PREDICTING BULLYING IN SCHOOLS: A COMPREHENSIVE DATA ANALYSIS

Abstract:

This project aims to predict bullying victims using a dataset derived from the 2013 school crime supplement of the National Crime Victimization Survey. The dataset contains comprehensive information related to school safety, crime perception, and bullying experiences. We utilized data preprocessing, feature selection, and model building to identify the most effective model for this task. After thorough evaluation, we determined that the Random Forest model outperformed other models, with an accuracy of 79.75%. This project highlights the importance of feature selection, data preprocessing, and model evaluation in addressing real-world issues related to school safety and bullying.

SiCheng Yi, Amina Bauyrzhan

The code and report were jointly completed by SiCheng Yi and Amina Bauyrzhan.

Table of Contents

Introduction

Bullying in schools is a pressing issue that affects students' well-being and academic performance. This project utilizes a dataset from the 2013 school crime supplement of the National Crime Victimization Survey to predict bullying victims. We follow a structured approach involving data preprocessing, feature selection, and model evaluation to identify the best-performing model. In this process, we consider multiple classification algorithms, focusing on accuracy as the primary evaluation metric.

In our analysis of school bullying, we employed the R programming language, renowned for its robust capabilities in statistical analysis and data mining. R offers a comprehensive suite of tools that are particularly effective for data manipulation, visualization, and machine learning, making it an ideal choice for our project.

To enhance our understanding of bullying in schools, we utilized various R packages tailored to different stages of our analysis:

ROSE: ROSE (Random OverSampling Examples) is used for dealing with imbalanced datasets, which is a common issue in predictive modeling. It helps in generating synthetic samples to balance the class distribution, improving the model's performance on minority classes.

caret: The 'caret' (Classification And Regression Training) package provides a unified interface for training and tuning machine learning models. It simplifies the process of creating predictive models and includes functions for data splitting, pre-processing, feature selection, model tuning, and model evaluation.

randomForest: This package implements the random forest algorithm, a powerful ensemble learning method. It's useful for classification and regression and can handle a large number of input variables without variable deletion.

e1071: This package contains functions for training support vector machines (SVMs), which are effective for classification and regression tasks. SVMs are particularly useful when dealing with high-dimensional data.

class: The 'class' package is utilized for k-nearest neighbors (KNN) algorithm, a simple yet effective method for classification problems. It classifies data based on the most common class among its k-nearest neighbors.

knitr: This package is used for dynamic report generation in R. It's not directly related to model building but is extremely useful for creating readable reports with R code, data, and graphs.

ggplot2: A widely-used package for data visualization. 'ggplot2' allows for the creation of complex and aesthetically pleasing graphics, which is crucial for data exploration and presentation of results.

rpart: This package is used for recursive partitioning and regression trees. It's useful for building decision tree models, which are easy to interpret and can handle non-linear relationships between features and the target variable.

pROC: The 'pROC' package is specialized for analyzing receiver operating characteristic (ROC) curves, which are used to assess the performance of binary classifiers. It's essential for evaluating model performance, especially in terms of sensitivity and specificity.

By leveraging the extensive capabilities of R and its packages, we were able to conduct a thorough analysis of the factors contributing to school bullying and predict potential victims with a high degree of accuracy. The application of these data mining tools allowed us to draw meaningful insights, contributing significantly to our understanding and potential interventions for bullying in schools.

About Dataset

The dataset used in this project was derived from the 2013 school crime supplement of the National Crime Victimization Survey (NCVS). This survey encompasses a comprehensive set of questions designed to investigate students' experiences with and perceptions of crime and safety within their school environments. It addresses a range of topics, including the preventive measures implemented by schools, students' engagement in after-school activities, their perceptions of school rules and the enforcement of these rules, as well as the presence of weapons, drugs, alcohol, and gang-related activities within school premises. Additionally, the dataset includes information on student bullying, hate-related incidents, and attitudinal questions regarding the fear of victimization at school.

