American Research Journal of Humanities and Social Sciences

Volume 10, Issue 1, 130-137 Pages
Research Article | Open Access
ISSN (Online)- 2379-1047
DOI : 10.21694/2378-7031.24020

AMERICAN RESEARCH JOURNALS
An Academic Publishing House

Unmasking Risk Factors of Bullying Behaviors among Adolescents in Schools Using Logistic Regression

Yilin Zhang

Holderness School, Holderness, NH.

ABSTRACT

The bulling behaviors among adolescents has become a serious issue in the United States. According to the U.S. Department of Education's National Center for Education Statistics (NCES), 22% of students from 12-18 years old are the victims of bullying and about 28% of middle schools, 15% of high schools, and 10% of elementary schools reported bullying at least once a week.Victims of bullying suffer from issues like anxiety, depression, and poor academic performance, while bullies are more prone to future violence or substance abuse. Thus it is important to find the factors causing school bullying and prevent it from happening.

In this research, we investigated possible risk factors for bullying behaviors at school among adolescents. 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 identify possible factors associated with school bullying. The logistic regression model has achieved an AUROC score of 0.76, with 74.5% true positive rate (TPR) and 11.2% false positive rate (FPR). The findings reveal that students who experience electronic bullying and racial discrimination are at a heightened risk of physical bullying, while those carrying weapons or having strong social connections in school are less likely to be bullied. From the findings, the research proposes some recommendations for teachers and parents to identify and support the victims of school bullying, preventing the happening of school bullying.

INTRODUCTION

Bullying is a harmful and aggressive action taken by some students or one student towards another person at school. Bullying is repeated and various, including physical bullying and cyberbullying, and it causes many different problems to happen to the receiver of bullying. From the information of NCES, during the 2019-2020 school year, 22% of students from 12-18 years old are the victims of bullying.[1] In a Pew Research Survey in 2022, 46% of students from 13-17 years old experienced cyberbullying.[2] According to the NCES 2021-22 School Survey on Crime and Safety, about 28% of middle schools, 15% of high schools, and 10% of elementary schools reported bullying at least once a week.[3] From the data, it can be known that bullying is common in schools, thus it is also important to find out what are the effects brought by bullying, and what people can do to decrease the rate of bullying.

Bullying will bring negative impacts to both the victims and the students who bully. These impacts include harmful physical,

1 "Fast Facts: Public School Teacher Salaries." National Center for Education Statistics, U.S. Department of Education, n.d., https://nces.ed.gov/fastfacts/display.asp?id=719.Accessed 19 Oct. 2024.
2 Anderson, Monica, and Maureen D. Salamon. "Teens and Cyberbullying 2022." Pew Research Center, 15 Dec. 2022, https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/.Accessed 19 Oct. 2024.
3 "National Center for Education Statistics Releases New Data on Education." National Center for Education Statistics, U.S. Department of Education, 17 Jan. 2024, https://nces.ed.gov/whatsnew/press_releases/1_17_2024.asp.Accessed 19 Oct. 2024.

mental, and academic outcomes. For students who bully others, they are more likely to conduct domestic violence or substance abuse.[4] For victims, they need to experience physical harm and many physical injuries under bullying. For example, they are more likely to have headaches, sleep disturbances, and somatization.[5] Mentally, they will have problems like anxiety, timidity, and depression.[6] These mental health issues let targeted students disengage from school and society, causing self-harm in some cases and some negative academic impacts like skipping classes and low GPA.[7] Moreover, in the long-run version, bullying causes negative social impacts. Students who bully are more likely to have criminal convictions. Students who are bullied may retaliate toward schools or society. Overall, bullying impacts not only the students who are involved in the bullying events, but it also impacts social well-being.

