The European Journal of Humanities and Social Sciences 2024, No 5
ISSN 2414-2344

PREMIER Publishing
ppublishing.org

Section 5. Psychology

DOI:10.29013/EJHSS-24-5-52-60

UNMASKING RISK FACTORS OF BULLYING BEHAVIORS AMONG ADOLESCENTS IN SCHOOLS USING LOGISTIC REGRESSION

Yixuan Yang [1]

[1] Shenzhen College of International Education, China

Cite: *Yixuan Yang. (2024). Unmasking Risk Factors of Bullying Behaviors Among Adolescents in Schools Using Logistic Regression. European Journal of Humanities and Social Sciences 2024, No 5.* https://doi.org/10.29013/EJHSS-24-5-52-60

Abstract

The bulling behaviors among adolescnets has become a serious issue in the United States. According to the U.S. Department of Education's National Center for Education Statistics (NCES), one out of every five (20.2%) students report being bullied at school and 41% of students who reported being bullied at school indicated that they think the bullying would happen again.

In this research, we investigated possible risk factors for bullying behaviors at school 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 been bullied at school during the past 12 months.

After processing the dataset, we built a logistic regression model to predict whether an adolescent is likely to be bullied. By investigating the logistic regression coefficients, we found that parents' attitude toward the adolescent, gender, race, and the adolescents' relationship to people at school are all risk factors. Specifically, we found that female white adolescents are more likely to be bullied at school. The logistic regression model has achieved an AUROC score of 0.74, with 62.1% true positive rate (TPR) and 30.9% 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 be bullied and thus developing mental health related issues.

Keywords: *bullying behaviors, model to predict whether, risk factors, family member's attitude toward the adolescent, difficulty in concentration*

1. Introduction

Bullying is unwanted, aggressive behavior among school aged children that involves a real or perceived power imbalance. Nowadays, the bulling behaviors among adolescnets has become a common yet serious issue in the United States. According to the U.S. Department of Education's Nation-

al Center for Education Statistics (NCES), one out of every five (20.2%) students report being bullied at school and 41% of students who reported being bullied at school indicated that they think the bullying would happen again (National Center for Educational Statistics. 2019). In addition, bullying behaviors can have serioes impacts on adolescents' development. According to NCES, students who experience bullying are at increased risk for depression, anxiety, sleep difficulties, lower academic achievement, and dropping out of school, and those who experience bullying are twice as likely as non-bullied peers to experience negative health effects such as headaches and stomachaches (National Center for Educational Statistics. 2019). Therefore, it is of great importance for healthcare professional to identify adolescents that are at high risk for being bullied at school 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 an adolescnts will be bullied and meanwhile identified top risk factors associted with such behaviors.

Specifically, we pre-processed the dataset, built a logistic regression model, and investigated factors most related to bullying behaviors at schools among adolescents. 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.

2. Method

2.1 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
Q5	What is your race?	0: American Indian, 1: Asian, 2: Black, 3: Native Hawaiian, 4: White
Q19	Have you ever been physically forced to have sexual intercourse when you did not want to?	0: Yes, 1: No
Q65	What's your sexual orientation?	0: Straight, 1: Gay or lesbian, 2: Bisexual, 3 or higher: others
Q66	How do you describe your weight?	Higher value indicates more overweighted
Q101	During the COVID-19 pandemic, did any adult in your home lose their job?	0: Yes, 1: No

Variable	Description	Comments
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
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
Q113	Do you agree or disagree that you feel close to people at your school?	Higher value indicates stronger disagree
Q114	How often do your parents or other adults in your family know where you are going or with whom you will be?	Higher value indicates higher frequency
Q115	Because of a physical, mental, or emotional problem, do you have serious difficulty concentrating, remembering, or making decisions?	0: Yes, 1: No
Q92	During the past 12 months, have you ever been bullied on school property?	0: No, 1: Yes

The dependent variable is a binary feature coded as "Q92," which indicates whether the respondent has been bullied at school in the past 12 months.