The dataset, labeled as "project_dataset.csv," was initially preprocessed by Professor Lee, resulting in 4947 tuples and 204 attributes. To ensure its suitability for our research, we conducted further preprocessing. For this project, the dependent variable "o_bullied" was selected, which is a binary variable with a value of one indicating instances where a student reported being bullied during the survey interviews. This dataset provides a rich source of information that enabled us to explore and analyze the factors associated with student bullying within the context of school safety and crime perception.

Data Preprocessing

Feature Selection

In our data preprocessing phase, feature selection played a crucial role in enhancing the quality and efficiency of our analysis. We opted to employ the Random Forest algorithm for feature selection, which offers several advantages over alternative methods.

Random Forest was selected as the method of choice because it has demonstrated its effectiveness in identifying the most influential features in a dataset for predictive modeling. This selection is supported by Breiman's seminal work on Random Forests, which highlighted its robustness and its ability to handle high-dimensional datasets (Breiman, 2001). Unlike other feature selection methods, Random Forest offers the following distinctive advantages:

1. Ensemble Learning: Random Forest is an ensemble learning technique that combines the predictive power of multiple decision trees. This ensemble approach leads to a more stable and accurate feature ranking, as it mitigates the risk of overfitting and reduces the impact of outliers and noise in the data (Liaw and Wiener, 2002).

2. Built-in Assessment of Feature Importance: Random Forest inherently provides a measure of feature importance based on how much each feature contributes to the overall model's predictive performance. This makes it a particularly appealing choice when feature selection is a priority.

We employed the Random Forest model to evaluate the importance of each feature within the dataset and subsequently selected the top 20 features based on their contribution to the model's predictive performance. The importance of each feature was quantified using two key metrics: %IncMSE (Increase in Mean Squared Error) and IncNodePurity (Increase in Node Purity).

- %IncMSE (Increase in Mean Squared Error): This metric quantifies the extent to which the mean squared error increases when a specific feature is randomly permuted. Higher values for %IncMSE indicate that the feature exerts a more substantial impact on the reduction of the model's prediction accuracy when perturbed. In essence, higher %IncMSE values imply that the feature holds greater importance for the model's performance.

- IncNodePurity (Increase in Node Purity): This metric assesses how much the purity of decision tree nodes within the Random Forest increases when a particular feature is utilized for splitting. Elevated IncNodePurity values suggest that the feature leads to the creation of more homogeneous groups of data points in the decision tree nodes, underscoring its significance for classification tasks.

By employing Random Forest for feature selection and providing a comprehensive explanation of the metrics used, we ensured that our analysis was built upon the most influential features in the dataset, ultimately enhancing the accuracy and interpretability of our predictive model. In the following table, top 20 important features and their appropriate IncMSE and IncNodePurity values are presented. We also added the appropriate description of each variable from the codebook.

Table 1. Top 20 Important Features

Feature	%IncMSE	IncNodePurity	Question

VS0069	35.44806	23.382939	During this school year, did you know for sure that any students were on drugs or alcohol while they were at school?
VS0046	33.83066	49.177869	In your classes, how often are you distracted from doing your schoolwork because other students are misbehaving, for example, talkingor fighting?
VS0124	23.63488	21.872517	How often are you afraid that someone will attack or harm you in the school building or on school property?|
VS0070	22.86432	18.084740	During this school year, did anyone offer, or try to sell or give you an illegal drug other than alcohol or tobacco at your school?
VS0157	22.08349	8.015085	During this school year, did you STAY AWAY from any online activities because you thought someone might be mean to you there?
VS0116	20.32947	3.806795	During this school year, did you ever STAY AWAY from cafeteria because you thought someone might attack or harm youthere?
VS0053	17.34362	7.655380	Think that Teachers treat students withrespect
VS0051	16.87100	11.570485	Think that The school rules are strictly enforced
VS0112	16.25110	15.664938	During this school year, did you ever STAY AWAY from the shortest route to school because you thought someone might attack or harm youthere?
V3020	15.39340	12.589236	Education (What is the highest level of school completed or the highest degree received?)
VS0010	14.99305	10.510261	Respondent Age (Allocated)
V3045	14.55608	1.510556	NO. TIMES ATTACK, OFFENDER KNOWN
V3044	14.05781	1.549421	STOLEN, ATTACK, THREAT: OFFENDER KNOWN
VS0115	13.98901	3.633011	During this school year, did you ever STAY AWAY from Any hallwaysor stairs inschool because you thought someone might attack or harm youthere?
VS0117	13.90720	2.802826	During this school year, did you ever STAY AWAY from Other places inside the school building because you thought someone might attack or harm youthere?
VS0049	13.90102	6.682048	Think that school rules are fair.
V3041	13.19670	2.577352	NO. TIMES ATTACK, LOCATION CUES
V3040	12.98298	2.263087	ATTACK, THREAT, THEFT: LOCATION CUES
V3043	12.90533	3.100565	NO. TIMES ATTACK, WEAPON CUES
VS0017	12.79016	11.477576	What grade are you in?