The research about bullying is varied. Some focused on the solutions to bullying, like *Best Practices to address (or reduce) Bullying in Schools* published on The Phi Delta Kappan and written by Nadia S.Ansary, Maurice J.Elias, Michael B.Greene, and Stuart Green. This paper introduces the important factors for bullying solutions to be effective in schools. Also, it gives some strategies that schools can use for solving bullying problems like building a HIBAT group. This type of research doesn't analyze the characteristics of people involved in bullying, which can make more suitable solutions. The research from ILLINOIS CRIMINAL JUSTICE INFORMATION AUTHORITY (ICJIA) uses this way to help stop bullying. It presents how common bullying is, and it analyzes the influence of sex, sexual orientation/gender identity, and race/ethnicity on the frequency of bullying.

METHOD

Data

We used 2021 Adolescent Behaviors and Experiences Survey (ABES) data for this study. In 2021, during the COVID-19 pandemic period, the Center for Disease and Prevention(CDC) produced the Adolescent Behaviors and Experiences Survey(ABES). This shows nationally representative data by giving online questionnaires on health-related experiences and behaviors among high school students who attend school remotely because of the pandemic. The experiences and behaviors focused on include[8]:

- COVID-19 related experiences
- Emotional well-being
- Experiences related to perceived racism
- Behaviors that contribute to violence and unintentional injuries
- Sexual behaviors that contribute to unintended pregnancy and sexually transmitted infections, including HIV infection
- Alcohol and other drug use
- Tobacco use
- Unhealthy dietary behaviors
- Inadequate physical activity

The purpose of the ABES was to assess the health impact of COVID-19, find the prevalence of health risk behaviors, and examine the co-occurrence of health risk behaviors. The target population of this survey is all students through grades 9 to 12, regardless of race, gender, religion, or school type. In 50 States, all regular public, Catholic, and private high schools were included in the sample frame. Three hundred and thirty-nine schools were sampled in total. The sampling frame included all classes in a required subject, and systematic equal probability sampling was used for picking the classes from every school to complete the survey. From ABES, we picked the following factors that may influence the bullying rate.

4 American Psychological Association. "Bullying: A Primer for Schools." APA, 2021, www.apa.org/ed/schools/primer/bullying.
5 National Academies of Sciences, Engineering, and Medicine. *Preventing Bullying Through Science, Policy, and Practice*. National Academies Press, 2016, www.ncbi.nlm.nih.gov/books/NBK390414/.
6 American Psychological Association. "Bullying: A Primer for Schools." *APA*, 2021, www.apa.org/ed/schools/primer/bullying.
7 "Effects of Bullying." StopBullying.gov, U.S. Department of Health & Human Services, www.stopbullying.gov/bullying/effects. Accessed 21 Oct. 2023.
8 Copied from the ABES survey description

Variable	Description	Response Type	Reasoning
Q5	What's your race	A. American Indian or Alaska Native B. Asian C. Black or African American D. Native Hawaiian or Other Pacific Islander E. White	This shows the influence of racism
Q6	How tall are you without shoes	Height in feet, inches	This shows the possible height discrimination
Q7	What's your weight without shoes	Weight in pounds	This shows the possible weight discrimination
Q13	What's the frequency of carrying a gun	higher frequency, higher value in the case	Bullying people may be scared by the gun
Q24	Whether have been electronically bullied	Yes or No	Some possible positive relevance between the electronic bully and physical bully
Q65	What's your sexual orientation	A. Heterosexual B. Gay or lesbian C. Bisexual D. Some other identity E. Questioning identity F. Do not understand	Discrimination towards the sex-orientation
Q87	Where is the usual sleep place	A. Home of parent or guardian B. Home of other people C. Shelter or emergency housing D. Motel or hotel E. Public place F. No usual place G. Somewhere else	This can show the family background and financial situation
Q88	What's the frequency of carrying a weapon on school property	higher frequency, higher value in the case	Bullying people may be scared by the weapon
Q89	What's the frequency of quitting school days due to not feeling safe	higher frequency, higher value in the case	Quitting school possibly due to threats or bullying
Q90	What's the frequency of being threatened or injured by someone with weapons on school property	higher frequency, higher value in the case	Threaten causes bullying
Q93	What's the frequency of feeling being treated unfairly due to race or ethnicity	higher frequency, higher value in the case	This shows the influence of racism
Q99	What's the frequency of exercising to strengthen or tone muscles	higher frequency, higher value in the case	Bullying people may be scared by the strong body shape or strong muscles.
Q111	What's the frequency of using electronic devices for communicating	higher frequency, higher value in the case	Possibility of being excluded if some students cannot use electronic devices to communicate with other classmates
Q113	How close are you with people at your school	higher frequency, higher value in the case	People who are excluded or not popular are more likely to be bullied.
Q92	If you have been school bullied in the last 12 months.	higher frequency, higher value in the case	Dependent variable

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.

