School of Computer Science

Data Science Group Project - Final Report - May 4, 2023

Nidhi Priyadarshini, Devi Chandran, Safia Elmi, Shamia Ali

How does bullying impact the overall development of children in schools in the UK?

Supervisor: Hayatullahi Bolaji Adeyemo. With thanks to Dr.Qamar Natsheh.

Abstract:

This study examines the impact of bullying on children's overall development in schools in England. This study is based on a detailed scientific study of the effects of bullying on children's academic, social, and emotional development. The major area of study is the categorization of bullying as On-campus, Off-campus, and cyber-bullying. The findings show that bullying has a negative impact on children's academic, social, and emotional well-being. This study also highlights the importance of early detection and prevention of bullying to minimize the impact on children's development with the help of using the appropriate machine learning algorithm. The findings of this study have important implications for teachers, parents, and policymakers in developing strategies to prevent and address bullying in schools and promote positive development for all children.

Keywords:

Bullying; Prediction Model; Classification; Machine Learning.

Table of Contents

1. Introduction

Bullying is a reprehensible behavior that entails a person's physical or mental abuse, leading to severe outcomes for its victims. This issue becomes even more delicate when young minds are susceptible to such malicious acts, particularly in school settings. In a survey performed by the Department for Education from 2017 to 2018, 17% of young students between the ages of 10 and 15 claimed to have experienced harassment in the 12 months prior [10]. Schools are intended to be safe and nurturing environments where young minds can thrive and flourish and incidents such as these can have serious effects on a child's life.

Apart from occurring in various forms, bullying can also take place in diverse settings. It can be physical, verbal, or psychological, and can take place anywhere from school hallways to the online realm. In this digital age, cyberbullying has emerged as the most dominant and dangerous form of bullying. According to the National Society for the Prevention of Cruelty to Children (NSPCC), online bullying is on the rise. As per the NSPCC's study for the 2015–2016 academic year, the number of people seeking counselling for online bullying has increased by 88% in the previous five years [11]. These statistics are a stark reminder that cyberbullying has become an increasingly common problem in recent years and requires greater attention and action for prevention.

Furthermore, it is essential to acknowledge that cyberbullying is not only hurtful but can have serious consequences on the mental and emotional well-being of young people. The rise of social media platforms and technology has created a breeding ground for online harassment, making it easier for bullies to target their victims. Therefore, there is an urgent need to take proactive measures to combat cyberbullying and create a safer online environment for our youth.

A study by King's College London revealed that bullying has long-term effects on children, such as a greater likelihood of experiencing mental health issues, problems in relationships, and an overall hindrance in their development [12]. Motivated by these alarming findings, we were intimidated to see which backgrounds contribute most significantly to the issue of bullying.

Our study is centered on exploring the factors that contribute to bullying and the ways in which it can impact a child's overall development. To this end, we analyzed several factors, including economic background, gender, age, body weight or obesity, social life, and parental communication. Our investigation into the matter also throws light onto the effectiveness of anti-bullying organizations and laws.

In conclusion, our research emphasizes the importance of providing a conducive and secure environment in schools and a safe online atmosphere for young individuals. By addressing the problem of bullying head-on, we can create a more inclusive and safer learning environment for all students. It is our hope that our findings will inspire action toward a brighter future, where every child can develop and grow without the threat of bullying.

2. Background Research

Bullying is defined as repeated aggressive behavior intended to harm or intimidate others and can take many forms, including physical, verbal, and psychological bullying. The impact of bullying on children's development has been the subject of many studies in recent years, with many researchers highlighting the negative effects bullying can have on children's academic, social, and emotional well-being. [1]

With our analysis, we found that school performance is one area that is negatively affected by bullying. Children who are bullied are more likely to have lower academic achievement, higher rates of absenteeism, and lower levels of engagement in school. This is because bullying can create a hostile learning environment that interferes with children's ability to learn and focus. In addition, bullying can lead to stress and anxiety, which can impair cognitive function and academic performance. In the research article [3], the results of the study revealed that bullying is a common problem in UK schools. More than 80% of students reported being bullied at least once. The most common forms of bullying were swearing, teasing, and spreading rumors. Physical bullying, such as hitting or pushing, was less common but still a significant problem. There have been a large number of reported cases of cyberbullying leading to suicidal tendencies in children.

Based on these studies conducted through research, we found that bullying also has a significant impact on children's social relationships. Children who are bullied may have difficulty making friends, forming healthy relationships, and communicating effectively with others. They may also experience more social isolation and loneliness, which can negatively affect their mental health.

The emotional impact of bullying is perhaps the most significant. Children who are bullied are more likely to experience anxiety, depression, and low self-esteem. They may also be at greater risk of developing psychological problems later in life, such as anxiety disorders and depression. The emotional wreck of bullying can be excruciating if it lasts longer or if the child does not receive adequate support from parents or carers.

Overall, bullying has a consequential adverse impact on the overall development of children in UK schools. It can affect their academic performance, social relationships, and emotional well-being and have long-term effects that persist into adulthood. Therefore, it is essential to develop effective strategies to prevent and address bullying in schools to promote the positive development of all children. These research and facts led us to study furthermore about this subject area and build our algorithm to a) detect what are the most affected class of children w.r.t age, gender, and background b) Classification of bullying which is common among the schools in the UK.

3. Question Development

Our key goal in formulating the question was to formulate a precise investigation that could be addressed by data analysis. Due to this project being within the field of data science, we considered that our question was supported by the necessary datasets for the study. This section will illustrate our rationale for stating and refining our question and describe the steps in obtaining the final question.

3.1 Rationale for developing the question

3.1.1 Determining the Type of question to investigate

The first step in developing a question was to understand what our goal of this study is as this would fundamentally determine the interpretation of our results. This involved exploring types of questions through research and through thorough communication of goals. Different papers such as [7] gave us an insight into classifying types of questions based on our predicted outcome of the interpretation of our results. Spending time to know the different types of questions shaped our study giving us a streamlined focus on the result.

3.1.2 Understanding the Characteristics of a good question

By looking at types of questions, we inherently thought about what makes a good question. This would later become a guideline for us in shaping and redefining our question to ensure this criterion was satisfied. [9] details more on the characteristics of a good question, mentioning qualities such as Specificity, Relevance, and Interesting. Ensuring that our question satisfied these attributes came through communicating with experts and revisiting the relevance as well as reiterating the plausibility and specificity of our ideas.

The first step in ensuring that these characteristics were met involved selecting the issue to investigate. Through research and communication, many suggestions were made, such as looking into crime rates, heart disease, and other health and social issues. Based on these different ideas, narrowing down to a single topic involved alternating between searching for relevant data and formulating a comprehensive and precise question that could answer an interesting issue.

3.2 Translating the question to a data science problem

As stated previously one important objective was ensuring data analysis could be applied to the question and can use data to obtain interpretable results. Identifying which data science problem to investigate was a heavily impactful variable in finding the relevant datasets as well as deciding on the main issue. Through research for different datasets and relevant topics, our common interests became investigating bullying within schools, a prominent problem that we found many young people in the UK still encounter.

3.2.1 Type of data science problem

Once we identified the main issue, we focused on identifying the type of data science problem to investigate as this also shaped how we stated our question. Initially, we had the idea of working on a logistic regression problem, where our model is trained on a bullying dataset and later predicts whether a person will be bullied based on certain factors and features. However, this idea was halted as we faced certain limitations including a lack of availability in datasets as well as quality in datasets we found. Hence investigating a classification problem was proposed which allowed us to think about what factors might affect bullying, such as economic status, gender, age, and religion. Thus, we proposed an initial question putting importance on certain factors like gender and exploring how they affect bullying within schools in the UK.