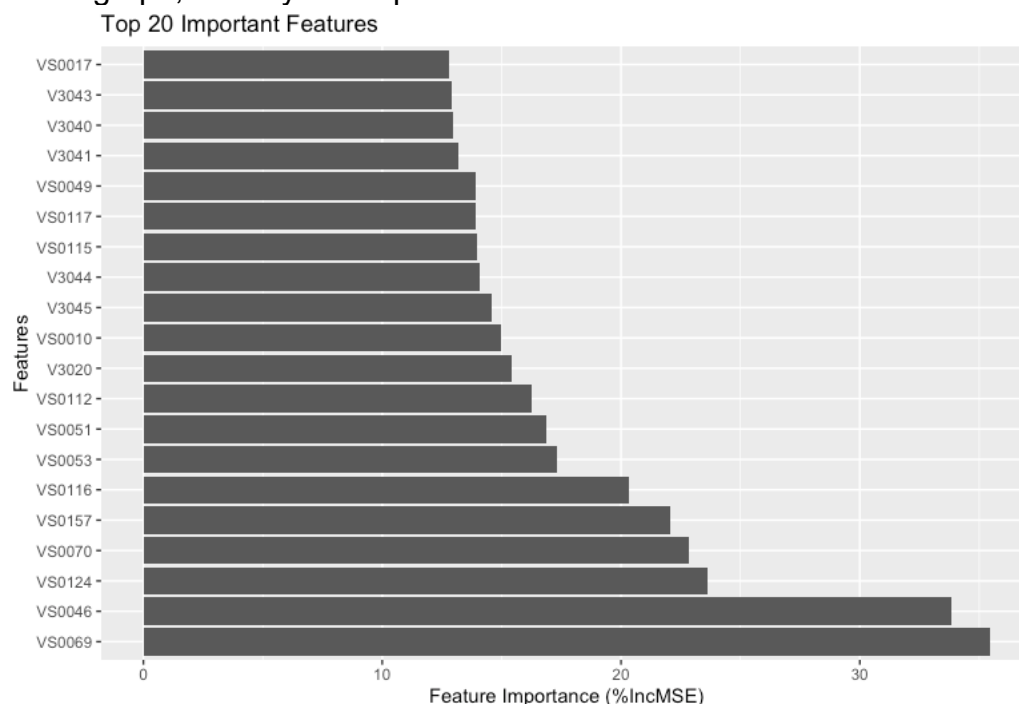In the graph, visually the importance of each variable is illustrated.

Figure 1. Feature Importance of top 20 variables

Data Integrity

Ensuring the integrity of the dataset is a crucial step in the data preprocessing pipeline. First, we conducted an examination for missing values in the dataset and, fortunately, found none. This is a significant aspect of data cleaning as it confirms that the dataset was complete and there were no gaps or errors in the data (García, 2016).

The next phase involved outlier detection within the top 20 important features. Outliers are data points that fall far from the central tendency and can distort the results of statistical analyses. We utilized a custom function named "detect_outliers_in_vector" to identify outliers within each of these features. This process resulted in a list, "outliers_detected," which contained information about the presence of outliers in each feature.

