# Factors Associated with Higher Rates of Anxiety and Depression in College Students

## Introduction

Our goal for this project was to implement a variety of regression models in an attempt to find important characteristics of the university and university life to predict the proportion of students with anxiety and depression. The data used for this analysis came from the American College Health Association's national annual survey, specifically the data from the spring semester of 2019, the most recently collected data from the ACHA. Since this data is self-reported and the survey was not mandatory, the results should not be looked at as causation for mental illness, but instead as factors that are associated with higher rates of mental illness or as warning signs that mental illness rates may be higher than what is believed. The results of all models were compared in a meta-analysis in order to find the most important factors that colleges should be aware of that correlate with higher rates of mental illness.

The analysis was completely performed in Python, using Google Colaboratory, VSCode Editor, and Spyder as well as packages including pandas, numpy, sklearn, and statsmodels.api. Our cleaned data and code can be found on GitHub for replication, however the raw data can only be accessed through the ACHA. The variables chosen for use do not represent the entirety of the data collected by the ACHA in their survey, but were selected in an attempt to minimize multicollinearity, strong relationships between explanatory variables, as well as for dimensionality reduction to encourage models to find significant factors. These variables are further outlined in the Data Dictionary section. Our analysis included three response variables: proportion of students reporting anxiety and/or depression, proportion of students reporting depression, and proportion of students reporting anxiety. The aim of this was to see potential factors that are associated with one type of mental illness, but not with the other. All of these response variables are quantitative variables, allowing us to deploy models including Multiple Linear Regression, Random Forest Classifier, and LASSO Regression using Grid Search methods to identify parameters.

## Data Decisions

In order to implement these models, the raw data needed to be cleaned and preprocessed. The raw data given by the ACHA was formatted as individual binary responses, meaning the data contained 60,000+ observations with 273 columns, since many of the questions on the survey have multiple levels. We decided to group the data by school, using the anonymous school ID assigned by the ACHA in order to maintain privacy, and find proportions of positive responses. A handful of the variables were information regarding the objective characteristics of the college (ex. location, size, etc) which were consistent for all respondents from that college. Many variables were dropped to once again reduce the dimensionality of the data as well as remove columns with large amounts of missing data since the respondent was not required to answer every question on the survey.

Many of the variables selected by the team had to be further transformed in order to reduce multicollinearity or to accurately reflect the relationship of the levels. Four variables regarding safety in the community and on campus were averaged out to reflect the general perception of safety in the area since the variables were highly correlated with each other. The variable containing information about the region of the school was split into dummy variables as well as the variable containing information about the public/private standing and religious affiliation of the school.

After grouping the data by school, we were left with a dataset containing 98 observations, one for each college represented, and 24 explanatory variables. To increase reliability, we split our data into a 20% training set and a 80% testing set. To do this without favoring schools with higher respondent numbers, we subsetted 20% of each school’s responses and found the proportions for the training data, then found the proportions for the remaining 80% for the testing data. As a result, each set contained an observation for each of the 98 schools. The training data was used for variable selection, identifying significant or important variables that were then modelled onto the testing data. This method was used for all three models for each response variable, creating nine models in total.

Finally, in order to implement the Random Forest Classifier, the response variables were transformed into binary responses: above average or below average. This was done in an attempt to highlight overall relationships in the small sample set that may be missed in the regression models. To do this, a mean value of each response was calculated using the entire dataset.

## Data Dictionary

| Covariate Name | Description |
|---|---|
| campus_setting | population size of area campus is located |
| school_size | enrollment size of school |
| cigarettes_last_30 | proportion of responding students who smoked cigarettes in the last 30 days |
| ecigs_last_30 | proportion of responding students who smoked e-cigarettes in the last 30 days |