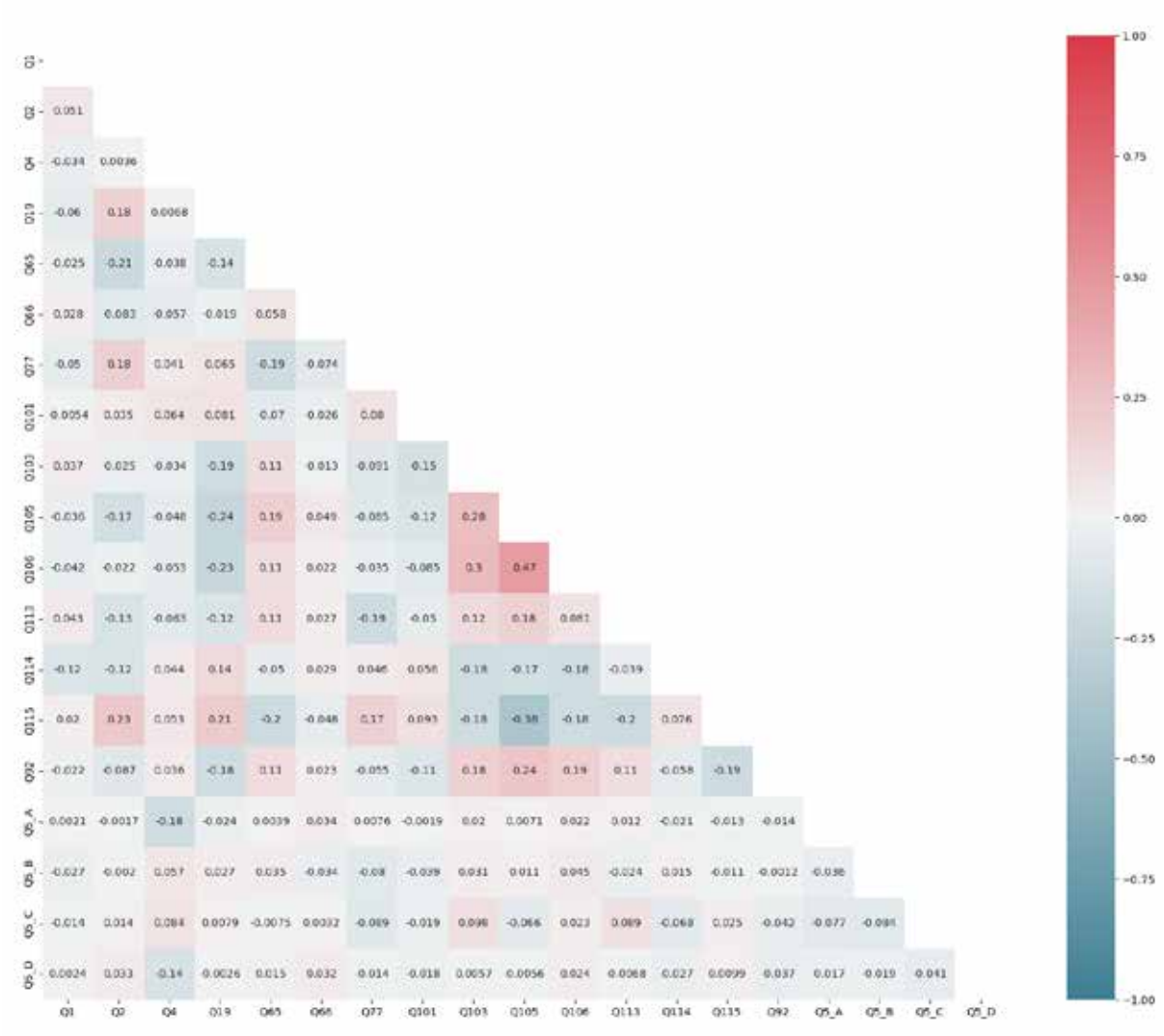
2.2 Exploratory Analysis

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.