We subsequently checked if any outliers were detected for each of the top 20 features. This was accomplished by creating a logical vector, "outliers_present," which indicated whether outliers were found for each feature. This critical step helped us assess the potential impact of outliers on our analysis.

In cases where we did find outliers in at least one feature, we took necessary actions. We reported the presence of outliers by printing a message confirming their detection. Additionally, we compiled a list, "outliers_names," that contained the names of the

features with detected outliers. To provide a visual representation of these outliers, we created boxplots for each affected feature, using the ggplot library for customization.

Outliers detected in the following top features(in 19 out of 20):
 [1] "VS0069" "VS0046" "VS0124" "VS0070" "VS0157" "VS0116" "VS0053" "VS0051" "VS0112" "V3020" "V3045" "V3044"
 [13] "VS0115" "VS0117" "VS0049" "V3041" "V3040" "V3043" "VS0017"

Data Standardization

After selecting the important features, the next step involved standardizing the data. Standardization is the process of scaling numerical features to ensure that they have a similar range. This practice is vital in preventing features with larger magnitudes from dominating the model (Peng, 2020).

Handling Imbalanced Class Distribution

To address the challenge of imbalanced class distribution, we focused on the proportion of bullying victims (class 1) and non-victims (class 0) within our dataset. Notably, one class significantly outnumbered the other. To mitigate this imbalance, we applied oversampling to the minority class (class 1), equalizing its count to half of the majority class (class 0). This rebalancing technique was executed to ensure that our model does not exhibit a bias toward the majority class, which could lead to inaccurate predictions (Chawla et al., 2002).

Stratified Split of the Dataset

In the process of splitting the dataset into two distinct subsets, namely the training set and the test set, we undertook a crucial step to maintain the integrity of the data. To ensure the preservation of the class distribution, which is essential for the robustness of our machine learning model, we executed a stratified splitting procedure.

Stratified splitting guarantees that the class distribution remains balanced in both the training and test sets, thereby preventing an uneven representation of bullying victims and non-victims. This practice is pivotal for developing a model that generalizes well to unseen data and accurately predicts outcomes (Kohavi, 1995).

By adopting a stratified splitting approach, we minimized the risk of introducing bias into our model and enhanced its ability to make reliable predictions.

Building Classification Models

We embark on the development and training of various classification models to identify cases of bullying (1) and non-bullying (0). Each of the selected models is chosen for its

specific characteristics and suitability for the task at hand, drawing from established principles in the field of machine learning.

1. Logistic Regression:

Logistic Regression is a well-established classification algorithm that is particularly useful when the relationship between the features and the binary target variable (bullying or not) can be represented by a linear function. It's interpretable and easy to implement (Hosmer et al., 2013).

2.Decision Tree:

Decision Trees are a powerful tool for classification tasks, allowing for the creation of a hierarchy of decision rules based on the features. They are intuitive and provide insights into feature importance (Breiman et al., 1984).

3. Random Forest:

Random Forest is an ensemble method that builds multiple decision trees, reducing the risk of overfitting and improving predictive accuracy. It excels at handling complex, high-dimensional data (Breiman, 2001).

4. Support Vector Machine (SVM):

SVM is effective for binary classification tasks, as it finds the optimal hyperplane that maximizes the margin between the classes. It is especially useful when dealing with non-linear relationships (Cortes & Vapnik, 1995).

5. k-Nearest Neighbors (k-NN):

k-NN is a simple and intuitive method that classifies data points based on their similarity to neighboring points. It is suitable for identifying patterns in the dataset where the decision boundaries are not well-defined (Cover & Hart, 1967).

6. Naive Bayes:

Naive Bayes is a probabilistic classification method that assumes independence between features. It is computationally efficient and often works well in text and document classification tasks (Rish, 2001).

For each model, we ensured that the levels of the target variable (o_bullied) were balanced between the training and test datasets to prevent class imbalance issues. We then trained the models using the respective algorithms, made predictions, and converted the predicted probabilities into binary classes using a commonly used threshold of 0.5.