| | |
|---|---|
| physical_fight_last_year | proportion of responding students who were involved in a fight within the last 12 months |
| unprescribed_antidepressants | proportion of responding students who use unprescribed antidepressants |
| unprescribed_ed | proportion of responding students who use unprescribed erectile dysfunction medication |
| unprescribed_painkillers | proportion of responding students who use unprescribed painkillers |
| unprescribed_sedatives | proportion of responding students who use unprescribed sedatives |
| unprescribed_stimulants | proportion of responding students who use unprescribed stimulants |
| anorexia | proportion of responding students who have diagnosed or treated for anorexia |
| bulimia | proportion of responding students who have diagnosed or treated for bulimia |
| difficult_academics | proportion of responding students who have had difficulty with academics in the last 12 months |
| difficult_career | proportion of responding students who have had difficulty with career-related issues in the last 12 months |
| difficult_finances | proportion of responding students who have had difficulty with finances in the last 12 months |
| seek_help | proportion of responding students who would consider seeking help from a mental health professional |
| safety_average | perception of safety in community and on campus |
| northeast_region | located in northeast region |
| south_region | located in south region |
| west_region | located in west region |
| midwest_region | located in midwest region |
| public | public college |
| private | private college, not religiously affiliated |
| private_religious | private college, religiously affiliated |

## Description of Models

### Multiple Linear Regression

Multiple Linear Regression (MLR) fits two or more explanatory variables against the response by fitting a linear equation. Also known as the least squares model (an extension of OLS), MLR takes residuals and sums them up to zero. Mean Squared Error is derived from this concept and is also known as the variance, the square root of this variance is the standard deviation of the model. If the model is performing well the residuals should sum up to as stated, and be normally distributed. We will also be evaluating Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). For this model, variables were chosen to be included in the model if the resulting p-value associated with the variable was below 0.10. This means that there is less than a 10% chance that the relationship would occur if there was actually no correlation between the variable and the response. The standard cutoff is 0.05, however given the previously mentioned interpretability issues regarding the survey-format, we increased the cutoff to potentially show more relationships.

### Random Forest

A Random Forest is a type of machine learning model built off of numerous structures called decision trees. A decision tree is a modeling approach in which a tree's branches represent the observations of an item, while the leaves represent the predicted value. To decorrelate all these features being present on just one decision tree, Random Forest aggregates many of these decision trees so that the factors we are evaluating are distributed and not too highly correlated which can lead to variance being present in our model. This model can be used as a classifier or regressor, but in our case we will be using it for classification. To fully evaluate this model we are using RMSE and Gini Index/Entropy to be able to identify significant predictors through a normal and permutative approach. For this model, variables were chosen to be included in the model if their feature importance was above 0.001.

### LASSO Regression

LASSO Regression is a model that performs both variable selection and regularization, meaning it is often used as a form of dimensionality reduction or variable selection. An insignificant feature with a large coefficient is penalized and the coefficient is set to zero. One characteristic of LASSO Regression models is the parameter alpha which needs to be chosen before implementation. An alpha of 0 indicates that the LASSO will perform the same as OLS Linear Regression, but raising the alpha level will increase the number of coefficients that get set to zero. In order to find the appropriate alpha level, a Grid Search was performed, using multiple alphas values and finding which value produced the best score. The result of this Grid Search led us to use an alpha value of 0.001 for all three response variables. For this model, variables were chosen to be included in the model if their coefficients were nonzero, meaning that the LASSO Regression found them to be significant.

## Results and Metrics

After running all nine models, the results were compared using the number of significant or important covariates, variance represented by MAE, and accuracy of predictions. The MAE was chosen to represent the variance since the MSE and RMSE give more weight to those further from the mean and in our case, this is not necessary. It's important to note as well that the trends in the MAE values were replicated in the RMSE and MSE values. The number of covariates was included as a measure of interpretability since our end goal is to give a recommendation of key associations. Figure 1 contains these metrics for all 9 models.

Figure 1

| Response Variable | Model Description | Number of Covariates | Variance (MAE) | Accuracy |
|---|---|---|---|---|
| Anxiety and Depression Combination | Linear Regression | 5 | 0.031675 | 88.56 % |
| | Random Forest | 10 | 0.268824 | 80.0 % |
| | LASSO Regression | 19 | 0.327757 | 75.32 % |
| Depression | Linear Regression | 4 | 0.026438 | 86.74 % |
| | Random Forest | 10 | 0.275819 | 79.41 % |
| | LASSO Regression | 17 | 0.345911 | 73.9 % |
| Anxiety | Linear Regression | 5 | 0.030090 | 87.53 % |
| | Random Forest | 10 | 0.269934 | 79.98 % |
| | LASSO Regression | 17 | 0.345911 | 73.9 % |

By comparing these models, we found that the most accurate predictions were found using the Multiple Linear Regression models, indicated by the accuracy levels as well as the variance. These models were also much more reduced models with higher interpretability, therefore we decided to use these covariates in our recommendations. LASSO and Random Forest models still performed well with high accuracy rates and low variance, however they may have struggled with overfitting with the high number of covariates. Figures 2-4 show the covariates included in the 9 models.

