The European Journal of Humanities and Social Sciences 2023, No 5
ISSN 2414-2344

PREMIER Publishing
ppublishing.org

Section 5. Psychology

DOI: 10.29013/EJHSS-23-5-39-46

RISK FACTORS FOR MENTAL HEALTH AMONG ADOLESCENTS DURING THE COVID-19 PANDEMIC

Anthony Ju [1]

[1] Sharon High School, MA, USA

Cite: *Anthony Ju. (2023). Risk Factors for Mental Health Among Adolescents During the Covid-19 Pandemic. European Journal of Humanities and Social Sciences 2023, No 5.* https://doi.org/10.29013/EJHSS-23-5-39-46

Abstract

The COVID-19 pandemic has greatly disrupted the daily life of adolescents nationalwide. According to the Centers of Disease Control and Prevention (CDC), more than half of students experienced emotional abuse in the home, more than 1 in 3 high school students experienced poor mental health during the pandemic and nearly half of students felt persistently sad or hopeless.

In this research, we investigated possible risk factors for mental health during the COVID-19 pandemic among adolescents and identified the most significant positive and negative factors through logistic regression. We used the 2021 Adolescent Behaviors and Experiences Survey data with features ranging from demographic information to the adolescents' family condition. The response variable is whether an adolescent has good or bad mental health during the COVID-19 pandemic.

After processing the dataset, we built a logistic regression model to predict whether an adolescent is likely to develop meatal health problems. By investigating the logistic regression coefficients, we found that parents' attitude toward the adolescent, the gender of the adolescent, and the family's financial ability to cover food are all risk factors. The logistic regression model has achieved an AUROC score of 0.78, with 68.2% true positive rate (TPR) and 28.0% false positive rate (FPR). This predictive model is helpful for healthcare professionals to identify and reduce the risk for the adolescents that are prone to the mental problems during the pandemic.

Keywords: *mental health, COVID-19 pandemic, logistic regression model*

Introduction

The COVID-19 pandemic has had a seismic effect on communities across the country, and young people have been especially impacted by the ways in which their everyday lives have been altered. The disruptions were widespread – school buildings closed, opportunities for connecting with peers were

limited, communities were dealing with loss and upheaval.

While the pandemic has affected all students, the experiences of disruption and adversity have not affected all students equally. According to CDC, more than 1 in 3 high school students experienced poor mental health during the pandemic and nearly half of students felt persistently sad or hopeless (Centers for Disease Control and Prevention (2022, March 31). In addition, the daily life of many adolescents has also been disrupted due to the pandemic, as more than half of students experienced emotional abuse in the home and more than 10% reported physical abuse in the home (Centers for Disease Control and Prevention (2022, March 31). Therefore, it is of great importance for healthcare professional to identify children that are at high risk for developing mental problems and help address problems at an early stage. To fulfill this task, this report discussed the machine learning techniques that can be applied to build predictive models on whether a child will have mental health issues.

Specifically, we pre-processed the dataset, built a logistic regression model, and investigated factors most related to the goodness of mental health during the COVID-19 pandemic. We also measured the model performance using various validation techniques and analyzed the model coefficients to find the variables that contribute most to our predicted results.

Method

Data

We used 2021 Adolescent Behaviors and Experiences Survey (ABES) data for this study. The ABES is a 110-question online survey completed by US high school students in early-mid 2021. It is a national survey conducted by Centers for Disease Control and Prevention (CDC) that provides rich data on health-related experiences and behaviors among high school students and was designed to assess the impacts of the COVID-19 pandemic on adolescents. In addition, ABES is also the first nationally representative survey looking at the effects of the COVID-19 pandemic on the health of adolescents. The 2021 ABES data contains 7,705 complete data samples. We used the following variables as independent variables.

Table 1. *Features used for analysis*

Variable	Description	Comments
Q1	How old are you?	Range: 12–18
Q2	What is your sex?	0: Female, 1: Male
Q4	Are you Hispanic or Latino?	0: Yes, 1: No
Q101	During the COVID-19 pandemic, did any adult in your home lose their job?	0: Yes, 1: No
Q102	During the COVID-19 pandemic, did you lose your paying job?	0: Yes, 1: No
Q103	During the COVID-19 pandemic, how often did you go hungry because there was not enough food in your home?	Higher value indicates higher frequency
Q104	Do you agree that doing your schoolwork was more difficult during the COVID-19 pandemic than before?	Higher value indicates stronger agree
Q105	During the COVID-19 pandemic, how often did any adult in your home swear at you, insult you, or put you down?	Higher value indicates higher frequency
Q106	During the COVID-19 pandemic, how often did any adult in your home hit, beat, kick, or physically hurt you in any way?	Higher value indicates higher frequency

