and key risk factors visually.

Evaluation:

- ROC curves were generated to assess model discrimination.
- Variable importance was reviewed to interpret key risk and protective factors.
- Precision, Accuracy, and F1 scores were also compared in choosing the best models

Compare Best Boosting Model

```
[139]: print('ADA Boosting Classification\n', ada_results)
       print('Gradient Boosting Classification\n', gradient_results)
       print('XGB Classification\n', xgb_results)
```

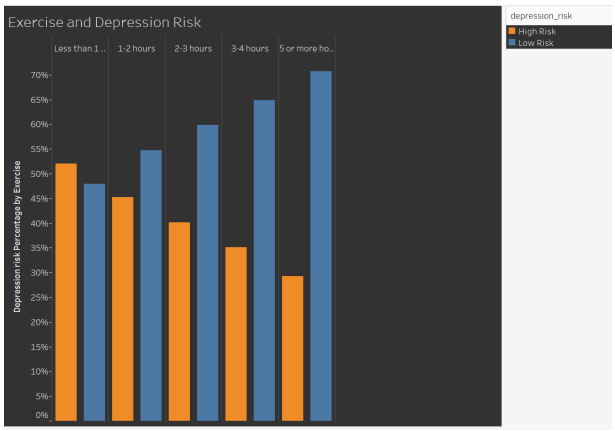
```
ADA Boosting Classification
              precision    recall  f1-score   support

           0       0.82      0.78      0.80       930
           1       0.70      0.75      0.72       638

    accuracy                           0.77      1568
   macro avg       0.76      0.76      0.76      1568
weighted avg       0.77      0.77      0.77      1568

Gradient Boosting Classification
              precision    recall  f1-score   support

           0       0.81      0.77      0.79       930
           1       0.69      0.74      0.71       638

    accuracy                           0.76      1568
   macro avg       0.75      0.75      0.75      1568
weighted avg       0.76      0.76      0.76      1568

XGB Classification
              precision    recall  f1-score   support

           0       0.80      0.77      0.78       930
           1       0.68      0.71      0.70       638

    accuracy                           0.75      1568
   macro avg       0.74      0.74      0.74      1568
weighted avg       0.75      0.75      0.75      1568
```

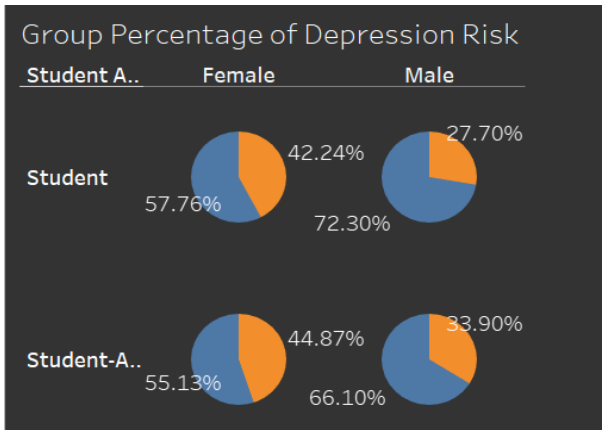
```
logistic regression result
              precision    recall  f1-score   support

           0       0.83      0.76      0.79       930
           1       0.68      0.77      0.72       638

    accuracy                           0.76      1568
   macro avg       0.76      0.76      0.76      1568
weighted avg       0.77      0.76      0.76      1568

KNN results
              precision    recall  f1-score   support

           0       0.71      0.74      0.72       930
           1       0.59      0.56      0.57       638

    accuracy                           0.66      1568
   macro avg       0.65      0.65      0.65      1568
weighted avg       0.66      0.66      0.66      1568

SVM results
              precision    recall  f1-score   support

           0       0.83      0.75      0.79       930
           1       0.68      0.77      0.72       638

    accuracy                           0.76      1568
   macro avg       0.75      0.76      0.76      1568
weighted avg       0.77      0.76      0.76      1568

decision tree results
              precision    recall  f1-score   support

           0       0.73      0.72      0.72       930
           1       0.59      0.61      0.60       638

    accuracy                           0.67      1568
   macro avg       0.66      0.66      0.66      1568
weighted avg       0.67      0.67      0.67      1568

Random forest results
              precision    recall  f1-score   support

           0       0.77      0.83      0.80       930
           1       0.73      0.64      0.68       638

    accuracy                           0.76      1568
   macro avg       0.75      0.74      0.74      1568
weighted avg       0.75      0.76      0.75      1568
```