3.2.2 Exploring relevant literature to the Problem

A crucial step to understanding our issue was by looking at the literature relevant to the topic. A research paper [8] gave another perspective on bullying, which put importance on the fact that bullying occurs in both females and males, however, there were different means of bullying between each gender. Taking this into consideration, rather than putting importance on the effect of gender on bullying, we considered the different methods of bullying and how that affects the development of children within schools in the UK. This gave room for us to explore cyberbullying as well as on-site bullying and consider the effects on children's development, i.e., psychological health, and social life.

3.2.3 Our question

This point of gender not being an important factor in bullying was later proven when we did a data exploration of our datasets and found out that gender did not play a role in bullying, as opposed to our initial thoughts of girls being affected by bullying more than boys. The details can be found in sections 6 and 8. Thus, refining our question based on these aspects gave us our question
"How does bullying impact the overall development of children in schools in the UK?"

4. Retrieving The Data

In our study of finding the socioeconomic factors in the cases of bullying, we collected a dataset to analyze the pattern and the behavior of bullies in schools and workplaces in the United Kingdom. For our analysis, we considered the features of gender, religion, economic background of the victims, etc. to check what is the factor that causes the most bullying cases. To support our project question and provide us with a clear answer to our research question, we initially looked for several datasets. We have struggled initially to set the research question especially because of the unavailability of the correct dataset. Whilst we knew what kind of question we wanted to research, the features were not available such as mental issues faced by the victims, different types of bullying, if the parents understood the problems, etc. Most of them also were not up to date and the years they covered were not recent which would have led to an incorrect representation of the current situation. We, therefore, had to spend a lot of time researching the dataset and finalizing the research question.

During the period of our dataset research, we made sure that the dataset should be large enough so that it demonstrates a better representation of the population and provides accurate results. It must have a few thousands of rows and more columns which our dataset consisted of such as fifty thousand rows and nineteen columns. We also checked whether it was possible to transform non-numeric features into numeric for example if the students were cyberbullied in the past 12 months which had yes or no answers that could be transformed into 0s and 1s. This way we ensured that our data were mapped into useful features so that it could be used to train the model.

A good dataset should have disaggregated data and our dataset meets this requirement as the data has been broken down into detailed sub-categories. If it was aggregated, then we would not have had much data for our analysis whereas disaggregated data offers many benefits which include giving an accurate analysis of the situation and providing enhanced details of the population's specific characteristics. This means that our dataset would allow us to visualize the issues caused by different features and characteristics of an individual.

While choosing the dataset, we considered the following guidelines to make sure our dataset was reliable and not biased:

- Not to use Kaggle-like websites to avoid plagiarism as the datasets that are published on those websites have been used by the publisher.
- Found a relevant dataset that answers the main research question and features such as the 11-19 age group and data from the years 2018-2022.
- As bullying is a delicate issue, we collected a reliable dataset so that the cases have been reported by the victims and the source is not fake.
- Checked whether the researchers have followed legal and ethical considerations and if the respondents have completed the questionnaires by themselves.

To find a suitable and accurate dataset, we used 'Google dataset search' to carry out our research project and we used websites such as ONS and NHS to collect a dataset that demonstrates a true representation of the bullies and the mental health of the victims.

The dataset we collected is in CSV file format and it is in the form of rows and columns which allowed us to structure our queries. The CSV file then were read into pandas DataFrame using Jupyter and pandas library was used to construct our data analysis.

5. Preparing the Data

To produce high-quality research, we decided to prepare our dataset and remove any null values or irrelevant features that have been considered before on the dataset. The main reason why it was necessary is that raw data usually could have many inconsistencies and data preparation allowed us to remove them before conducting our research. There are many issues that we have encountered before modeling the dataset which is:

5.1 Handling missing values:

Our dataset contained many missing values that needed to be removed before modeling and therefore, we used different strategies to deal with them. The methods that we used included replacing missing values with mean, median, and mode of the total column depending on the queries that we wanted to evaluate. The columns that we worked on were Gender, Physically_attacked, Physical_fighting, and Felt_lonely. These columns had missing values and we replaced them with the median value. On the other hand, there were missing values in the other columns which were less than 1% of the total data and therefore, we decided to delete them.

5.2 Dropping irrelevant columns:

There were a few columns that were irrelevant to our analysis, and we decided to drop those columns. For that, we first identified the columns that were not relevant to our analysis and then delete them to reduce the complexity of the dataset. This helped us improve the accuracy of the dataset and enables a faster data analysis. 'Record' was the column that has been removed as it did not give any related information to our analysis. Also, column 'Miss_school_0_permission' was removed for having the same values as the 'Missed_classes_or_school_without_permission' column. These were the only two columns that has been removed.

5.3 Data cleaning:

This is the process that we performed to remove any duplicates or incorrect data within the dataset. This process would allow us to get an algorithm that is more reliable and filter out any unwanted outliers. When we checked for duplicates in our dataset, we used the pandas library and then used it to remove any duplicates that were present.
d) Data exploration:
This part is one of the first steps for any data analysis and we have examined our dataset and visualized our data to gain an insight into the data we were working with. It helped us understand which area of the dataset we wanted to focus more on and get more information out of it. The main features of visualization that we have used were histograms and scatter plots. This allowed us to visualize our data and check which data could be more important and which might distort our data analysis.

6. Rationale for the group's approach to exploring the Data

The data we worked with was intrinsically categorical and presented in a non-numerical format. To conduct our analysis, we transformed it into a numerical format, using Boolean encoding as described in section 7. The data pertained to three primary categories of bullying: bullying inside school, bullying outside of school, and cyberbullying.

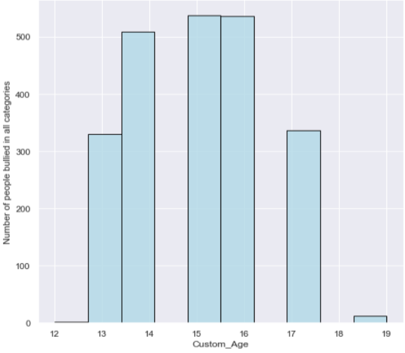
6.1 Analysis of bullying in the three categories:

6.1.1 Children who were commonly bullied in all spaces (Fig 6.1)

We combined the data into a single data frame comprising a record of persons who were victimized by bullying in each of the three categories. To present this information in a clear and accessible manner in the form of the age of the victims, we generated a histogram using the Seaborn and Matplotlib libraries. The histogram was an optimal choice due to its ability to directly display the count of each variable without the need for additional calculations.

```
bullied=df.loc[(df['Cyber_bullied_in_past_12_months']==1)&(df['Bullied_
on_school_property_in_past_12_months']==1)&(df['Bullied_0t_on_school_pr
operty_in_past_12_months']==1)]

sns.histplot(bullied['Custom_Age'],bins=10,color='lightblue',edgecolor=
'black')
plt.ylabel('Number of people bullied in all categories')
```



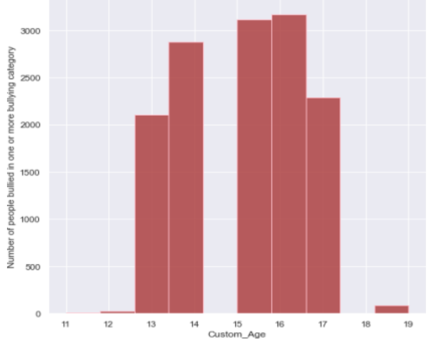


Fig. 6.1 Histogram showing the number of people bullied in all the three bullying categories(left) (Fig. 6.2) Histogram showing number of people bullied in at least one of the three categories(right)

6.1.2 Children who were bullied in at least one of the three categories (Fig. 6.2)

A detailed count of children who were bullied in either one of the three, two or all of the categories was seen, and they were examined to see the number of children in each age group who were bullied at least once in their life. The histogram (Fig. 6.2) reveals that children in the age of 13-17 were the major victims of bullying, with age 19 being the highest.

```
bulliedinany=df.loc[(df['Cyber_bullied_in_past_12_months']==1)|(df['Bul
lied_on_school_property_in_past_12_months']==1)|(df['Bullied_0t_on_scho
ol_property_in_past_12_months']==1)]

sns.histplot(bulliedinany['Custom_Age'],bins=10,color='brown',edgecolor
='white')
plt.ylabel('Number of people bullied in one or more bullying category')
```

6.2 Feature selection:

There were several features included among which we had to select the most relevant. To prepare the data for modeling and predictions, we performed certain feature selection techniques such as correlation matrix (Fig. 7.2) and pair plot of each feature with the different target variables (bullying inside school, outside, and cyberbullying) as described in Fig. 6.3 and 7.3. This step was necessary as to avoid the problem of overfitting (described in detail in section 7).