The graph above shows that the dependent variable (has been bullied at school in the past 12 months) has the highest positive correlation with Q105, while having the highest negative correlation with Q115, indicating that family members' attitude toward the adolescent and the difficulty in concentration play a significant role in their 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 Q105 and Q106 is 0.47, indicating that the two variables are significantly positively correlated and adolescnets whose family memebrs treat them badly verbally are also likely to beat them physically.

Figure 1. *Correlation among variables*



2.3 Statistical Method

2.3.1 Pre-processing

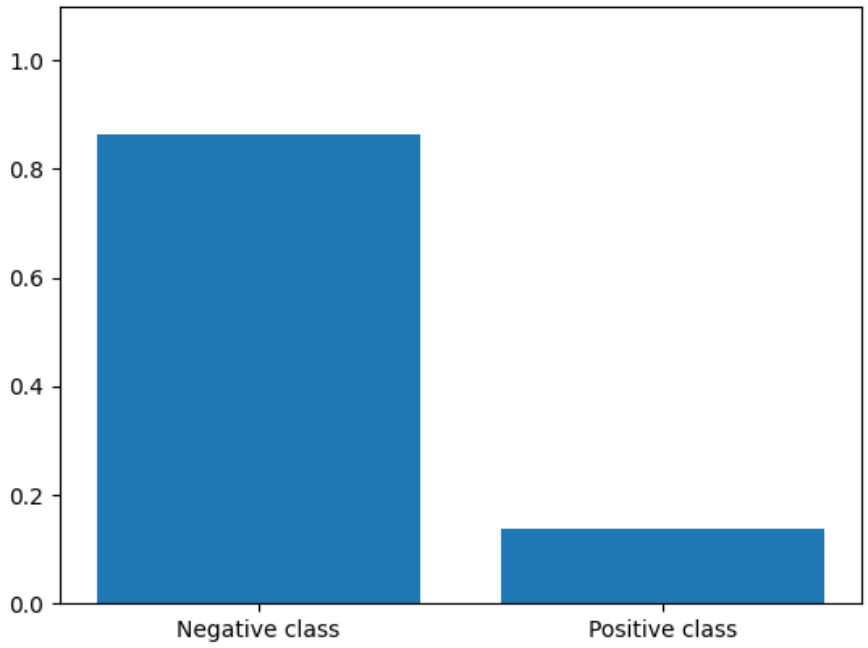
The data set is pre-processed in this step to improve both the training speed and accuracy. As most machine learning algorithms are not able to deal with missing values, all the data points with missing entries or invalid responses to the dependent variable are excluded from training and testing. In 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:

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

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.

As the distritbution of the positive class and negative class is highly unbalanced in the training set, we further appleid the over-sampling techniqeu to rebalnce the data. Over-sampling is done by randomly selecting samples from the minority class, duplicating it, and then putting back into the dataset till both classes are balanced.

Figure 2. *Distribution of class in the training set*



2.3.2 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:

$$\ln\left(\frac{\mathrm{h_w}\left(x^{\mathrm{i}}\right)}{1-\mathrm{h_w}\left(x^{\mathrm{i}}\right)}\right)=w_0+w_1x_1+\ldots+w_mx_m$$

In the logistic regression, $\mathrm{h_w}\left(x^{\mathrm{i}}\right)$ 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 cross-entropy cost function between the output $\mathrm{h_w}\left(x^{\mathrm{i}}\right)$ and the actual values y^i is minimized.

$$J(w)=-\frac{1}{m}\left[\sum_{i=1}^{m}y^i\log\left(\mathrm{h_w}(x^i)\right)+(1-y^i)\log(1-(\mathrm{h_w}(x^i))\right]$$

is minimized.

In addition, we applied elastic-net regularization to constrain model complexity and prevent model over-fitting problems with L-1 ratio equaling 0.5. We applied the grid search technique with 5-fold cross validation to find the optimal regularization strength. The 5-fold cross-validation divides the training data into five equal partitions and conducts five separate experiments to assess the model's performance with different regularization parameters. In each experiment, four folds are used for training, and one is reserved for validation, cycling through all the folds so that each is used once for validation. The set of regularization parameters that gives the best average performance across all experiments is then selected.