Logistic Regression is best because it also had highest recall for class 1 which is more important because we'd rather not miss people who are at high risk

Results

- Exercise is the strongest protective factor — students who exercise more than 5 hours/week are 77.8% less likely to be at high risk for depression than those who exercise less than 1 hour a week.
- Student-athletes, despite higher activity levels, face slightly higher depression risk—likely due to academic and athletic pressure.
- Women student-athletes are especially at risk: nearly 1 in 2 are classified as high-risk.
- Financial stress is the most significant risk factor—students with "always stressful" finances are 3x more likely to be at high risk.
- Protective Factors: exercise, GPA, knowing where to get help, and low stigma.
- Logistic regression performed best, with 76% prediction accuracy.





Exercise has clear effects on depression, and Female Student Athletes were most at risk. **The orange color means a student is at high risk.**

Tableau Linked below to view visualized results:

https://public.tableau.com/views/Story_Draft_revised_51/Dashboard1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

Discussion

This study confirms and extends findings from previous research linking financial stress and exercise habits to mental health outcomes among college students. Consistent with past studies, financial stress emerged as the strongest risk factor for depression, while higher GPA and regular exercise were protective factors. The machine learning models demonstrated that predictive analytics can successfully identify students at higher risk, offering a scalable tool for early intervention. Subgroup analysis revealed that women student-athletes were particularly vulnerable to higher depression risk. However, this study also challenges some traditional assumptions about mental health stigma among student-athletes. Contrary to previous beliefs that athletes may experience higher stigma around seeking mental health help, the results show that student-athletes actually report lower levels of self-stigma compared to non-athletes. Additionally, student-athletes were slightly more likely to say they would seek help if needed and more likely to know where to find mental health resources on campus. Despite these positive indicators related to stigma and help-seeking behaviors, student-athletes still showed higher rates of depression risk in the model. This suggests that their elevated depression risk is not primarily due to stigma or lack of access to resources, but rather may stem from unique stressors related to balancing the demands of academics, athletics, and social expectations. These findings add nuance to the understanding of mental health among student-athletes and highlight the importance of addressing structural and time-related pressures in mental health interventions. Overall, this project adds new evidence to the field by combining predictive modeling with a deeper exploration of subgroup disparities, showing that machine learning approaches can reveal complex and sometimes counterintuitive patterns in mental health outcomes.

Conclusion

This project demonstrated that machine learning models, particularly logistic regression, can effectively predict depression risk among college students using psychological, social, and demographic factors. Financial stress emerged as the strongest risk factor

for depression, while higher GPA, regular exercise, and stronger help-seeking behaviors served as major protective factors. Contrary to traditional assumptions, student-athletes exhibited lower stigma levels and greater awareness of mental health resources compared to non-athletes. However, they still showed higher depression risk, suggesting that unique academic and athletic pressures—not stigma—may be driving these outcomes.

These findings highlight the potential of predictive analytics as a powerful tool for early identification and intervention for students at risk of depression. Beyond identifying at-risk individuals, machine learning models can guide universities in targeting resources more effectively toward groups that may otherwise be overlooked.

Future work could involve expanding the predictor set to capture additional non-health variables, such as financial aid status, employment, or academic load. Additionally, building models to predict anxiety risk alongside depression could provide a more comprehensive picture of student mental health challenges. With these enhancements, predictive systems could become a vital part of mental health strategies on college campuses.

References

Here are my references, click the link below to see them
https://public.tableau.com/views/AnnotatedbibliographyFinal/Dashboard1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link