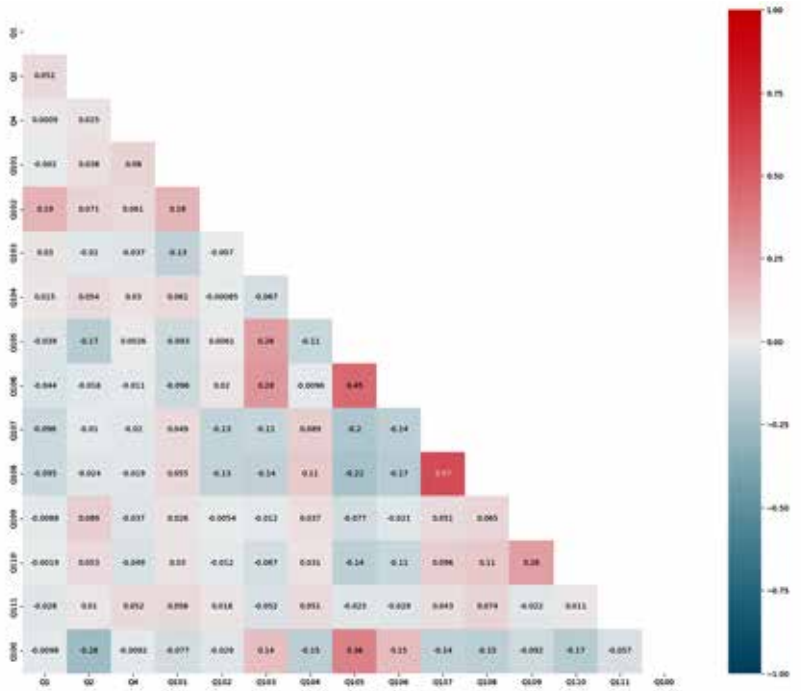
Q107	Do you agree that you drank more alcohol during the COVID-19 pandemic than before?	Higher value indicates stronger agree
Q108	Do you agree that you used drugs more during the COVID-19 pandemic than before?	Higher value indicates stronger agree
Q109	During the COVID-19 pandemic, did you get medical care from a doctor or nurse using a computer, phone, or other device (also called telemedicine)?	0: Yes, 1: No
Q110	During the COVID-19 pandemic, did you get mental health care using a computer, phone, or other device (also called telemedicine)?	0: Yes, 1: No
Q111	During the COVID-19 pandemic, how often were you able to spend time with family, friends by using a computer, phone, or other device?	Higher value indicates higher frequency
Q100	During the COVID-19 pandemic, how often was your mental health not good?	Higher value indicates higher frequency and worse mental health

The dependent variable is a binary feature coded as "Q100," which indicates the respondent's mental health. Responses of "most of the time" and "always" were recoded as having poor mental health and were used as positive samples in this study.

Exploratory Analysis

A correlation graph is a primitive yet straightforward representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes by using visual thinning and correlation-based variable ordering. Moreover, the matrix cells can be shaded or colored to show the correlation value. The positive correlations are shown in red, while the negative correlations are shown in blue; the darker the hue, the greater the magnitude of the correlation.

Figure 1. *Correlation among variables*



The graph above shows that the dependent variable (has bad mental health) has the highest positive correlation with Q105, while having the highest negative correlation with Q2 (sex), indicating that family members' attitude toward the adolescent and the gender of the adolescnet play a significant role in his or her mental health. In addition, we discovered that the variable Q103 and Q106 also have positive correlation with the dependent variable.

In addition, the correlation graph also provides valuable information regarding the relationship among features. For example, the correlation between Q108 and Q107 is 0.57, indicating that the two variables are significantly positively correlated and adolescnets who abuse alcohol are more likely to abuse drugs as well during the COVID-19 pandemic.

Statistical Method

Pre-processing

The data set is pre-processed in this step to improve both the training speed and accuracy. As most machine learning algorithms are not able to deal with missing values, all the data points with missing entries or invalid responses to the dependent variable are excluded from training and testing. In addition, as different features usually have remarkedly different value ranges, we applied the feature standardization technique to transform different features into comparable scales. This measure ensures that different features weigh equally in the training process. For each feature, its mean value and standard deviation are first computed as *avg(x)* and *std(x)*. Then each data point x with respect to that feature is replaced by y_i calculated as:

$$y_i = \frac{x - avg(x)}{std(x)}.$$

Finally, the dataset is partitioned into two datasets for training and test purposes: the training dataset (70%) for model development and the test dataset (30%) for model test and validation.