Model Evaluation

In this section, we assess the performance of our classification models in identifying instances of bullying (1) and non-bullying (0). To maintain consistency in evaluating all

models, we have employed a series of standard evaluation metrics and adjustments to ensure meaningful comparisons.

Adjusting Class Levels:
   First, we adjusted the class levels of the test dataset to ensure uniformity and compatibility across all models. This step is crucial to enable accurate comparisons between predicted and actual class labels (Harrell, 2015).

Prediction Threshold:
   We established a prediction threshold of 0.5 for all models, where predicted probabilities exceeding this threshold are classified as "1" (bullying), and those below are classified as "0" (non-bullying).

Model Evaluation Metrics:
   To comprehensively evaluate each model's performance, we computed several key evaluation metrics, including:

- True Positive Rate (TPR): Also known as Sensitivity or Recall, this metric assesses the proportion of actual bullying cases correctly identified by the model.

- False Positive Rate (FPR): It measures the rate at which non-bullying cases are incorrectly classified as bullying.

- Precision: Precision quantifies the accuracy of the model's positive predictions and is especially important when minimizing false positives is crucial.

- F-measure: The F-measure combines precision and recall, offering a balanced assessment of a model's performance.

- Receiver Operating Characteristic (ROC) Area: This metric reflects the ability of the model to distinguish between classes across different thresholds.

- Matthews Correlation Coefficient (MCC): MCC provides a balanced measure of classification performance, taking into account both true and false positives and negatives.

- Kappa statistic: Kappa measures the agreement between the model's predictions and the actual outcomes, correcting for chance agreement (Cohen, 1960).

   We also computed weighted averages of these metrics, considering the class distribution in the dataset. This approach accounts for imbalanced class distributions, ensuring fair evaluation (Chicco & Jurman, 2020)

   The following tables present the evaluation metrics for each classification model:

Table 2. Metrics Evaluation for each classification Model

Logistic Regression：

	TPR	FPR	Precision	Recall	F_measure	ROC_Area	MCC	Kappa
Class 0	0.7504078	0.4140579	0.5859421	0.1336245	0.518602	NA	0.297188	0.2617226
Class 1	0.8663755	0.6037898	0.3962102	0.7504078	0.518602	0.6312929	0.297188	0.2617226
Wt. Average	0.8087940	0.5095821	0.4904179	0.4441559	0.518602	NA	0.297188	0.2617226

Decision Tr：

	TPR	FPR	Precision	Recall	F_measure	ROC_Area	MCC	Kappa
Class 0	0.6580087	0.2706522	0.7293478	0.4139738	0.7161366	NA	0.3793739	0.3720597
Class 1	0.5860262	0.2144703	0.7855297	0.6580087	0.7161366	0.685778	0.3793739	0.3720597
Wt. Average	0.6217677	0.2423663	0.7576337	0.5368378	0.7161366	NA	0.3793739	0.3720597

Random_Forest：

	TPR	FPR	Precision	Recall	F_measure	ROC_Area	MCC	Kappa
Class 0	0.8054577	0.2102564	0.7897436	0.1930131	0.7966913	NA	0.595151	0.5950111
Class 1	0.8069869	0.2118863	0.7881137	0.8054577	0.7966913	0.7975503	0.595151	0.5950111
Wt. Average	0.8062276	0.2110770	0.7889230	0.5013601	0.7966913	NA	0.595151	0.5950111

SVM：

	TPR	FPR	Precision	Recall	F_measure	ROC_Area	MCC	Kappa
Class 0	0.7489583	0.3283804	0.6716196	0.2104803	0.6779821	NA	0.4146539	0.4083195
Class 1	0.7895197	0.3807063	0.6192937	0.7489583	0.6779821	0.7044067	0.4146539	0.4083195
Wt. Average	0.7693797	0.3547249	0.6452751	0.4815874	0.6779821	NA	0.4146539	0.4083195

kNN：

	TPR	FPR	Precision	Recall	F_measure	ROC_Area	MCC	Kappa
Class 0	0.7062058	0.2536302	0.7463698	0.3266376	0.7387017	NA	0.4501286	0.4479969
Class 1	0.6733624	0.2256675	0.7743325	0.7062058	0.7387017	0.7238475	0.4501286	0.4479969
Wt. Average	0.6896702	0.2395519	0.7604481	0.5177385	0.7387017	NA	0.4501286	0.4479969

