Predicting Bullying Victims in Schools

Leveraging the 2013 National Crime Victimization Survey

SiCheng Yi, Amina Bauyrzhan

Boston University

BOSTON UNIVERSITY

Outline

Objective: Utilizing the 2013 school crime supplement of the National Crime Victimization Survey dataset to predict bullying victims.

Approach: Structured methodology involving data preprocessing, feature selection, and model evaluation.

Tool of Choice: R, packages: ROSE, caret, e1071, randomForest, class, knitr, ggplot2, rpart, pROC.

Key Focus: Exploring multiple classification algorithms, emphasizing accuracy as the primary evaluation metric

BOSTON UNIVERSITY

About Dataset

Source: 2013 School Crime Supplement of the National Crime Victimization Survey (NCVS)

Derived from a comprehensive set of questions on students' experiences with and perceptions of crime and safety within school environments.

Explores various aspects such as preventive measures, after-school activities, school rules enforcement, presence of weapons, drugs, alcohol, gang-related activities, bullying, hate-related incidents, and fear of victimization.

Dataset Name: project_dataset.csv

Further processed from the original dataset (4947 tuples, 204 attributes) to focus on the dependent variable "o_bullied" indicating instances of reported bullying.

BOSTON UNIVERSITY

Data Preprocessing and Feature Selection

Cleaning: Addressing missing values and outliers.

Normalization: Scaling features for uniformity.

Encoding: Converting categorical data into numerical format.

Feature Selection: Identifying Key Predictors

Method: Utilized Random Forest for feature importance analysis.

Outcome: Identified top 20 features impacting bullying prediction.

Significance: Streamlined model by focusing on most influential variables.

BOSTON UNIVERSITY

Feature Selection with Random Forest

Importance of Feature Selection: Crucial role in enhancing analysis quality and efficiency.

Methodology: Utilizing Random Forest algorithm for its robustness and ability to handle high-dimensional datasets.

Metrics: %IncMSE and IncNodePurity to quantify feature importance.

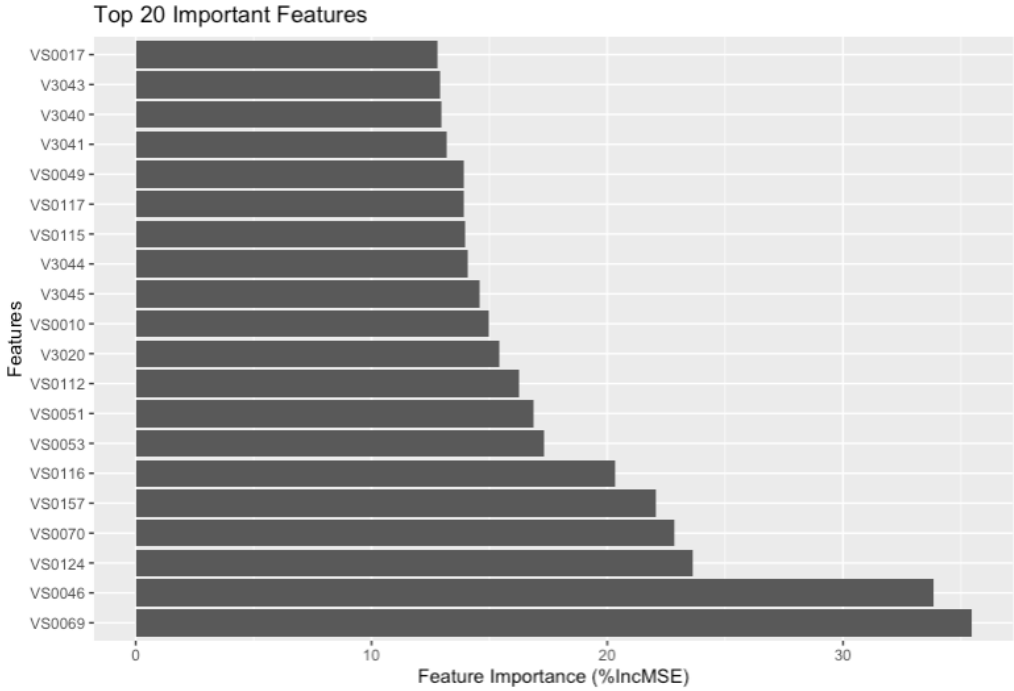

Figure 1. Feature Importance of top 20 variables

BOSTON UNIVERSITY

Classification Models

Logistic Regression

Type: Linear model for binary classification.

Characteristics: Predicts the probability of occurrence of an event by fitting data to a logistic curve.

Suitability: Good for understanding the influence of several independent variables.

Decision Tree

Type: Tree-like model of decisions.

Characteristics: Uses a tree-like graph to model decisions and their possible consequences.