Figure 2

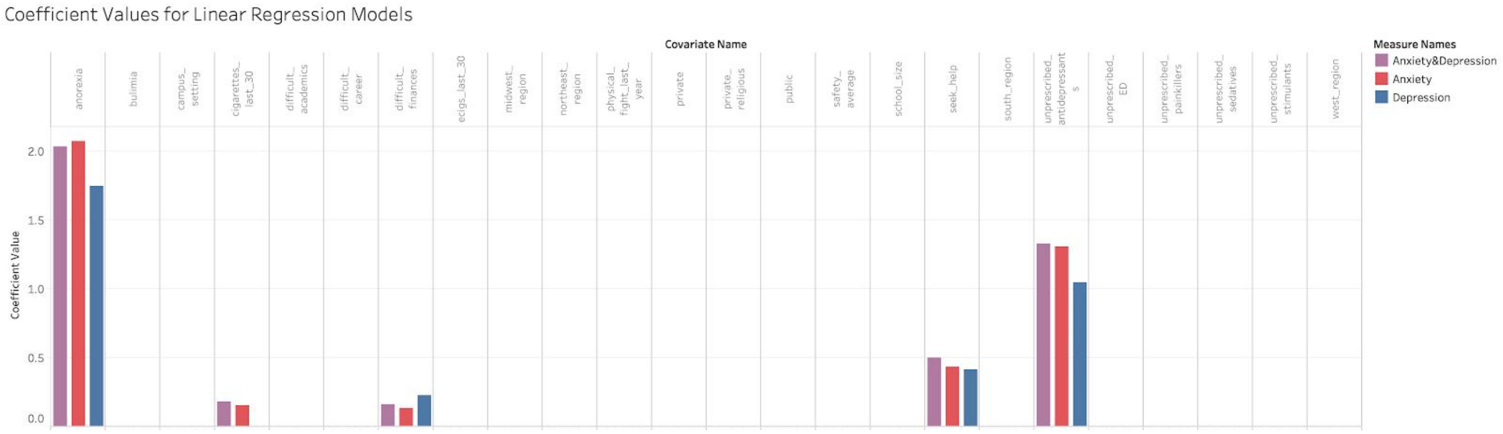


Figure 3

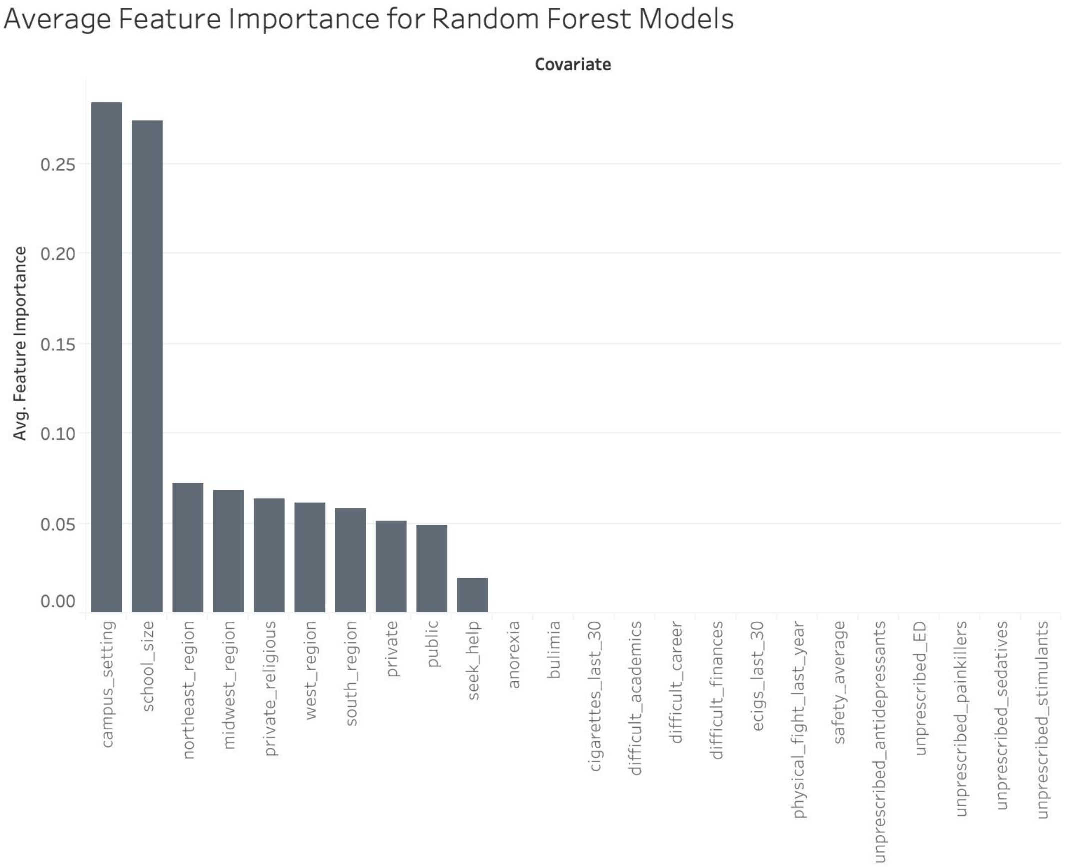
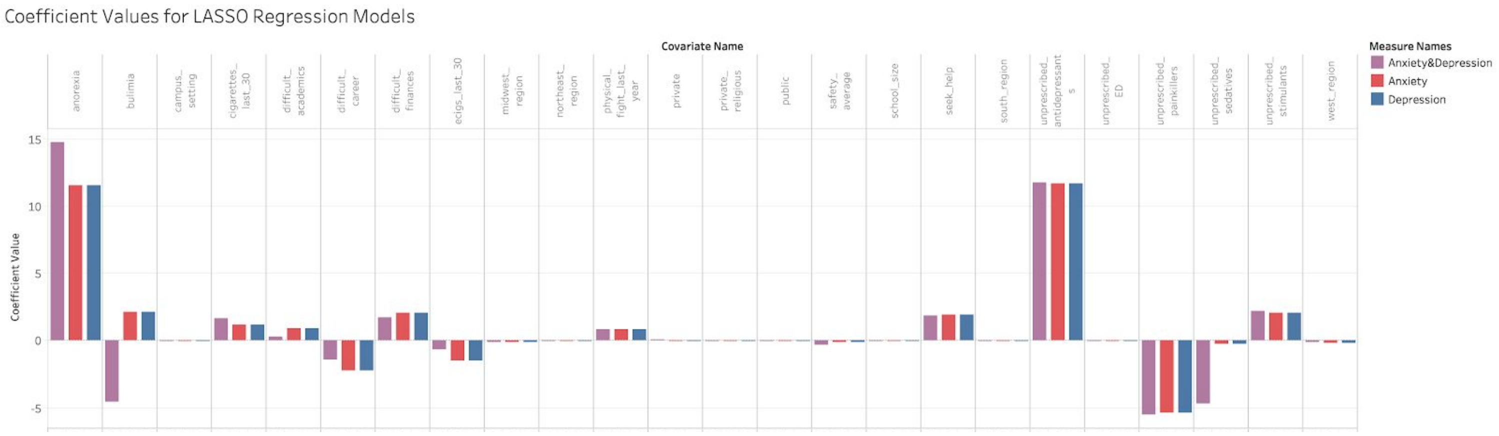


Figure 4



Our statistical analysis found that the factors most associated with elevated rates of anxiety and depression were the rate of anorexia, the consistent use of cigarettes, the abuse of antidepressants, financial difficulties, and the perception of being able to seek help from a medical professional. These factors are associated with both mental illness individually as well as the overall rate.