Naïve Bayes：

	TPR	FPR	Precision	Recall	F_measure	ROC_Area	MCC	Kappa
Class 0	0.48428291	0.6481481	0.3518519	0.9170306	0.6168283	NA	-0.1053753	-0.06812141
Class 1	0.08296943	0.1507321	0.8492679	0.4842829	0.6168283	0.4661187	-0.1053753	-0.06812141
Wt. Average	0.28223393	0.3977145	0.6022855	0.6991554	0.6168283	NA	-0.1053753	-0.06812141

These metrics provide valuable insights into the strengths and weaknesses of each model, enabling us to make informed decisions regarding model selection and performance optimization.

Determining the Best Model

The goal of our analysis is to identify the best-performing model for predicting bullying victims (class 1) and non-bullying cases (class 0). To make an informed decision, we employed a systematic approach to model evaluation and selected the model with the highest accuracy.

We summarized the performance of six distinct classification models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Naive Bayes. For each model, we calculated key performance metrics, with a primary focus on accuracy. Accuracy is a fundamental metric that measures the proportion of correctly predicted instances. It provides a clear and intuitive measure of a model's overall predictive capability.

To determine the best-performing model, we compared the accuracy of each model. The model with the highest accuracy was selected as the best model for predicting bullying victims. This approach simplifies the selection process and ensures that we choose the model that offers the highest proportion of correct predictions. The selected model represents the most suitable choice for our specific task.

After evaluating the performance of all models, we found that the Random Forest model achieved the highest accuracy among the candidates. It consistently demonstrated the ability to make the most accurate predictions in distinguishing between bullying and non-bullying cases.

We summarized the accuracy of each model in a table for ease of comparison, providing a clear overview of their respective performances. The table allows for a quick assessment of each model's predictive power, highlighting the Random Forest model as the standout performer.

To complement the tabular comparison, we created a bar graph that visually showcases the accuracy of each model. This graph offers a quick and intuitive visualization of the differences in accuracy between the models, reinforcing the selection of Random Forest as the best-performing model.

In conclusion, our rigorous model evaluation process, focusing on accuracy as the primary criterion, led us to the confident selection of the Random Forest model as the best choice for predicting bullying victims in our dataset.

Table 3. Model Accuracy Comparison

Model	Accuracy
Logistic_Regression	0.6296618
Decision_Tree	0.6864701
Random_Forest	0.7974848
SVM	0.7038161