Statistical Method

Pre-Processing

Data pre-processing, or "data cleaning" and "data wrangling." It transforms row data into a format that is suitable for analysis by machine learning algorithms. The purpose is to improve the quality of the data and make it better for the analysis to use. 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:

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

In addition, this technique can also improve the convergence in some algorithms, like gradient descent.

For evaluating the model's performance, it is important to test it on the data that the model has never seen before, which helps in understanding how well the model generalizes to the unseen data. For this reason, 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 distribution of the positive class and negative class is highly unbalanced in the training set, we further applied the over-sampling technique to rebalance the data. Oversampling is a method used for class imbalance in datasets. When the minority class has significantly fewer instances than another class, oversampling can balance the class distribution by increasing the minority class's instances amount and letting the biased models perform better. The methods of oversampling include random oversampling, synthetic minority oversampling technique(SMOTE), and adaptive synthetic sampling(ADASYN).

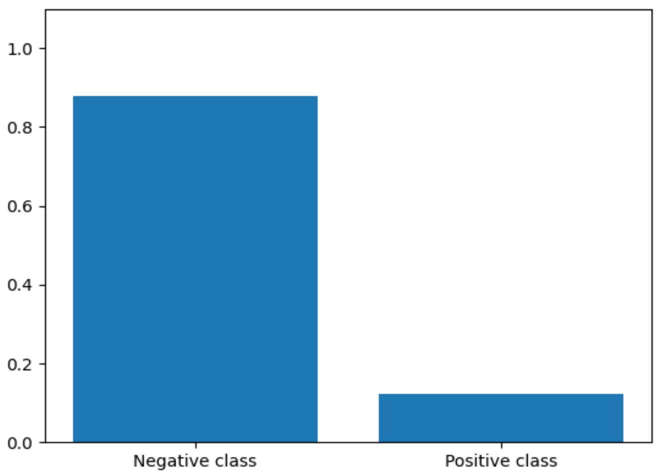

Figure 1. Distribution Graph Sample

Logistic Regression

Logistic regression is a statistical method that models the probability of a binary outcome based on one or more predictor variables. Instead of being utilized primarily for regression problems, logistic regression is employed for classification tasks. It is a fundamental algorithm in statistics and machine learning that is well-known for being simple and straightforward. The dependent variable in a logistic regression is a categorical, binary variable with two possible outcomes. These results are often categorized as 0, or 1 for two classes.

In logistic regression, the relationship between the input and the output is modeled using a linear combination of the input features. It is passed through the sigmoid function to produce a probability.

Sigmoid Formula:

$$\sigma(z) = \frac{1}{(1+e^{-z})} \quad \text{where z = w . x + b}$$

The logistic regression model can be expressed with the formula:

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

In the logistic regression, $hw(x^i)$ 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.

In logistic regression, regularization is a technique used to prevent over fitting in machine learning models. L1 regularization adds a penalty term that is proportional to the absolute value of the coefficients of the model. This lets some coefficients of the model be zero, making a more sparse model. L2 regularization adds a penalty term that is proportional to the square of the coefficients of the model. This encourages smaller coefficients instead of many large ones, which makes the model simpler.

L1 formula: $Cost(\beta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)\right] + \lambda\sum_{j=1}^{n}|\beta_j|$

L2 formula: $Cost(\beta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)\right] + \frac{\lambda}{2}\sum_{j=1}^{n}\beta_j^2$

The cost function, or loss function, Cost(β) measures how well a model fits the data by comparing predicted outputs to actual outputs, with β representing the model's parameters (weights). The dataset has m training examples, and for each example, y_i is the actual class (1 for positive, 0 for negative), $\hat{y}_i$ while is the predicted probability for the positive class, generated by the sigmoid function in logistic regression. The cost function includes two main terms: $y_i log(\hat{y}_i)$, which penalizes wrong predictions for the positive class, and $(1-yi)$ $log(1-\hat{y}_i)$, which handles penalties for wrong predictions in the negative class. Regularization is applied using λ to penalize model complexity, preventing overfitting by adding a penalty to the cost.