6.2.1 Correlation matrix:

The correlation matrix is one of the finest tools to find out the features that best explain the data. It can be used to derive the correlation coefficient which has a value that lies between -1 and 1, with 1 being a perfect positive correlation, -1 showing a perfect negative correlation and 0 indicating uncorrelated features. Ideally, the uncorrelated features can be removed for modeling to avoid the problem of overfitting.

6.2.2 Pairplot

Pairplots are graphical interpretations of the relationship between features (linear or non-linear) and help us choose the input variables which are the most closely related to our output variable of interest. They can visually represent the correlation among features which makes it easier for us to make decisions for our data.

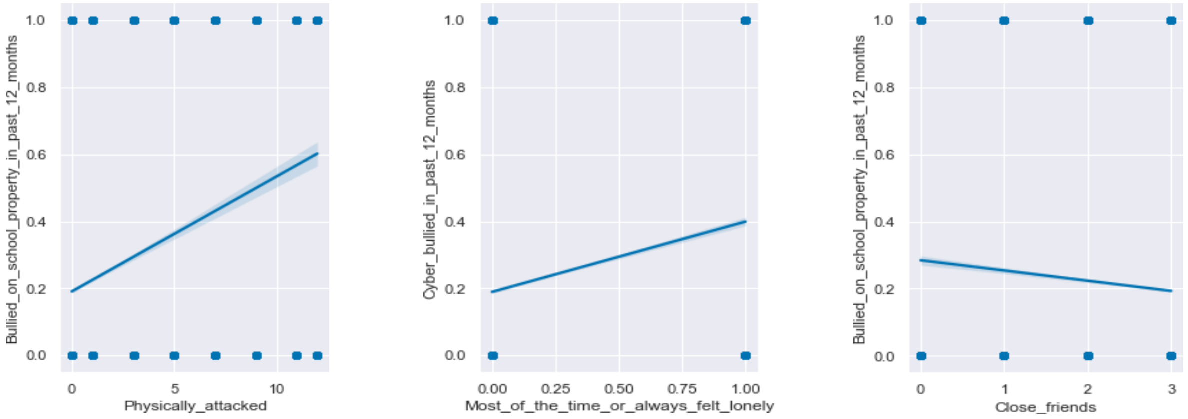


Fig. 6.3 Pairplots showing the relationship between different variables

6.3 Analysis of the mental health of Children:

As part of analyzing the data, we conducted exploratory data analyses. We were eager to find out the proportion of the data which contributed to the non-numerical attributes. There were two main attributes from the data which gave us deep insights into our problem at hand: a) whether parents understand the problems of students (Fig. 6.4), b) how often students felt lonely (Fig. 6.5). Both these attributes had five main categories: ‘Always’, ‘Sometimes’, ’Rarely’, ‘Most of the time’, ‘Never’.

6.3.1 Pie chart

Pie charts (Fig. 6.4, Fig 6.5, and Fig. 6.6(a)) help in emphasizing the differences in the proportion of the data in the five categories and can easily tell us the result.

Example code of pie chart for Fig. 6.5:

```
Always = df_copied ['Felt_lonely']=='Always'
Sometimes = df_copied ['Felt_lonely']=='Sometimes'
Rarely = df_copied ['Felt_lonely']=='Rarely'
Never = df_copied ['Felt_lonely']=='Never'
Most_times = df_copied ['Felt_lonely']=='Most of the time'


labels = ['Always', 'Sometimes', 'Rarely', 'Never', 'Most of the
time']
bulliedData = [(Always==True).sum(),(Sometimes==True).sum() ,
(Rarely==True).sum(),
         (Never==True).sum(), (Most_times==True).sum()]
plt.figure(figsize=(10,10))
plt.axis("equal")
plt.pie(bulliedData, labels=labels, autopct='%1.1f%%')
plt.legend(loc='upper right')
plt.title('Felt lonely',fontsize=20)
```

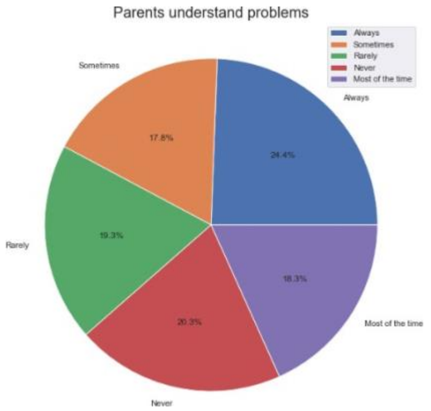


Fig. 6.4 Parents understand problems

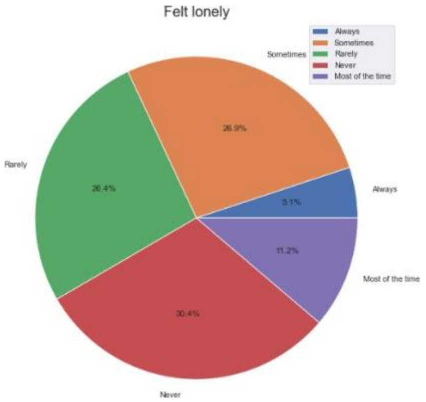


Fig. 6.5 Felt lonely

From Fig. 6.4 it is evident that the percentage of ‘Never’, ’Rarely’, and ‘Sometimes’ is 57.4%. This information reveals that parents not understanding the children’s issues could be a possible reason for them being bullied and suggests that students need to be given adequate support from their families for healthy development. Fig. 6.5, reveals that about 16.3 % of children in these schools felt lonely ‘always’ and ‘most of the time’. Consequently, it implies that about a quarter of children tend to feel lonely when they are faced with bullying and that bullying can affect their mental health in more ways than imaginable.

6.4 Analysis of cyberbullying faced by Children

A major part of our investigation aimed at cyberbullying and the potential reasons behind it. There were several conclusions drawn from the data many of which pointed to features such as age, economic background, and gender.

6.4.1 Pie chart

As mentioned in the pie chart below (Fig. 6.6(a)), among the other age groups, 13-17 werethe age groups most affected by cyberbullying.

6.4.2 Bar Chart

Another major conclusion that can be drawn from Fig. 6.6(b) is that, even among the age group of 13-17, the gender-wise distribution in each age group is unequal. This portrays that woman are slightly more likely to be cyberbullied and targeted than men, even though the difference is not very prominent. Thus, by far, anyone can be a victim of cyberbullying or bullying in general and gender does not draw a line of certainty.