Suitability: Easy to interpret and can handle both numerical and categorical data.

Random Forest

Type: Ensemble of Decision Trees.

Characteristics: Operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes of individual trees.

Suitability: Reduces overfitting in decision trees and improves accuracy.

BOSTON UNIVERSITY

Classification Models

Support Vector Machine (SVM)

Type: Supervised learning model.

Characteristics: Identifies a hyperplane in an N-dimensional space that distinctly classifies the data points.

Suitability: Effective in high dimensional spaces and in cases where the number of dimensions is greater than the number of samples.

k-Nearest Neighbors (k-NN)

Type: Non-parametric method.

Characteristics: Classifies data points based on how their neighbors are classified.

Suitability: Simple and effective, good for data with little or no information about distribution.

Naive Bayes

Type: Probabilistic classifier.

Characteristics: Applies Bayes' theorem with the assumption of independence between every pair of features.

Suitability: Performs well in multi-class prediction and when the assumption of independence holds.

BOSTON UNIVERSITY

Model Evaluation Metrics

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

BOSTON UNIVERSITY

Model Evaluation: 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

BOSTON UNIVERSITY

Model Evaluation: Decision Tree

	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

BOSTON UNIVERSITY

Model Evaluation: 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

BOSTON UNIVERSITY

Model Evaluation: Support Vector Machine

	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

BOSTON UNIVERSITY

Model Evaluation: k-Nearest Neighbors

	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

BOSTON UNIVERSITY

Model Evaluation: Naive 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

BOSTON UNIVERSITY

Best Model Determination

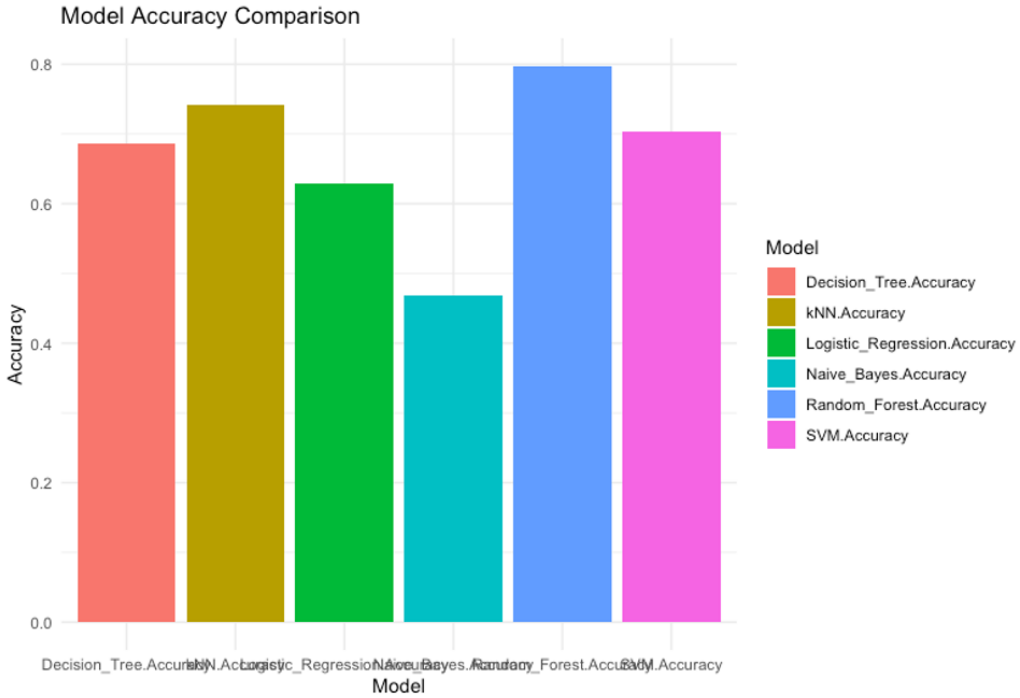

Table 1. Model Accuracy Comparison

Model	Accuracy
Logistic_Regression	0.6296618
Decision_Tree	0.6864701
Random_Forest	0.7974848
SVM	0.7038161
kNN	0.7241977
Naive_Bayes	0.4687771

The model with the highest accuracy was selected as the best model for predicting bullying victims.

After evaluating the performance of all models, we found that the Random Forest model achieved the highest accuracy among the candidates.

BOSTON UNIVERSITY

Conclusion

Key Findings: Random Forest emerges as the best-performing model with 79.75% accuracy.

Real-world Impact: Applicability in improving school safety and addressing bullying-related concerns.

Structured Approach: Demonstrating the effectiveness of a systematic data science methodology.

BOSTON UNIVERSITY

Any Questions?

BOSTON UNIVERSITY

Thank You!

BOSTON UNIVERSITY