We applied elastic-net regularization, which is a combination of L1 and L2 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.

Cross-validation (CV) is a technique used to evaluate the performance of a model. It works to estimate how well a model trained on a subset of the data will perform on unseen data. The subset of the data that it is trained on is called the training set, and another one is called the validation or test set. CV separates the data into several subsets, known as folds. This process repeats multiple times, and every fold serves only once as the validation set. The results are averaged across the different folds, acquiring an overall performance estimated for the model. Hyperparameters are parameters that are set before the training process begins and are not learned during training. Tuning the different hyperparameters is important for improving the performance of a model. 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.

Model Validation

A confusion matrix is a tabular representation for evaluating the classification model's performance. It includes the model's predictions and compares them with the actual labels. The confusion matrix for a binary classification problem is a 2x2 matrix consisting of four values: True Positive(TR), True Negative(TN), False Positive(FP), and False Negative(FN). 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), which is the rate of actual positives that are correctly identified by the model, can be calculated as follows:

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

And the false positive rate(FPR), the rate of actual negatives that are incorrectly identified as positives by the model, can be calculated as:

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

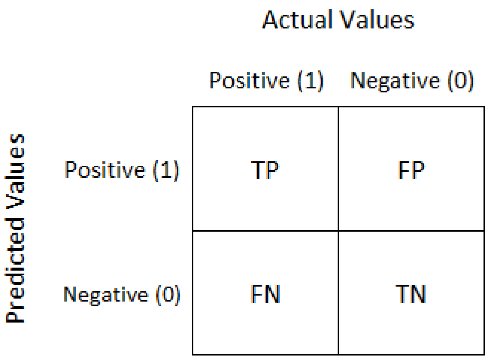

Figure 2. Confusion Matrix Sample[9]

A receiver operating characteristic curve, or ROC curve, is the graphical plot that illustrates the performance of a binary classification model across different thresholds. It is created by plotting the True Positive Rate (TPR), also known as sensitivity or recall, against the False Positive Rate (FPR), also known as the false alarm rate or 1-specificity, at various threshold settings. 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. Sometimes, it might be hard to identify which algorithm performs better by directly looking at ROC curves. Area Under Curve (AUROC) overcomes this drawback by finding the area under the ROC curve, making it easier to find the optimal model. A sample ROC plot is shown in Figure 3 below.

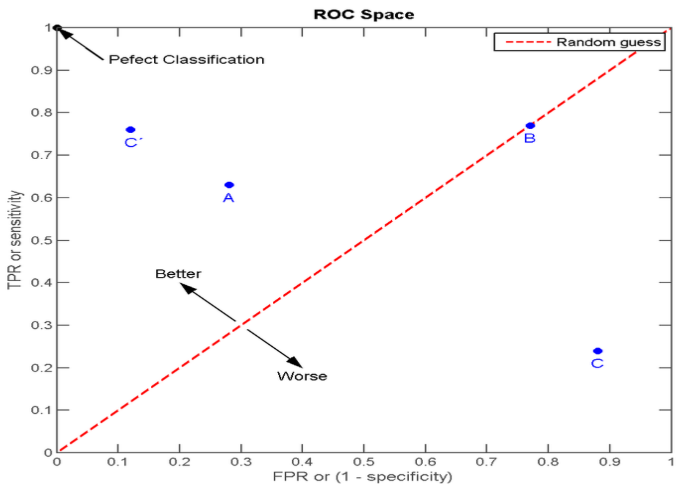

Figure 3. ROC Curve Sample[10]

RESULTS

Exploratory Analaysis

Correlation is a statistical measure that depicts the degree of relationship between two variables. It is different from causation because correlation cannot show that one variable causes another variable, but shows that there is a consistent pattern in their movements. A high positive correlation shows a strong direct relationship between two variables. It means that the growth of one variable goes along with the growth of another variable. A high negative correlation indicates a strong inverse relationship between two variables. One variable increases, and another variable tends to decrease. The near-zero correlation indicates no consistent correlation between the two variables.