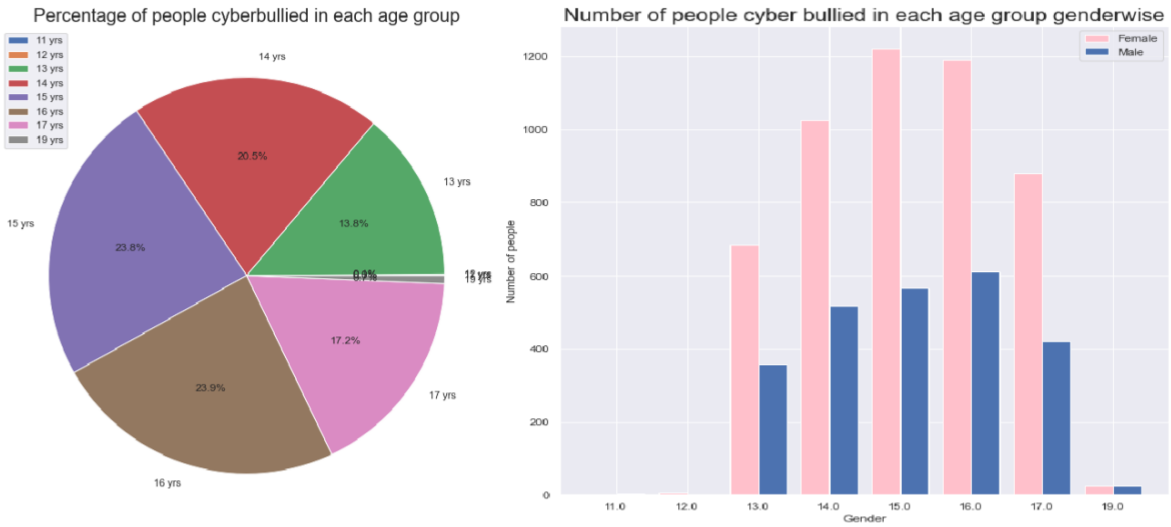


Fig. 6.6(a) Pie Chart showing Percentage of cyberbullied in each age group(left)
Fig. 6.6(b) Bar Chart showing Count of people cyberbullied in each age group gender wise(right)

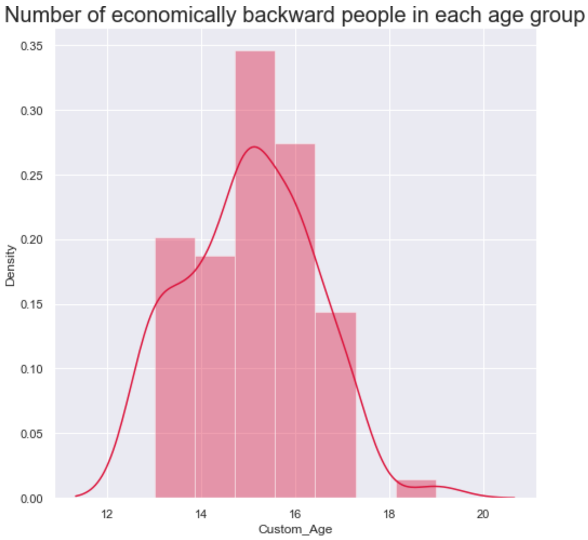


Fig. 6.7 Distribution plot showing the density of economically backward

6.4.3 Distribution Plot

The distribution plot (Fig. 6.7) was created to analyze the density distribution of the number of people cyberbullied in each age group. As seen from the figure, the plot showcases that among the age group of 11-19, the people in ages 13-17 and especially 15 were the most targeted.

7. Rationale for Data Modelling/Experimentation

As part of our project, the essential element after we figure out the right dataset and its reliability is to apply the right algorithm. The dataset was initially complex and required some cleanup. There was missing information, duplicates, and a little formatting was required. After we completed the collection of data, data preprocessing was performed. In this step as mentioned in the above section 6, the null values were replaced by the median of the individual features which had the null values. There could be several reasons for the missing information and instead of blindly dropping the missing values we thought of performing a small analysis by taking the sum of all the missing records available in each column, to understand if these columns correlate. We saw that the missing values are not of the type Missing Completely at Random (MCAR) where we could have easily dropped the missing values, rather it is of the type Missing data Not at Random (MNAR). In this type, it is vital not to drop the rows as it can adversely impact the analysis and can impact the modeling. So, we took the best approach to replace the nulls with the median of these columns with the help of the "Pandas" library in Python.

There were also a few duplicate records that we dropped. Since the duplicate records were less than 1% of the total data count, it was convenient to drop the duplicates as it did not have much of an impact on the dataset.

Since we had a fair number of features in our dataset, it was essential for us to pick the most impactful features as we wanted to avoid the issue of overfitting. Initially, when we started our modeling, we considered all the features that were present in our dataset. When we applied the Logistic regression and used the metric "Accuracy" to calculate the goodness of the model, it was coming as 99.6%. This was not a very good number as a lot of features just made our model overfit. This is when we understood that using all the features is not giving us the desired output and so we performed the feature selection technique. In the feature selection technique, we used various plots and graphs, and the correlation matrix is shown in Fig 7.1 and Fig 7.2 respectively to highlight the correlated features. By this, we were able to identify the most important features concerning the below 3 different types of output features:

i. Bullied_on school property
ii. Bullied_out of school property
iii. Cyber_bullied

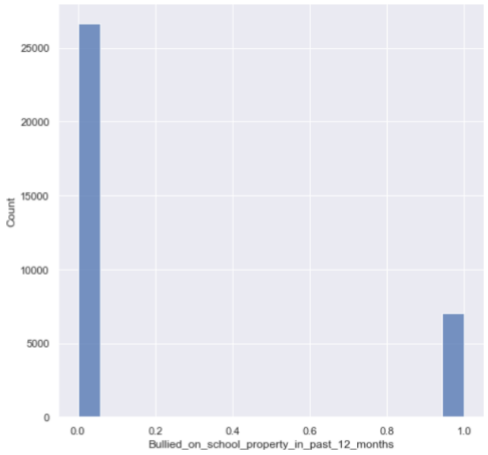
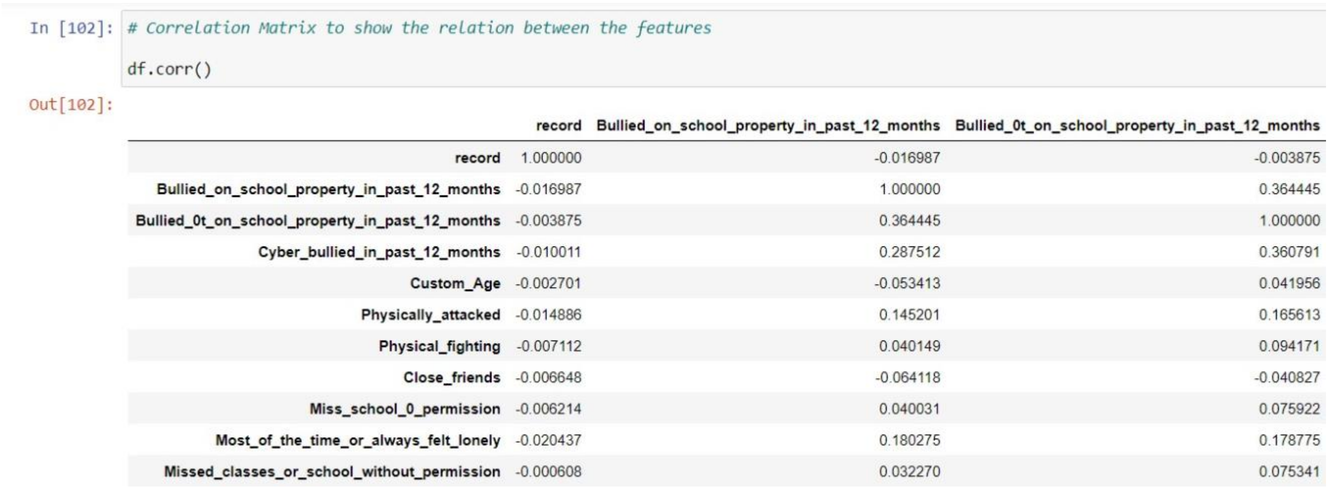


