MET CS699 Project Assignments - Predicting Bullying Victims

By Nakul Shahdadpuri and Lekh Shetty

Boston University

10/5/2023

I. Introduction

Bullying is a pervasive and concerning issue within educational settings, impacting the lives of countless students and affecting their mental, emotional, and physical well-being. The detrimental consequences of bullying, ranging from increased stress and anxiety to academic decline and, in severe cases, self-harm or suicide, emphasize the critical need for effective prevention and intervention strategies.

Bullying is a widespread issue affecting educational institutions globally. According to the National Center for Education Statistics (NCES), in the United States alone, in the 2019 school year, approximately 22% of students aged 12-18 reported being bullied at school[1]. These distressing statistics underscore the urgency of addressing this problem effectively.

Understanding the prevalence and gravity of bullying, we recognize the need for data-driven approaches to combat this issue. By leveraging data analysis and classification modeling, we aim to not only predict bullying instances but also identify potential risk factors and patterns associated with bullying. Such insights are invaluable for educators, policymakers, and stakeholders to tailor intervention strategies, allocate resources efficiently, and foster safer and more supportive educational environments.

In an effort to contribute to the efforts aimed at addressing bullying, this project sets out to develop and evaluate multiple classification models using a dataset sourced from the 2013 school crime supplement of the National Crime Victimization Survey (NCVS). The dataset encompasses 4947 tuples, each representing a student, with 204 attributes. The crucial class attribute, denoted as 'o_bullied,' categorizes students based on whether they have experienced bullying (1) or not (0). By leveraging this dataset, our objective is to build robust classification models capable of predicting instances of bullying, enabling proactive measures to be taken to prevent or mitigate the adverse effects of bullying on students.

By undertaking this project and employing advanced analytical techniques, we seek to contribute to a growing body of research and initiatives aimed at combating bullying. Our efforts aim to provide data-driven insights that can inform the development of proactive anti-bullying strategies and contribute to creating a safer and more inclusive learning environment for students. Through the application of diverse classification algorithms and rigorous performance evaluation, we endeavor to identify the most effective predictive model that can significantly contribute to the ongoing fight against bullying in schools.

In this report, we have discussed our data preprocessing approach and data exploration findings. Section 1 entails the introduction of the report, Section 2 includes our data preprocessing approach and details about the tools used, Section 3 includes our findings after intensive data exploration and finally section 4 includes our references.

II. Data Preprocessing

For the dataset provided, we needed to undergo preprocessing in order to exclude redundant variables and better understand the information provided. Data preprocessing is a crucial step in the development of machine learning applications. Preprocessing involved a series of operations aimed at preparing the dataset for effective classification model predictions.

The number of columns in the given dataset were about 204. WIth such a large number of independent variables, classification algorithms tend to overfit and the data becomes too sparse for any distance based learning methods. Thus reduction of the number of independent variables becomes of paramount importance. In this project, we have reduced the number of columns from 204 to about 87 columns.

Methodology	**Columns Eliminated**
Removed duplicates	5 removed
Near Zero Variance Columns	74 removed
Remove Highly Correlated Columns	58 removed
Columns removed during data exploration	4 removed

First, we looked to remove any duplicate columns found in the dataset, ensuring that each column in the dataset provided is unique. From this, we saw that 5 variables were removed from the original dataset.

Next we tried to find columns which had little to no variation in their values. We identify them using the NearZeroVar function in R. After removing all the columns with zero var, we performed correlation analysis to remove columns with inter correlation more than 0.85.

Finally we manually perform data exploration to remove values and find variables of noted importance. Upon review of the remaining columns, we noticed there were 5 columns with a disproportionate number of nulls in these columns, meaning these columns would not contribute much to the predictive modeling. These columns were found to have a majority of null/Out of Universe values, and hence not being considered to be kept in the final preprocessed data sets.These variables were:

Column	Description
V2025	Household Information - Direct Outside Access
V2120	Public Housing
V2121	Verification of Public Housing
VS0021	School Crime Supplement Variables - Is your school associated with a religion?

Moreover, from the remaining columns, we noticed that there were variables present that were worth making a note of to be considered as vital factors that should be used for the predictive models. These columns were well distributed with significant values of information from the survey and, in theory, sound like factors that could heavily determine the outcome of someone being bullied or not. Some of these variables of note include:

Column	Description
VS0070	During the school year, did anyone offer, or try to sell or give you an illegal drug other than alcohol or tobacco at your school?
VS0046	In your classes, how often are you distracted from doing your schoolwork because other students are misbehaving?
VS0112	Have you seen any hate-related words or symbols written in school?
VS0148	Would you agree there is an adult at school who… C. listens to you when you have something to say
….	…

III. References

[1] The NCES Fast Facts Tool provides quick answers to many education questions (National Center for Education Statistics). National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. (n.d.). https://nces.ed.gov/fastfacts/display.asp?id=719

Introduction

Abstract

Literature review

Data preprocessing

Data Exploration