From the Correlation Matrix show in Figure 4, Q24's correlation with Q92 is -0.41, which means whether the students have been electronically bullied has a strong negative correlation with the state of students being bullied on school property. The decrease in electronic bullying goes with the increase in physical bullying on school property.

9 Hussnain, Ali. "Understanding Confusion Matrix." *Towards Data Science*, 22 Nov. 2020, towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

10 Kaur, Simran. "Understanding AUC-ROC Curve." *Towards Data Science*, 23 Mar. 2021, towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Q88 has a 0.24 correlation with Q89, showing that as the frequency of carrying a gun increases, the frequency of quitting school days due to not feeling safe increase. Q89 and Q90 has a correlation of 0.17, meaning that the frequency of quitting school days due to not feeling safe also has the same positive correlation with the frequency of being threatened by someone with weapons in school.

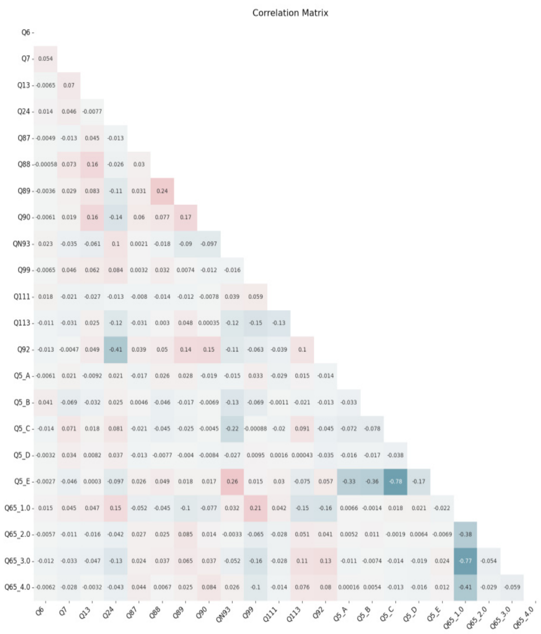

Figure 4. Correlation Graph

Confusion Matrix and ROC Curve

Figure 2 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 (~74.5%) true positive rate and a relatively low (~11.2%) false positive rate.

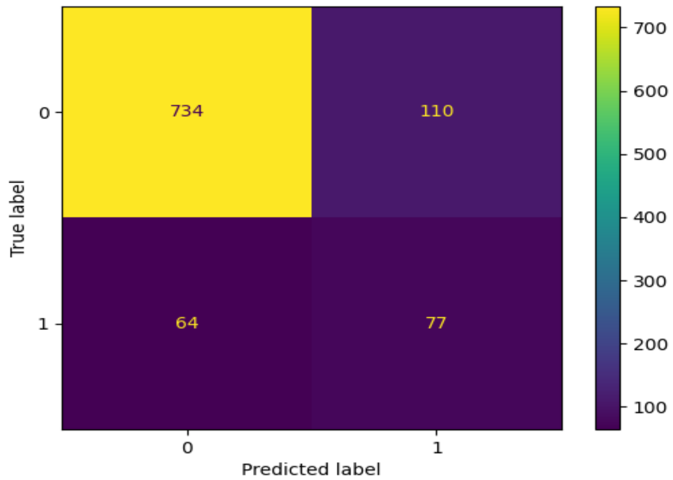

Figure 5. Confusion Matrix Graph

From the Receiver operating characteristic plot in Figure 6, the AUROC is 0.76, showing the model performs well and has a good ability to distinguish between positive and negative classes.