2.3.3 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:

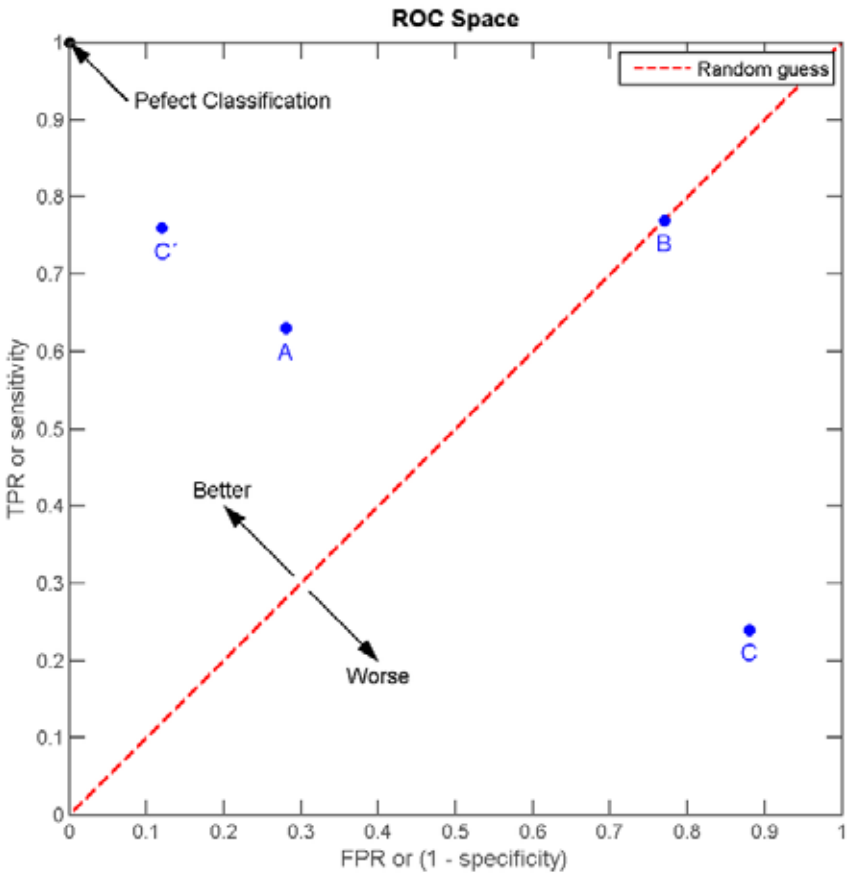
$$FPR = \frac{FP}{TN + FP}$$

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 3. *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)

Figure 4. *A sample ROC plot*



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: ROC Curve and AUC | Machine Learning Crash Course. Accessed November 25, 2021). 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 4. 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.

3. Results

3.1 Confusion matrix and ROC curve

Figure 5 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 5, the logistic regression model has a relatively high (~62.9%) true positive rate and a relatively low (~30.8%) false positive rate.