Fig: 7.1 Example of a graph plotted for analysis



```
In [102]: # Correlation Matrix to show the relation between the features

          df.corr()
```

Out[102]:

	record	Bullied_on_school_property_in_past_12_months	Bullied_0t_on_school_property_in_past_12_months
record	1.000000	-0.016987	-0.003875
Bullied_on_school_property_in_past_12_months	-0.016987	1.000000	0.364445
Bullied_0t_on_school_property_in_past_12_months	-0.003875	0.364445	1.000000
Cyber_bullied_in_past_12_months	-0.010011	0.287512	0.360791
Custom_Age	-0.002701	-0.053413	0.041956
Physically_attacked	-0.014886	0.145201	0.165613
Physical_fighting	-0.007112	0.040149	0.094171
Close_friends	-0.006648	-0.064118	-0.040827
Miss_school_0_permission	-0.006214	0.040031	0.075922
Most_of_the_time_or_always_felt_lonely	-0.020437	0.180275	0.178775
Missed_classes_or_school_without_permission	-0.000608	0.032270	0.075341

Fig: 7.2 Correlation Matrix

Another thing that we observed in our dependent feature or output column was that the output was given as ‘True’ and ‘False’. Since with these types of outputs, it is difficult to apply the algorithm, so to make it simpler **Boolean encoding** technique was used to encode the ‘True’ values to ‘1’ and ‘False’ to ‘0’. We also tried to find out the presence of outliers but after we replaced the missing data with the median, we did not find any outliers in the dataset.

Split the data: In the next step, we split the dataset into training and testing. Initially, to split the dataset we used **Randomization** technique. But realized that it is not an optimal approach as it was creating bias and variance in the dataset. So, we have used **Cross Validation** with 10 iterations. Cross-validation being the most popular and powerful technique helped us to improve the performance, robustness, and efficiency of the machine learning model.

Choose an algorithm: The next step was to choose the appropriate algorithm. Since our question was related to the supervised learning classification technique, so we tried different types of modeling techniques. In order to understand what the best model is based on the research question and the dataset that we are analyzing; the best approach is to create the pairplot with the help of the seaborn library to check the relation and pattern of each feature. We plotted the graph using pairplot in which we could see that the data was related and for a few features, the data could very well be divided into the decision boundary whereas, for some features, the data was completely overlapping Fig 7.3. Based on these findings, we tried three models **Logistic Regression**, **Decision Tree Classifier**, and **Random Forest Classifier**.

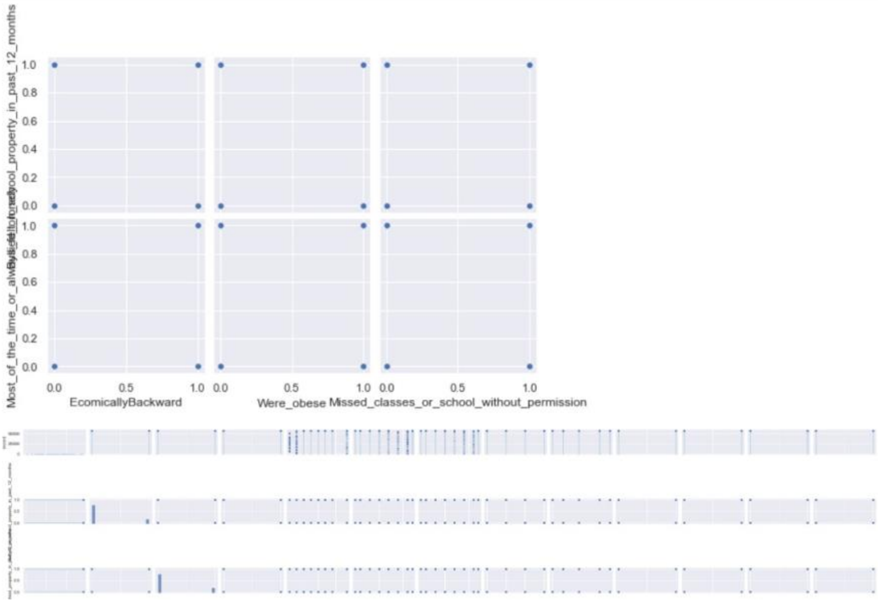
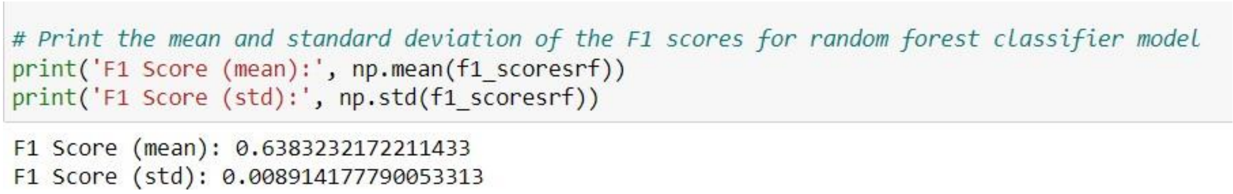


Fig 7.3: Pairplot showing the relationship between the features

Train the model: Using the cross-validation technique the data was split into training data and testing data. The model was trained on this training dataset. And then it was validated to tune the model to ensure that it is performing well.

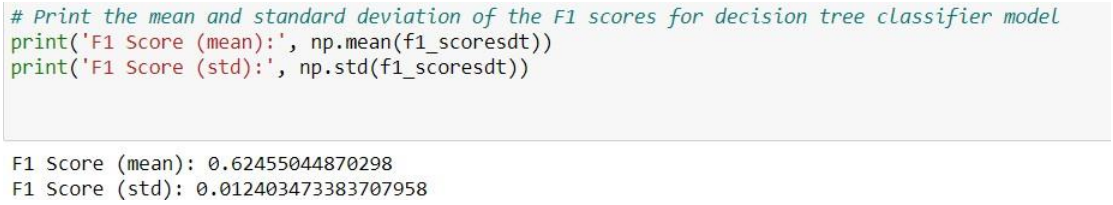
Test the model: Using the testing data the performance of the model is evaluated. Initially, we used the accuracy score as the evaluation metric to measure the performance of the model. Based on this Logistic regression was coming out to be the most accurate model as compared to Random Forest Classifier and Decision Tree Classifier with an accuracy of 44%. However, this was not a valid approach as this is a classification algorithm that we are using. So, we changed to **'F1 Score'** metric to evaluate the performance of the model. F1 Score for **Logistic Regression** is the **highest at 0.648** which can be seen in Fig 7.6. Whereas the F1 score for Random Forest Classifier and Decision Tree Classifier is 0.638 and 0.624 respectively (Fig 7.4 and Fig 7.5). So, our final model is Logistic Regression model.



```
# Print the mean and standard deviation of the F1 scores for random forest classifier model
print('F1 Score (mean):', np.mean(f1_scoresrf))
print('F1 Score (std):', np.std(f1_scoresrf))
```

```
F1 Score (mean): 0.6383232172211433
F1 Score (std): 0.008914177790053313
```

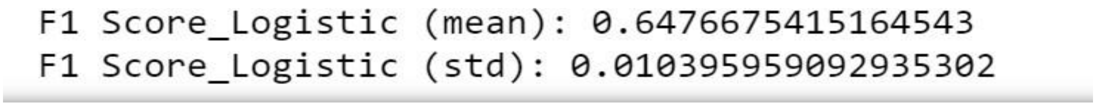
Fig: 7.4: F1 Score of Random Forest Classifier



```
# Print the mean and standard deviation of the F1 scores for decision tree classifier model
print('F1 Score (mean):', np.mean(f1_scoresdt))
print('F1 Score (std):', np.std(f1_scoresdt))
```

```
F1 Score (mean): 0.62455044870298
F1 Score (std): 0.012403473383707958
```

Fig: 7.5: F1 Score of Decision Tree Classifier



```
F1 Score_Logistic (mean): 0.6476675415164543
F1 Score_Logistic (std): 0.010395959092935302
```

Fig: 7.6: F1 Score of Logistic Regression

8. Results leading to answering the question

This section is entailed to showcase our outputs and results by analyzing figures and graphs leading to answering our initial research question.