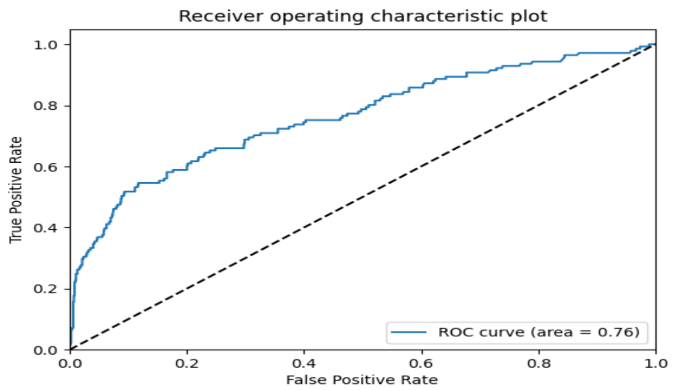

Figure 6. ROC Curve

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. From the figure 7 showing the coefficient of logistic regression, a higher positive value indicates a higher possibility of not being bullied. On the other hand, a higher negative value shows a higher possibility of being bullied. Therefore, the feature's importance can be known from observing this graph. Q24 has the highest negative coefficient compared with other variables, and QN93 has the second-highest negative coefficient. This means that being electronically bullied will cause physical bullying. Also, experiencing racial discrimination more often leads to an increased likelihood of being physically bullied. Among all the races, it can be told that Q5_D, which is Native Hawaiian or Other Pacific Islander, faces more school bullying than other races. Moreover, there are some factors that make the students less likely to be bullied, and they have a higher positive logistic regression coefficient. Among all the sexual orientations, Heterosexuals faces more school bullying than others, showing that minor sexual orientations will not cause more bullying. The graph reveals that Q13 and Q113 have a more significant influence on Q92, stating that if the students are carrying guns to school or being closer to people at their school, they are more protected in front of the school bullies.

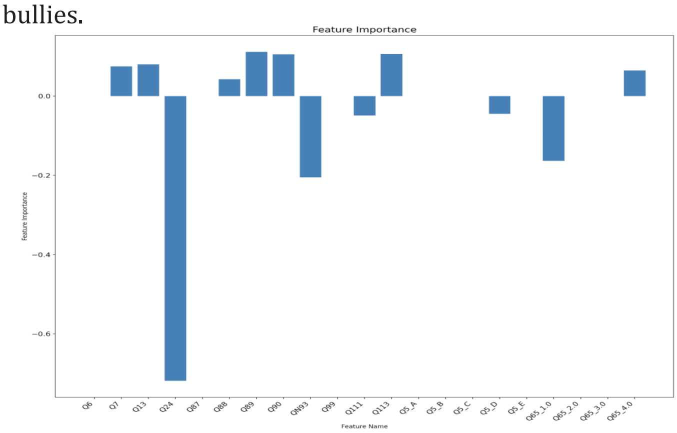

Figure 7. Feature Importance

CONCLUSION

In this research, the school bullying phenomenon is being studied. From the data, it can be known that bullying is a harmful action that happens in school which has a negative impact on the whole society. Both the bullying students and the bullied students are victims of school bullying because of the bad effects bullying has physically and mentally. Therefore, this research is focusing on finding the factors behind school bullying. As the teachers and parents get a better understanding of these factors, they can predict the possible coming of school bullying, and then stop it on time. During the research process, the Adolescent Behaviors and Experiences Survey(ABES) is used for applying data to train the model, and machine learning, especially logistic regression, is used to predict the correlations or potential causations between different variables with the bullying rate. From the model, we predict that students who are being electronically bullied and who are experiencing racial discrimination are more likely to be physically bullied. On the other hand, students who are carrying a gun to school and students who are close to other people in the school are less likely to be bullied.

Connecting with the research of Marci Feldman Hertz, Ingrid Donato, and James Wright titled *Bullying and Suicide: A Public Health Approach*[11], we can evaluate the importance of investigating the factors behind bullying. *Bullying and Suicide* claims three points: bullying is common, prevalent, and has a detrimental effect, bullying has a strong association with suicide-related behaviors, and public health strategies can prevent the happening of bullying and suicide-related actions. This research mostly focuses on the results of bullying, collecting and analyzing the data on how many students have mental problems and how many students have what kinds of suicide-related actions after bullying, but the factors of school bullying are ignored. Our study complements this gap, demonstrating that some variables like race cause school bullying more often, which helps the prohibition or prevention become more specific and useful.