Logistic Regression

Logistic regression models were used to calculate the predicted risk. Logistic regression is a part of a category of statistical models called generalized linear models, and it allows one to predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a combination of these. Typically, the dependent variable is dichotomous, and the independent variables are either categorical or continuous.

The logistic regression model can be expressed with the formula:

$$\ln\left(\frac{y}{1-y}\right) = w_0 + w_1 x_1 + \ldots w_m x_m$$

In the logistic regression, y is the probability of the sample classified as the positive class, and each feature x_i has its specific weight w_i, where w_0 is the intercept while w_1 through w_m are the coefficients of the independent variables.

Our task is to find a set of parameters $w_0,\ldots,w_m$ such that the loss function between the output y and the actual values u is minimized.

$$l(y,u) = \|y-u\|_2^2$$

In addition, we applied elastic-net regularization to constrain model complexity and prevent model over-fitting problems with L-1 ratio equaling 0.5.

Model Validation

Consider a two-class prediction problem, where the outcomes are labeled either as positive or negative. There are four possible outcomes from a binary classifier. If the outcome from a prediction is positive and the actual value is also positive, then it is called a true positive (TP); however, if the actual value is negative, then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are negative, and false negative (FN) is when the prediction outcome is negative while the actual value is positive. In this way, the true positive rate (TPR) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN}$$

And the false positive rate (FPR) can be calculated as:

$$FPR = \frac{FP}{TN + FP}$$

A confusion matrix is a table that allows visualization of the performance of an algorithm. Each row of the matrix represents the

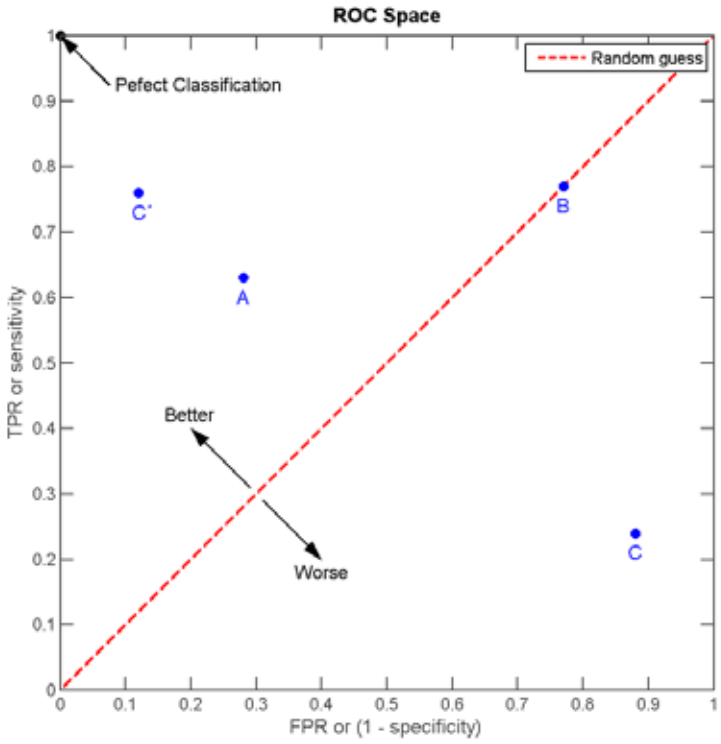
instances in an actual class while each column represents the instances in a predicted class. An example of the confusion matrix and the meaning of each cell within the table can be found in the graph below. Typically, the confusion matrix of a good predictive model has high true positive and true negative rates

Figure 2. *Confusion matrix example*

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings (Google. Classification: ROCC urve and AUC). The best possible prediction method would yield a point in the upper left corner of the ROC space. A random guess would give a point along a diagonal line from the left bottom to the top right corners. Points above the diagonal represent better than random classification results, while points below the line represent worse than random results. A sample ROC plot is shown in Figure 2. In general, ROC analysis is one tool to select possibly optimal models and to discard suboptimal ones independently from the class distribution. Sometimes, it might be hard to identify which algorithm performs better by directly looking at ROC curves. Area Under Curve (AUC) overcomes this drawback by finding the area under the ROC curve, making it easier to find the optimal model.

Figure 3. *A sample ROC plot*



Results

Confusion matrix and ROCcurve

Figure 4 shows the confusion matrix of the logistic regression model. The upper left region is true negative, the upper right region is false positive, the lower left region is false negative, and the lower right region is true positive. As shown in (Figure 4), the logistic regression model has a relatively high (~68.2%) true positive rate and a relatively low (~28.0%) false positive rate.