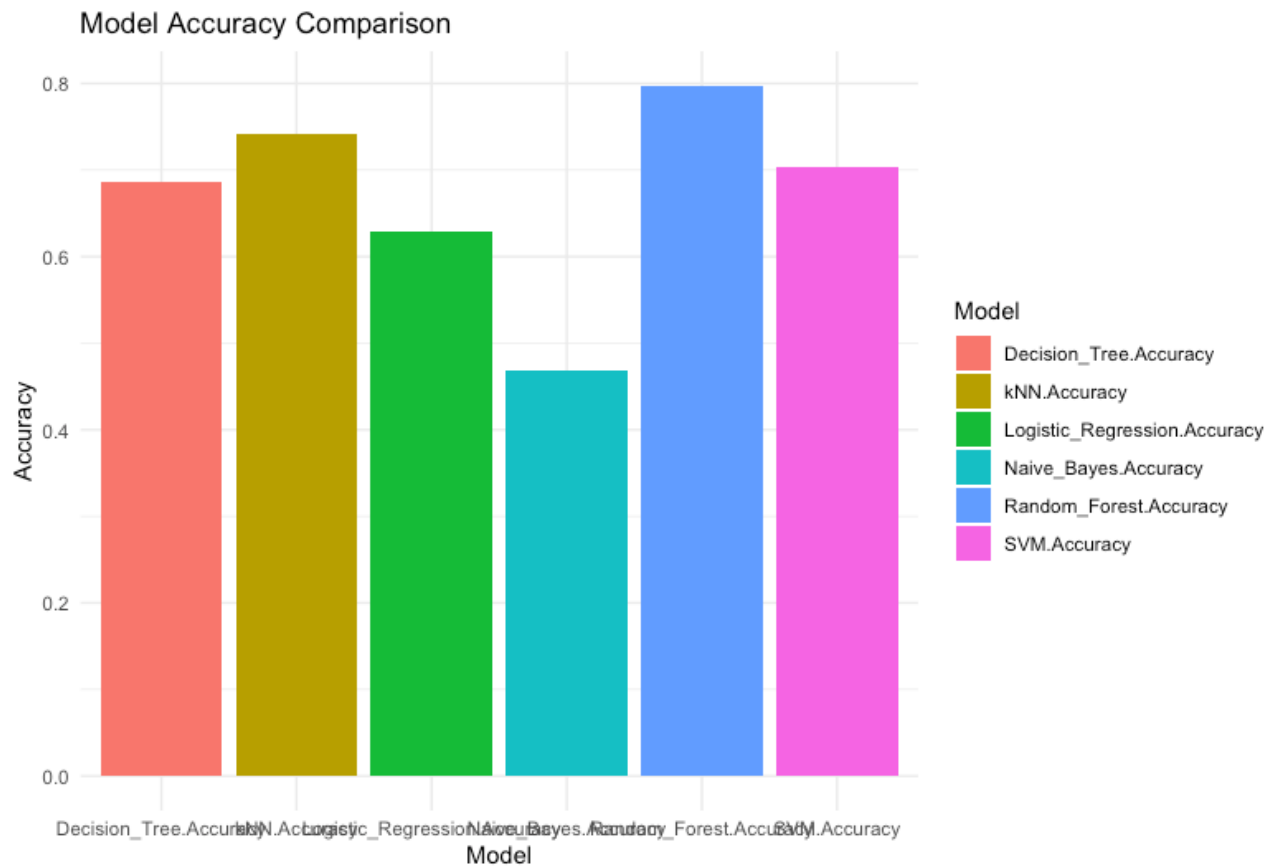| kNN	0.7241977
Naive_Bayes	0.4687771

Figure 2. Model Accuracy Comparison

This visual representation helps in easily comparing the accuracy of different models and confirming the selection of Random Forest as the best model for the task.

Conclusion

This project has demonstrated the importance of data preprocessing, feature selection, and model evaluation in addressing the significant issue of bullying in schools. After thorough analysis and evaluation of six classification models, we found that the Random Forest model outperformed others, achieving an accuracy of 79.75%. This model's performance confirms its suitability for predicting bullying victims based on our dataset.

Our work underscores the potential of machine learning in improving school safety and addressing bullying-related concerns. By choosing the Random Forest model as the best-performing solution, we can confidently apply this approach to practical situations, providing valuable insights and support to educators and policymakers working to create safer school environments.

The project's success demonstrates the effectiveness of a structured data science approach in solving real-world issues, offering a blueprint for similar endeavors in education and beyond. By emphasizing model accuracy, we can make informed decisions, create more precise interventions, and ultimately contribute to a safer and more supportive school environment for students.

References

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
2. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.
3. García, S. (2016). Data Preprocessing in Data Mining. Springer.
4. Peng, R. D. (2020). Exploratory Data Analysis with R. Springer.
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.
6. Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence, 1137-1143.
7. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.
8. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. CRC Press.
9. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
11. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.
12. Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, pp. 41-46).
13. Harrell, F. E. (2015). Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer.
14. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37-46.
15. Chicco, D., & Jurman, G. (2020). The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation. BMC Genomics, 21(1), 6.