In real life, the research result of this paper gives teachers and parents a direction of what kind of students they should treat more carefully in the context of stopping school bullying. For example, for students who do not belong to the major race and students who have been electronically bullied before, the teachers need to be sensitive if they have some abnormal actions, that might be caused by the bullying. Also, the teachers need to teach these students who are more susceptible to school bullying how to react if they become victims in the future. For instance, some students are being threatened to not tell teachers and parents after the first bullying, thus they do not dare to tell teachers about their bullying experience. If the teachers and parents had taught them how to react before, they would not have followed the threats but would have reported the bullying events.

One limitation of this research is the selection bias it carried. Since we removed all missing values from the data, the selection bias appeared. For example, for some students who are in the database but don't have the values or responses for the factor we evaluated, we just exclude these students. One improvement we can make is using missing value imputation, which imputes values to get the responses and replace the missing value. Another better one is the k-nearest neighbor (k-NN) method, which finds the most similar response from all the data, and gives the missing value of that similar response, which reduces the bias most effectively. This technique requires more effort but can generally achieve better performance.

11 Hertz MF, Donato I, Wright J. Bullying and suicide: a public health approach. J Adolesc Health. 2013 Jul;53(1 Suppl):S1-3. doi: 10.1016/j.jadohealth.2013.05.002. PMID: 23790194; PMCID: PMC4721504.

REFERENCES

1. U.S. Department of Education, National Center for Education Statistics. *The Condition of Education 2024*. Sept. 2024,https://nces.ed.gov/pubs2024/2024109.pdf.
2. U.S. Department of Health and Human Services. "What Is Bullying?" *StopBullying.gov*,https://www.stopbullying.gov/bullying/what-is-bullying. Accessed 21 Oct. 2024.
3. "Fast Facts: Public School Teacher Salaries." *National Center for Education Statistics, U.S. Department of Education, n.d.*, https://nces.ed.gov/fastfacts/display.asp?id=719 Accessed 19 Oct. 2024.
4. Anderson, Monica, and Maureen D. Salamon. "Teens and Cyberbullying 2022." *Pew Research Center*, 15 Dec. 2022, https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/ Accessed 19 Oct. 2024.
5. "National Center for Education Statistics Releases New Data on Education." *National Center for Education Statistics*, U.S. Department of Education, 17 Jan. 2024, https://nces.ed.gov/whatsnew/press_releases/1_17_2024.asp Accessed 19 Oct. 2024.
6. American Psychological Association. "Bullying: A Primer for Schools." *APA*, 2021, www.apa.org/ed/schools/primer/bullying.
7. National Academies of Sciences, Engineering, and Medicine. *Preventing Bullying Through Science, Policy, and Practice*. *National Academies Press*, 2016, www.ncbi.nlm.nih.gov/books/NBK390414/
8. American Psychological Association. "Bullying: A Primer for Schools." *APA*, 2021, www.apa.org/ed/schools/primer/bullying
9. "Effects of Bullying." *StopBullying.gov*, U.S. Department of Health & Human Services, www.stopbullying.gov/bullying/effects Accessed 21 Oct. 2023

10. Hui, Jonathan. “Normal Distributions in Machine Learning.” *Medium*, 25 Feb. 2020, jonathan-hui.medium.com/normal-distributions-in-machine-learning-c8a21c8ba8c9

11. Hussnain, Ali. “Understanding Confusion Matrix.” *TowardsDataScience*,22Nov.2020,towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

12. Kaur, Simran. “Understanding AUC-ROC Curve.” *Towards Data Science*, 23 Mar. 2021, towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Citation: Yilin Zhang, “Unmasking Risk Factors of Bullying Behaviors among Adolescents in Schools Using Logistic Regression”, American Research Journal of Humanities and Social Sciences, Vol 10, no. 1, 2024, pp. 130-137.