Figure 4. *Confusion matrix of the predicted results*

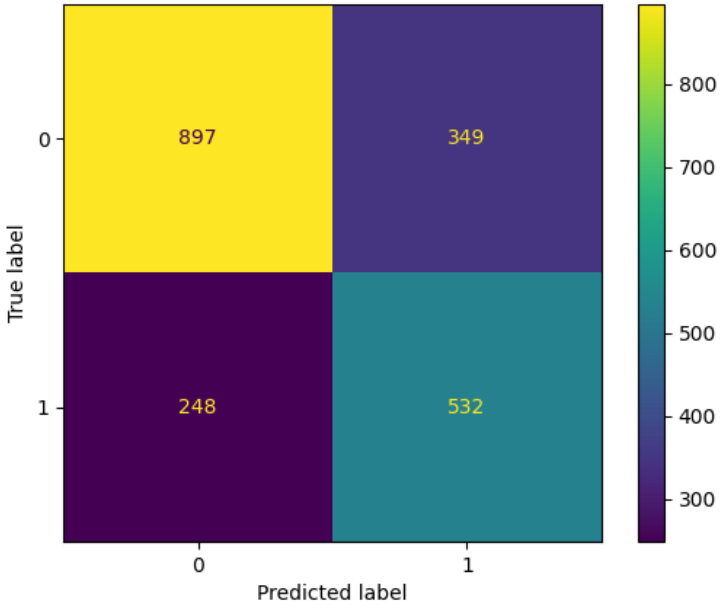
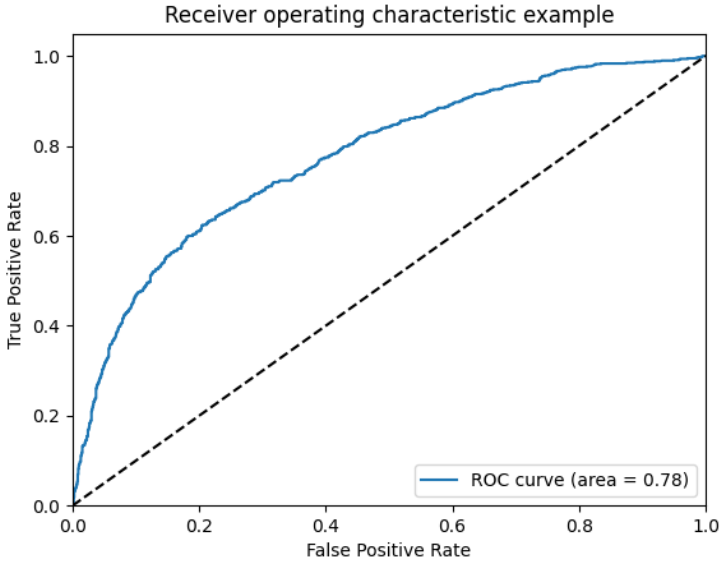


Figure 5 displays the ROC curve for the logistic regression model. It can be concluded that the model has results much better than random guessing and the AUROC score is 0.78.