Figure 5. *Confusion matrix of the predicted results.*

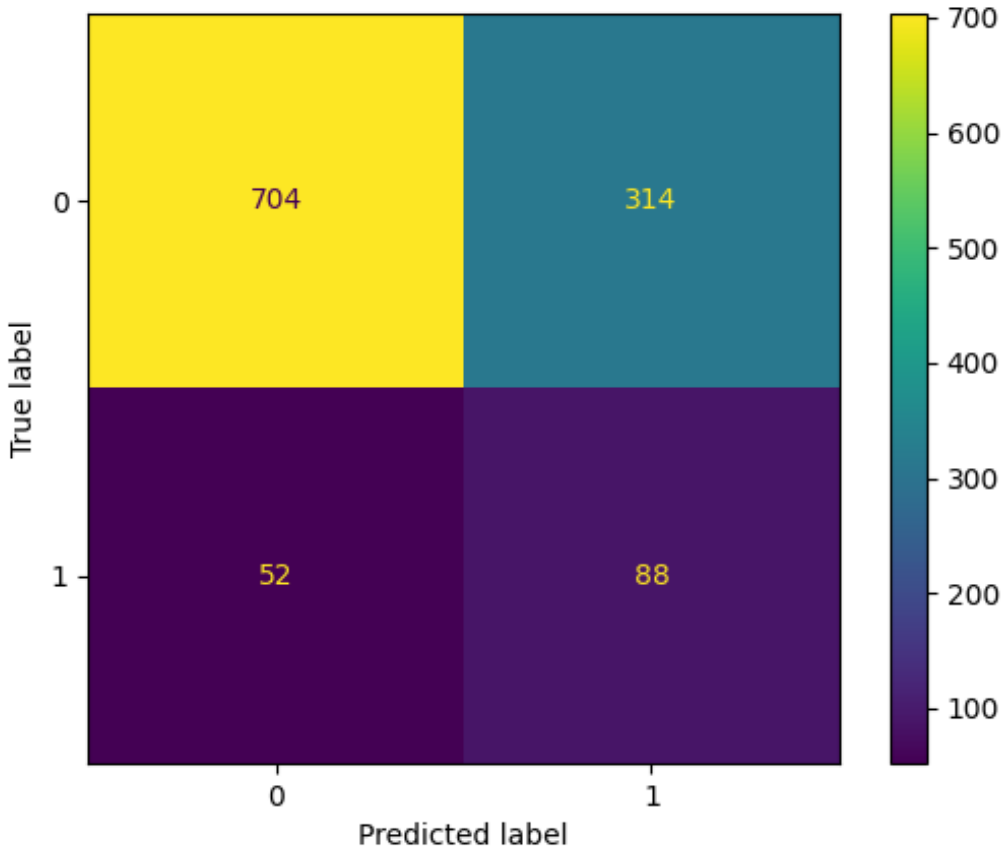
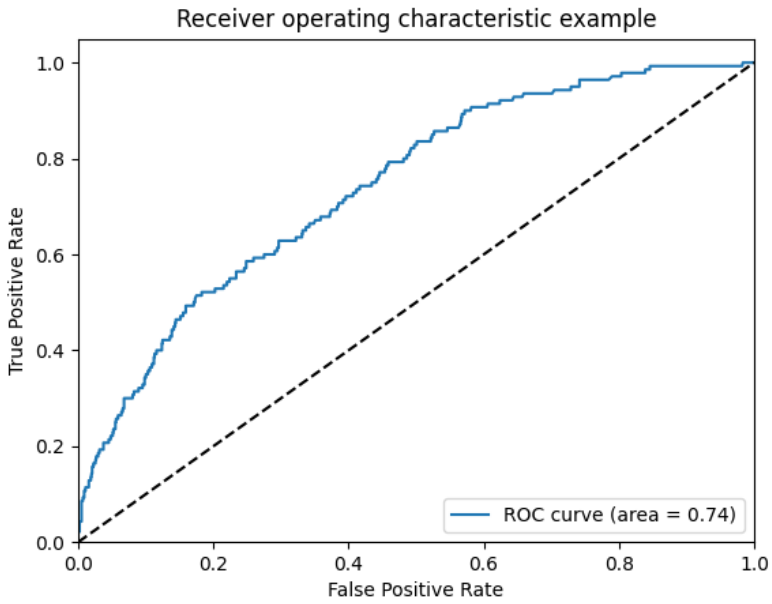


Figure 6 displays the ROC curve for the logistic regression model. It can be concluded that the model has results much better than random guessing and the AUROC score is 0.74.