8.1 Results of EDA

8.1.1 Analysis of all 3 bullying types.

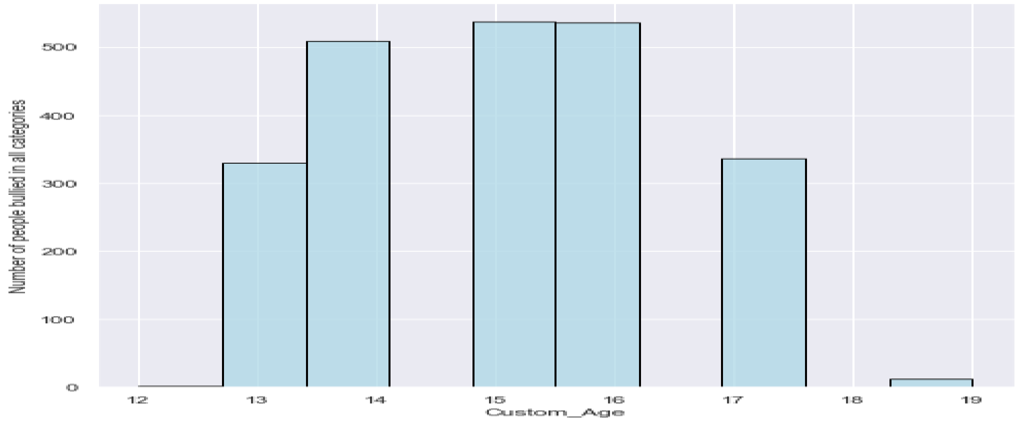


Fig. 8.1 Histogram showing the number of people bullied commonly in all three categories.

Fig 8.1 shows a histogram plot entitled to visualize the count of bullying that occurred based on age group. This was the total number of all types of bullying cases (on campus, out of campus, and cyberbullying). We can see that bullying was most prominent between ages 13 to 17 and it was even more so between the ages 15 and 16. We can also see that there were no bullying cases in the 18-year-old age group and a minimal amount in ages 11 and 12. Although our expectations were met when observing bullying cases to be smallest in larger age groups, a surprise to us was when bullying occurred in 17-year-olds, even more so than in 13-year-olds.

8.1.2 Results of Cyberbullying

We investigated this age factor further by exploring the dataset and classifying our data between males and females as shown in fig 6.6(b) previously. Here we also focused on cyberbullying and the proportionality between the bullying cases in the different genders. We see that both genders have high numbers of bullying cases with females being the greatest contributors. Although there was a slight difference between the genders, we see that they were both equally affected by bullying, especially observing this result in 19-year-olds.

8.1.3 Results of Bullying on and off the School Campus

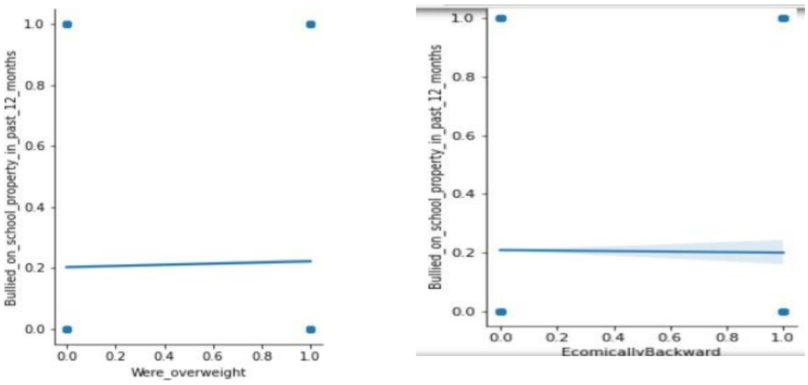


Fig 8.2: Pairplots of the feature economically backward and bullied in school(left) and overweight and bullied on school property(right).

Similarly, being overweight has proven to not have a strong correlation with being bullied on campus. This can be seen in Fig 8.2, illustrating how both features have almost a constant gradient meaning that there is a close to 0 correlation between these features and being bullied on the school campus. The authors could explain this as possibly being due to the changes in modern-day standards of body image having changed throughout the past 10 years. Reports made by the government from older years suggest that body image seemed to highly correlate with the issue of bullying especially among children of this age range[5]. Similarly, figure 8.2 also shows that being economically backward showed a similar correlation to being bullied on school campuses, which meant that victims were still bullied irrespective of their household financial status.

In contrast, our analysis showed victims who felt lonely and whose parents did not understand their problems were bullied. In fact, almost 50% of students felt like their parents did not understand them always or most of the time. Moreover, students who always or sometimes felt lonely showed to be almost 40% as shown previously in figures 6.4 and 6.5. This is such a significant discovery as this correlates with our pairplot findings in figure 6.3, showing a strong negative correlation where students who had more close friends faced less bullying compared to students who had no close friends.

Thus, with these findings, we had a clear idea of our dataset and the features which contributed towards bullying cases based on correlation matrices in fig 7.3 and pairplots shown in section 6. Our research's focal point became not factors like gender and financial status and their impact on bullying as opposed to our initial assumptions.

8.2 Results of modeling.

8.2.1 Metrics on Bullying on Campus.

Table 8.2.1: Performance metrics results for bullying on campus.

Model	F1 Score	ROC AUC
Logistic Regression	0.648	0.766
Decision Tree	0.623	0.644
Random Forest Classifier	0.639	0.724

Table 8.2.1 above shows the accuracy of our 3 models classifying the case of bullying on campus. Looking at the F1 Score, we see that the logistic regression model outperforms the other 3 models with an F1 score of 0.648 and the least accurate model being the Decision Tree Classifier with an F1 Score of 0.623. We have conducted the ROC AUC score, and although this metric gave us higher values, it was not a strong metric to use to determine which model to choose (discussed more in section 9). However, this metric still yielded the highest accuracy score for the logistic regression model with an accuracy score of 0.766. We can also see that the Decision Tree Model performed the worse with an ROC AUC score of 0.644. This is quite a significant difference between the 2 models as opposed to the difference seen using the F1 Score.

8.2.2 Model evaluation on classifying bullying out of campus.

Similarly, table 8.2.2 below shows the F1 Score was the highest with the logistic regression model in predicting the bullying out-of-campus case. Here we see that the model went up to 67% accuracy using the F1 score metric. Although the F1 Score showed a noticeable increase in showing the predictions for this case compared to the previous case above (Table 8.2.1), we can see that not much change had occurred with the ROC AUC scores across the 3 models. For this case, we can also see that logistic regression gave the highest accuracy based on the F1 Scores Metric.

Table 8.2.2: Performance metrics results for bullying out of campus.

Model	F1 Score	ROC AUC score
Logistic Regression	0.671	0.786
Decision Tree	0.648	0.659
Random Forest	0.665	0.748

8.2.3 Model Evaluation on Classifying cyber bullying

Finally, table 8.2.3 below shows the model performance for classifying cyberbullying. Here we can see an F1 Score of 0.656 for Logistic Regression, outperforming both the decision tree and random forest model performing worse. The ROC AUC score can predict this up to an accuracy of 0.751 for this case. Hence, we can conclude that the logistic regression model is the best model for predicting cyberbullying cases.

Table 8.2.3: Performance metrics results for cyberbullying.