Figure 5. *The ROC curve for the logistic regression model*



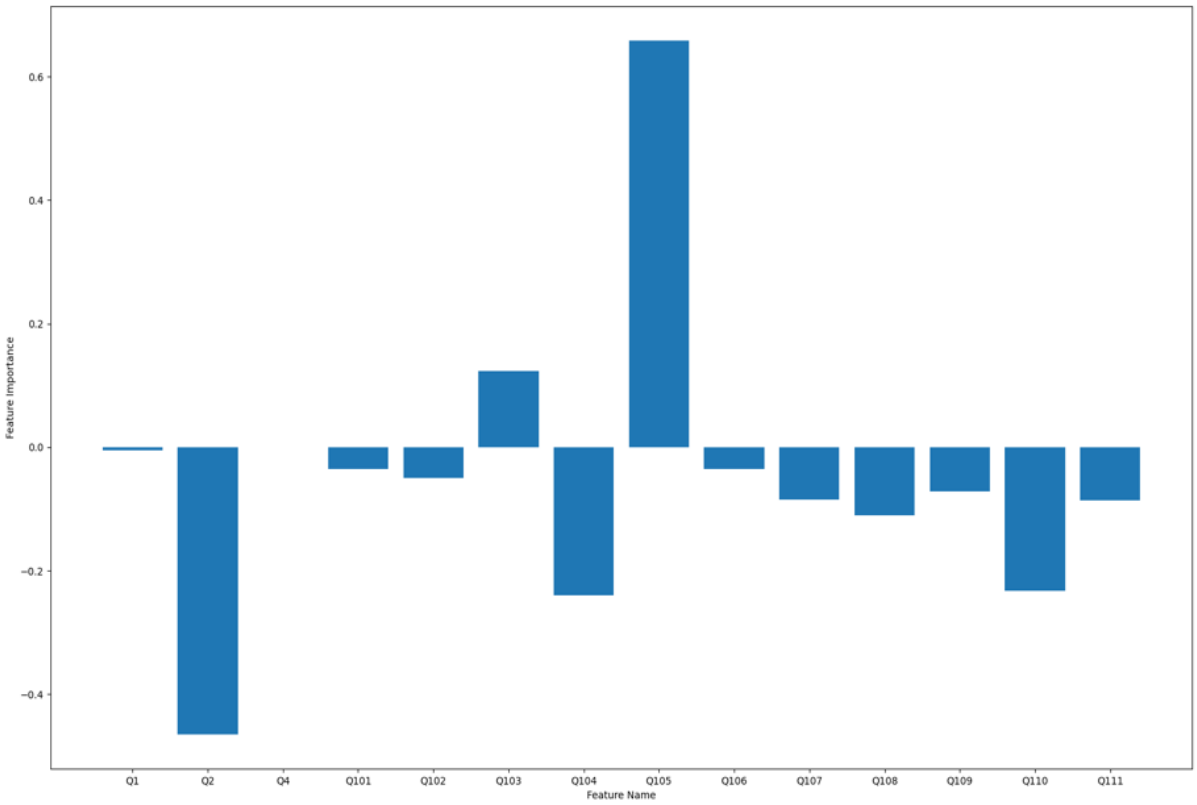
Feature Importance

Like in linear regression, the coefficients in the logistic regression model also provide valuable information about the direction and magnitude of the impact of each input variable on the dependent variable. In other words, these coefficients can provide the basis for a crude feature importance score. The figure below shows the coefficient of each input variable.

The chart below shows that variables Q2, Q103, Q104, Q105 all have relative-

ly large impact on the dependent variable (adolescents' mental health). These results align with our findings from the correlation analysis. By analyzing those relationships in detail, we also found that being female, having more difficult schoolwork compared to the pre-pandemic era, having abusing family members, and living in a family with insufficient food are all risk factors for developing mental health problems.

Figure 6. *The importance score for each feature*



Discussion

This study intends to build a predictive model to investigate the factors most related to the development of mental problems during the COVID-19 pandemic among adolescents. Through preliminary analysis, we discovered that family members' attitude and family's financial ability to cover basic living expenses are all risk factors for the adolescents' mental health. A logistic regression model was built, and the AUROC score is 0.78, indicating that the model has achieved relatively good performance in making accurate predictions on whether a child will develop mental issues. The predictive model suggests that Q105 (family member's attitude toward the adolescent) and Q2 (the gender of the adolescent) are top risk factors for mental health. A possible explanation of the results might be that adolescents with parents or family members frequently insulting them may receive much less love and encouragement and thus are more likely to develop mental health issues. This predictive model is helpful for healthcare professionals to identify children that are at higher risk for mental diseases and come up with specific plans to reduce their risk for long-term impacts.

One limitation of this study is that data entries with missing values are excluded from the analysis. This is a timesaving but defective approach. Depending on the number of data entries with missing values, we may have removed too many sample points, resulting in losing valuable information for the model to learn the critical relationship between the independent and dependent variables. Therefore, for future studies, we may use more advanced techniques such as mean value imputation or k-nearest neighbors (kNN) to impute a value for the missing entries. The mean value imputation method completes missing values with the mean of the entire feature. This is a simple and effec-

tive way to make those entries usable by the logistic regression model. Other techniques include the k-nearest neighbor approach, which replaces missing values with the mean of k (a value assigned by users) nearest neighbors of that sample (Kozma, Laszloб 2008). This technique requires more effort but can generally achieve better performance.

References

Centers for Disease Control and Prevention. (2022, March 31). Adolescent behaviors and experiences survey (ABES). Centers for Disease Control and Prevention. Retrieved July, 24, 2022. URL: from https://www.cdc.gov/healthyyouth/data/abes.htm

Google. Classification: ROC Curve and AUC | Machine Learning Crash Course. Accessed November, 25, 2021. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Kozma, Laszlo. “k Nearest Neighbors algorithm (kNN).” Helsinki University of Technology (2008).

submitted 22.08.2023;
accepted for publication 20.09.2023;
published 8.10.2023

Contact: Anthony.ju71651@gmail.com