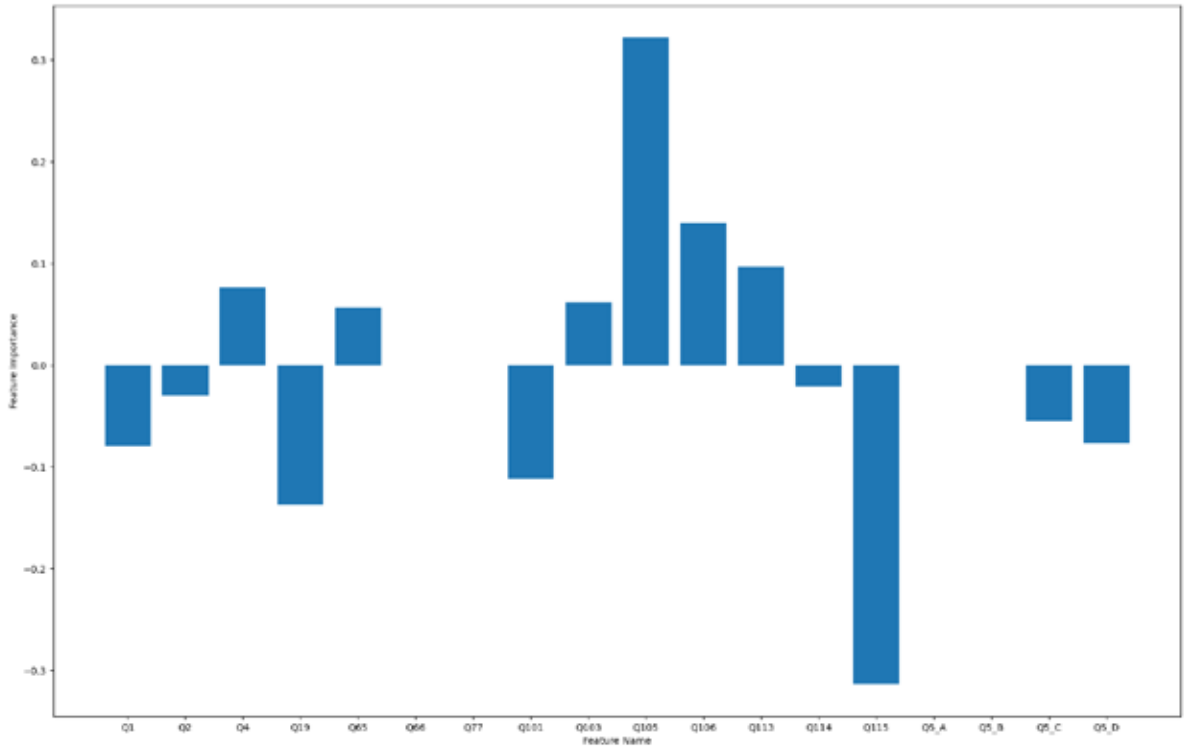
Figure 6. *The ROC curve for the logistic regression model*



3.2 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.

Figure 7. *The importance score for each feature*



The chart below shows that variables Q19, Q105, Q106, Q115 all have relatively large impact on the dependent variable (adolescents' being bullied at school). These

results align with our findings from the correlation analysis. By analyzing those relationships in detail, we also found that being female, white, having more difficulty in concentration, having abusing family members, and having been forced to have sexual intercourse are all risk factors for developing mental health problems.

4. Discussion

This study intends to build a predictive model to investigate the factors most related to the bullying behaviors among adolescents. Through preliminary analysis, we discovered that gender, race, family members' attitude, and the adolescent's difficulty in concentration are all risk factors for the adolescents' mental health. A logistic regression model was built, and the AUROC score is 0.74, indicating that the model has achieved relatively good performance in making accurate predictions on whether a child will be bullied at school. The predictive model suggests that Q105 (family member's attitude toward the adolescent) and Q115 (difficulty in concentration) are top risk factors. 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 less likely to interact with other students and more likely to be bullied. In addition, we also found that female white adolescents are more prone to bullying at schools. This predictive model is helpful for healthcare professionals to identify children that are at higher risk to be bullied and to develop 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 effective 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, Laszlo. 2008). This technique requires more effort but can generally achieve better performance.

References

National Center for Educational Statistics. (2019). Student reports of bullying: Results from the 2017 School Crime Supplement to the National Victimization Survey. US Department of Education. Retrieved from URL: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2015056

National Center for Educational Statistics. (2019). Student reports of bullying: Results from the 2017 School Crime Supplement to the National Victimization Survey. US Department of Education. Retrieved from URL: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2015056

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 02.09.2024;
accepted for publication 16.09.2024;
published 28.10.2024

Contact: xxjnicole@hotmail.com