Model	F1 Score	ROC AUC score
Logistic Regression	0.656	0.751
Decision Tree	0.618	0.629
Random Forest	0.629	0.706

Up to this point, we have seen that the model which overall outperformed in predicting all 3 cases of bullying was the Logistic regression model, with the highest prediction being the bullying out-of-campus case. The model performing the worst was the Decision Tree Model. Nonetheless, our decisions were based on the F1 score metrics giving a consistently higher accuracy in predicting the cases. Hence, we have chosen the Logistic Regression model as our final model. The details of the full code can be found on the below link:

Link: https://deepnote.com/workspace/techwiz-02ed677d-5309-4cd4-971d-a8b54076a1b5/project/Welcome-ffbca34b-97ca-467d-b67d-b4dc6d9457b9/notebook/Techwiz%20EDA%2BModelling%20(1)-df350cefb64641c691a07ae02a767a85

9. Discussion of interpretation of results

The project is a comprehensive evaluation of the issue of bullying in schools which may be a more serious issue than imagined. As we have aimed to find out the major factors behind the terrorizing of students by bullies, major discoveries were made from our data through exploratory data analysis and modelling predictions.

One of the most important facts which shocked us on analyzing bullying across all the three categories was that bullying was that being older isn’t necessarily a far cry from bullying. More than half of the bullying cases belonged to the older age group, ages 15 to 17 with the younger population much less prone to bullying than older ones. The trend was consistent across all bullying categories and may be partially explained by the accessibility of technology to older students and the prevalence of cyberbullying. Moreover, our study revealed that over half of the students lacked communication with their parents, and parents often fail to understand the concerns of their children. These results line up with the research on bullying, including its complex causes and consequences [6]. It emphasizes the requirement for educators and parents to be watchful and proactive in eliminating bullying and fostering environments that are welcoming and safe for kids of all ages.

We were concerned about several of the other features in our dataset to see if any of them had a strong correlation with the bullying cases. Surprisingly, we found that neither gender nor economic background had a significant impact on the likelihood of being bullied. While being a female may slightly increase the risk of being bullied, being a male does not provide any profound privileges in avoiding bullying. Section 8.1.3 also pertains to the fact that having close social connections could protect one from being vulnerable to such situations. As one may seem, economic backwardness or body image may not be as prominent features leading to bullying in our dataset. However, it is critical to note that our analysis did reveal a concerning trend: bullying on school property was strongly associated with physical violence. This underscores the need for schools to prioritize safety and to take swift and decisive action to prevent and address incidents of bullying. Every child deserves to feel safe and secure in their learning environment, and it is our collective responsibility to ensure that schools are safe spaces for all students.

The decision to choose the best model for our problem was one of the major tasks involved. We conducted an in-depth study on a variety of different models required for the best fit of our data. From the results (section 7), it is evident that logistic regression is the best model that predicts our data.

The F1 score is better than the accuracy

Our findings align with the fact that the f1 score is a better metric for evaluating our logistic model in all three categories of bullying. This was mainly because the f1 score combines the two measures: precision and recall for its calculation which makes it reliable in case of complex data such as ours, where the data is imbalanced. While accuracy only measures the correction predictions to overall predictions ratio, it may not be as much in compliance with our model.

F1 score is better than AUC ROC (Area under the Receiver operating characteristic curve)

In general, the F1 score is preferred over ROC AUC when there is a large difference between the false positive and false negative predictions. From Table 9.1, it is evident that all three modelling cases had a significant difference in false positive and false negative values, thus F1 score is a better metric for evaluating our models.

Table 9.1 Confusion matrix of the three bullying models

Confusion matrix(out of school bullying)-->
[[20065 950]
[3906 2002]]
Confusion matrix(cyber bullying)-->
[[19639 1253]
[4074 1957]]
Confusion matrix(inside school bullying)-->
[[19929 1385]
[3667 1942]]

10.Summary

The model examines "How does bullying impact the overall development of children in schools in the UK" and the effects of bullying on the academic, social, and emotional well-being of children in schools in the UK. Initially, when we were researching this topic, we had certain assumptions. There were certain parameters that we thought would be the most impacted ones contributing to the class of most of the victims being associated. Gender, Economically backward, and physical features to name a few which we had assumed would be the most important features. As we progressed our analysis, to our revelation, gender was not an essential element. The victims were both males and females with a very minute difference. Females are the most impacted victims of cyberbullying among the age range (13-17). Irrespective of the economic background of the victim, there were considerable numbers of victims from both economically backward as well as forward classes. Our algorithm predicts whether the event is a case of bullying and categorizes it as on-campus, off-campus, or cyber-bullying based on the inputs obtained from the victims. As we progressed, there were certain surprising facts that we observed. More than 50% of the reported cases felt that they are not comfortable sharing the details of the incident with even their parents.

However, we have certain limitations due to the lack of availability of more variety of datasets that could give us more opportunity to deep dive into this topic of our research. We tried getting the best accuracy based on the available data and could hit the accuracy of 64% by changing various things like changing the model from Random Forest Classifier which was giving us a lower accuracy score to a Logistic Regression Model. We were using the Accuracy score as the evaluation metric getting only 44% accuracy was not good enough and when we used F1 Score we could increase our score to 64%. The details can be viewed in our code. We could not increase the score further due to the limited features.

Overall, the answer to our research question is that bullying not only impacts physical health leading to issues like obesity and hormonal disbalances at the later stage in their lives, but it also impacts the children's mental health leading to depression which they mostly feel uncomfortable socializing. Due to their introverted nature, they tend also feel difficult to grow in their career as they are less confident. It takes a lot of effort for them to recover from what they have gone through. We do feel that there could be extended research on this topic in order to investigate the effectiveness of anti-bullying organizations in curbing the cases of bullying in schools. Nonetheless, of all the issues we steered in the project, it was a very enriching and informative investigation and overall, we felt very insightful.

11.Group Work Evaluation

From the early stage of our project, we have taken this project in a serious manner. After confirming our group members, we have made sure that we are in contact with each one of them. We wanted to create an effective team and for that, communication, organization, and planning the project were the core of the project. We identified the roles of each member clearly through their experience and that is how we set our goals for the upcoming weeks. We made sure that each one of us get to do every part of the project by splitting the work and shuffling it equally.

Creating an effective team means having supportive and understanding team members. Each member of our group added value to the team and took the initiative to produce a good-quality project. It is evident from our set task that we were up to date as a whole and we constantly had group meetings outside the set lectures and meetings we had. It was one of the strengths of our group that everyone was communicating which allowed us to check on the members if they were struggling with their part of the tasks.

From searching for the topic to modeling the project, every member has contributed equally. Our search for the topic required a lot of meetings and everyone turned up for each meeting. All of us researched the topic and finally, after a lot of research, we finalized our topic with research questions. Then we decided to divide each task amongst each member and make a weekly plan to meet deadlines. For that, we used the Gantt chart to plan out our upcoming weeks according to the deadlines and divided the tasks according to each other’s potential.

Another strength of our group was that we welcomed each one of the members and that helped us to build strong bonds with each other. The benefit of that bond is that none of the team members felt under-confident to suggest and point out any mistakes. Everyone felt comfortable raising recommendations to improve the tasks and everyone discussed through casual conversations. We also had structured meetings every week with our supervisor, and we were prepared to ask questions so that we are not struggling on any task.

12.Individual contributions

12.1 Shamia Ali

The eagerness and ambitious side of each member of the team has been reflected from the first day of our research project. Each member has come up with different topics and different questions that could have potentially been our research topic, however, everyone liked my suggestion and the stories of bullied children that I personally have experienced as well. Through my experience, I am aware of the effect of mental health of bullied children which intrigued me to propose the idea and without any further delay, my team finalized the topic as they were interested to find out how many more children would be facing the same issues.

The next step of our project was to work on the dataset and in that part, I and Nidhi have worked together to finalize the dataset which has been proposed by Devi and Safia. We looked at the different datasets and made sure that our dataset met all the requirements. Our week 3 was mainly focused on the progress check therefore, I worked on the data pre-processing with Nidhi so that data was transformed into a format that would be easier to use for modeling. I also have produced a Gantt chart with Safia to demonstrate our weekly plan for the presentation. The upcoming week involved us working together on EDA which Nidhi and Devi were the main leaders that week and I helped and learned from them. On week 5, me and Safia took the lead, and we completed the EDA which was in progress, and we created some pattern visualization. After completion of EDA, we decided to focus on the modeling, and on week 6, I started initiating modeling training and testing datasets with Devi with the help of the rest of the team. In week 7, I helped to work on the training and the testing model to train model and test whether the model is accurate or not. As we approached week 8, I started working on the preparation for the slides of the mid-term assessment. The next part of the project was to write a report and we divided each section of the report equally in order to complete it on time and that took us weeks 9, 10, and 11. I have written retrieving data, data preparation, and group work evaluation.

To conclude, all of us worked together to check the code and proofread each other's sections of the report. Through this project, I have had a great time working with my team members and we worked together to produce the best possible project.

12.2 Nidhi Priyadarshini

I thoroughly enjoyed working on this group project and was actively involved in all aspects of the project. From the very beginning, I was involved in the topic selection and finding the appropriate dataset for the research along with Devi, Safia, and Shamia. I was adding to their findings with relevant details that I could scrap from the internet. Finally, we were having a lot of datasets in multiple files in which few had irrelevant information as well. I along with Devi and Safia also narrowed down to the best and the most relevant dataset to be used. I also suggested to the team the best approach to work to obtain the best results. As soon as we finalized the dataset, I divided the task of creating the datasheet for the dataset among all the teammates before our Progress Check. I planned the entire timeline for the project and set smaller milestones to accomplish as a team. I always took the opinion and discussed with the entire team and worked as a team player. Some of my major contributions to this project are Finding datasets, Data cleansing, data preprocessing, and some parts of Exploratory data analysis along with Devi and modeling. I was furthermore responsible for the implementation of the machine learning models, and making recommendations to the team about which models to use. Additionally, I guided the team throughout the project and helped my teammates with the issues they faced. Also, I have continuously shared knowledge with the rest of my teammates to ensure collective learning. I have attempted to make everyone feel included in the project and opened a forum for discussion to share innovative ideas.

As it is a team effort we demarcated the task equally, and I was involved in the creation of the

final report's Abstract, Background Research, Rationale for Modeling, and Summary. With this, I also reviewed the work of each teammate and contributed to the content creation.
I took on the accountability of planning the timelines for the project and conducted regular discussions with my team members to ensure that we were on track to meet our goals which helped us to stay motivated and make steady progress towards the final deliverables.

Overall, I am proud of my contribution to the project and team and believe that my active involvement in all aspects of the project helped to ensure its success. It was a great learning experience for me, and I look forward to applying the skills and knowledge I gained in future projects.

12.3 Safia Elmi

During week 1 and week 2, I worked on selecting a topic and the data collection alongside Shamia, Devi, and Nidhi. This involved us all suggesting our ideas and thoroughly communicating with each other. We had regular meetings where I was also able to contribute my findings on the research and the ideas I had. I also worked on preparing the slides for the progress check presented in week 3 alongside Shamia. In week 3 I also worked on supporting the project plan which was in the form of a Gantt chart. I also worked on the visualization graphs of the Exploratory Data Analysis. Closer to the Midterm Demo, in week 7 I worked on training and testing the models alongside Nidhi and Devi. As we progressed with modeling I researched and contributed to choosing the models and the performance metrics.

In the report, I contributed to writing the question development section as well as the results contributing to answering the questions section. I also took part in proofreading the report and supporting any of the group members in their writing.

Overall, I have thoroughly enjoyed working on this report and feel that I have learned a lot as the project progressed. I have always dedicated myself to being a team player and doing my best in helping other members while also meeting my targets. I also feel that we have all worked together well and everyone has shown their different strengths when contributing.

12.4 Devi Chandran

The research idea was very intriguing to me and working on the project gave me good hands-on experience in working on a social issue such as bullying and applying data science algorithms for its analysis, modeling, and predictions.

The first target was to find a research question and the right dataset for it. All of us contributed equally to forming the research question. Safia and I worked on this part, and we found a dataset of interest aligning closely with our area of interest.

I have been involved in the Progress Check and created the slides for the presentation along with Nidhi. Later, I played a major part of exploratory data analysis and was involved in visualizing the plots, especially in the case of cyberbullying. I have tried my best to create all variations in the EDA and find all the trends within the dataset. I could make major discoveries through EDA and Safia, Shamia and Nidhi helped me with finalizing my findings of EDA. Once EDA was completed, I was involved in training and testing the dataset for modeling along with the Safia. I also tried several models on the dataset along with Nidhi and initiated an evaluation of the models using different metrics. I also suggested ideas for the evaluation of the models using different metrics and how to combine them.

A major part of documenting the results was creating the report. I created the sections of the report including the Introduction, Rationale to the group's approach for exploring the data, and Discussion. I have also engaged in checking the code along with the other group members.

Overall, the project has been a major milestone and I enjoyed the time working in a group and sharing my ideas. I always made sure that I complete my work on time to attain our goals consistently. I believe that my contributions were valuable and thank all the team members for their dedication and support throughout the project.

13.Link to Deepnote to access the Code:

The entire code can be viewed from the below link:

Link:
https://deepnote.com/workspace/techwiz-02ed677d-5309-4cd4-971d-a8b54076a1b5/project/Welcome-ffbca34b-97ca-467d-b67d-b4dc6d9457b9/notebook/Techwiz%20EDA%2BModelling%20(1)-df350cefb64641c691a07ae02a767a85

14. References

[1] Samaneh Mojtabaei, Habibah Lateh, and Ashutosh Tiwari. Cyberbullying in UK Schools: A Descriptive View of the Current Status and Its Effects"

[2] Simon K.S. Cheung, Ricky K.H. Law, and Zachary Y. Chan. "Classifying bullying texts using text-mining techniques: A study of UK school children's social media posts"

[3] M. Symeonides and C. Humphrey. "Understanding and Addressing Bullying of Students with Disabilities in Schools in the United Kingdom"

[4] Louise Arseneault, Andrea Danese, and Avshalom Caspi."The Longitudinal Relationship between Bullying and Mental Health among UK Adolescents: Depressive Symptoms as a Mechanism"

[5] Van Geel, M., Vedder, P. and Tanilon, J., 2014. Are overweight and obese youths more often bullied by their peers? A meta-analysis on the relation between weight status and bullying. *International journal of obesity*, *38*(10), pp.1263-1267.

[6] Swearer, S. M., Espelage, D. L., Koenig, B., Berry, B., Collins, A., & Lembeck, P. (2012). A socio-ecological model for bullying prevention and intervention in early adolescence

[7] Leek, J.T. and Peng, R.D., 2015. What is the question?. Science, 347(6228), pp.1314-1315.

[8] Carney, A.G. and Merrell, K.W., 2001. Bullying in schools: Perspectives on understanding and preventing an international problem. School Psychology International, 22(3), pp.364-382.

[9] Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D.G., Hearst, N. and Newman, T., 2001. Conceiving the research question. Designing clinical research, 335.

[10] Office of National Statistics (2018), Bullying in England, April 2013 to March 2018.

[11] Cloke, Christopher and NSPCC (2016) What children are telling us about bullying: Childline bullying report 2015/16. [London]: NSPCC.

[12] Takizawa R, Maughan B, Arseneault L. "Adult health outcomes of childhood bullying victimization: Evidence from a 5-decade longitudinal British birth cohort"