Optimizing Models for Predicting Bullying in Adolescents

Group 12

Yasi Asgari, Nikki Gharachorloo, Alex Herron
Aryann Dharamsey, Hailie Nguyen, Victor Cui

Data Description

- Global School-Based Student Health Survey (GSHS) conducted in Argentina
- **57,095** rows
- **18** self-reported columns
- Columns include age and gender of participants, where the participant was bullied (cyber, on-campus, off-campus), how many close friends the participant has, whether parents are understanding, etc.

record	Bullied_on_school_property_in_past_12_months	Bullied_not_on_school_property_in_past_12_months	Physically_attacked	Physical_fighting	Felt_lonely	Close_friends	Miss_school_no_permission	Other_students_kind_and_helpful	Parents_understand_problems
1	Yes	Yes	0 times	0 times	Always	2	10 or more days	Never	Always
2	No	No	0 times	0 times	Never	3 or more	0 days	Sometimes	Always
3	No	No	0 times	0 times	Never	3 or more	0 days	Sometimes	Always
4	No	No	0 times	2 or 3 times	Never	3 or more	0 days	Sometimes	
5	No	No	0 times	0 times	Rarely	3 or more	0 days	Most of the time	Most of the time
...	...	...	...	...	...	...	...	...	...
57091	No	Yes	0 times	4 or 5 times	Sometimes	3 or more	0 days	Sometimes	Sometimes
57092	No	No	0 times	0 times	Rarely	1	0 days	Sometimes	Never
57093	No	No	0 times	0 times	Sometimes	3 or more	0 days	Rarely	Sometimes
57094	No	Yes	0 times	0 times	Sometimes	2	0 days	Most of the time	Rarely
57095	No	Yes	0 times	0 times	Always	2	1 or 2 days	Rarely	Never

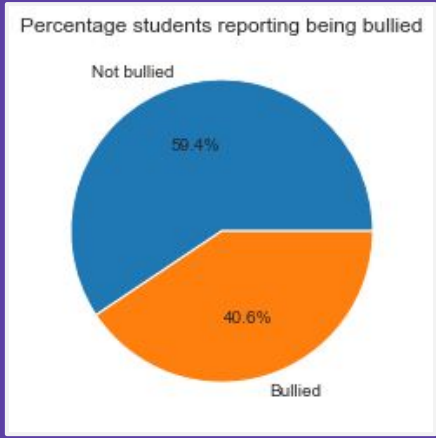
Data Preprocessing

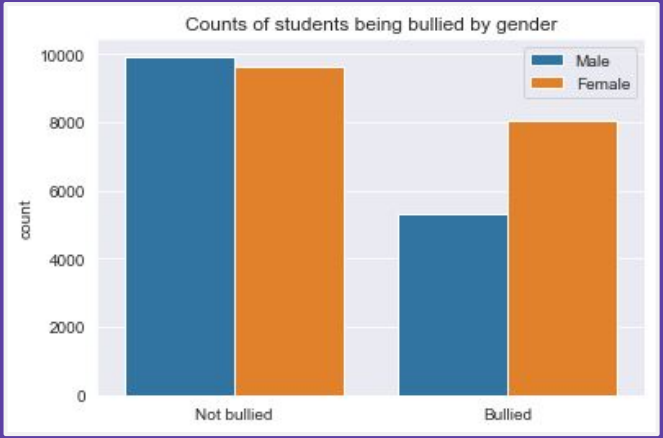
- Remove rows with null values
  - Left with **32,938** observations
- Encode categorical variables using numeric values
- Combined three classes of bullying **(cyber, on-campus, off-campus)** into one
- Split data into training **(0.8)** and test set **(0.2)**

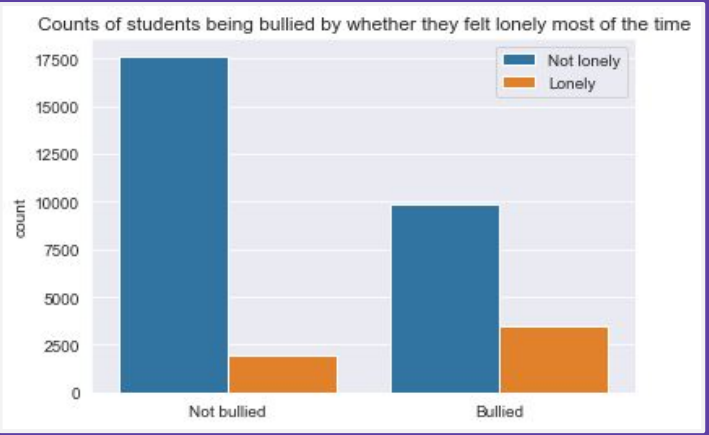
Cyber_bullied_in_past_12_months	Sex	Physically_attacked	Physical_fighting	Felt_lonely	Close_friends	Miss_school_no_permission
0	0	0	0	0	3	0
0	0	0	1	0	3	0
0	0	0	1	0	3	2
1	0	0	2	3	3	0
0	0	0	2	3	2	0
...	...	...	...	...	...	...
0	0	0	0	1	3	2
0	0	0	2	0	3	0
1	0	7	2	1	3	0
0	0	0	3	2	3	0
1	1	0	0	4	2	1

	count	mean	std	min	25%	50%	75%	max
Bullied_on_school_property_in_past_12_months	32938.0	0.208574	0.406295	0.0	0.0	0.0	0.0	1.0
Bullied_not_on_school_property_in_past_12_months	32938.0	0.222023	0.415613	0.0	0.0	0.0	0.0	1.0
Cyber_bullied_in_past_12_months	32938.0	0.224057	0.416966	0.0	0.0	0.0	0.0	1.0
Sex	32938.0	0.536736	0.498656	0.0	0.0	1.0	1.0	1.0
Physically_attacked	32938.0	0.340610	1.031779	0.0	0.0	0.0	0.0	7.0
Physical_fighting	32938.0	0.460319	1.150626	0.0	0.0	0.0	0.0	7.0
Felt_lonely	32938.0	1.345832	1.166598	0.0	0.0	1.0	2.0	4.0
Close_friends	32938.0	2.508440	0.855752	0.0	2.0	3.0	3.0	3.0
Miss_school_no_permission	32938.0	0.475591	0.906017	0.0	0.0	0.0	1.0	4.0
Other_students_kind_and_helpful	32938.0	2.335448	1.179191	0.0	1.0	2.0	3.0	4.0
Parents_understand_problems	32938.0	2.073077	1.469086	0.0	1.0	2.0	3.0	4.0
Most_of_the_time_or_always_felt_lonely	32938.0	0.163641	0.369955	0.0	0.0	0.0	0.0	1.0
Missed_classes_or_school_without_permission	32938.0	0.287571	0.452637	0.0	0.0	0.0	1.0	1.0
Were_underweight	32938.0	0.019855	0.139506	0.0	0.0	0.0	0.0	1.0
Were_overweight	32938.0	0.294766	0.455944	0.0	0.0	0.0	1.0	1.0
Were_obese	32938.0	0.072439	0.259218	0.0	0.0	0.0	0.0	1.0
Age	32938.0	15.048485	1.351078	11.0	14.0	15.0	16.0	18.0
bullied	32938.0	0.406370	0.491163	0.0	0.0	0.0	1.0	1.0

Exploratory Data Analysis



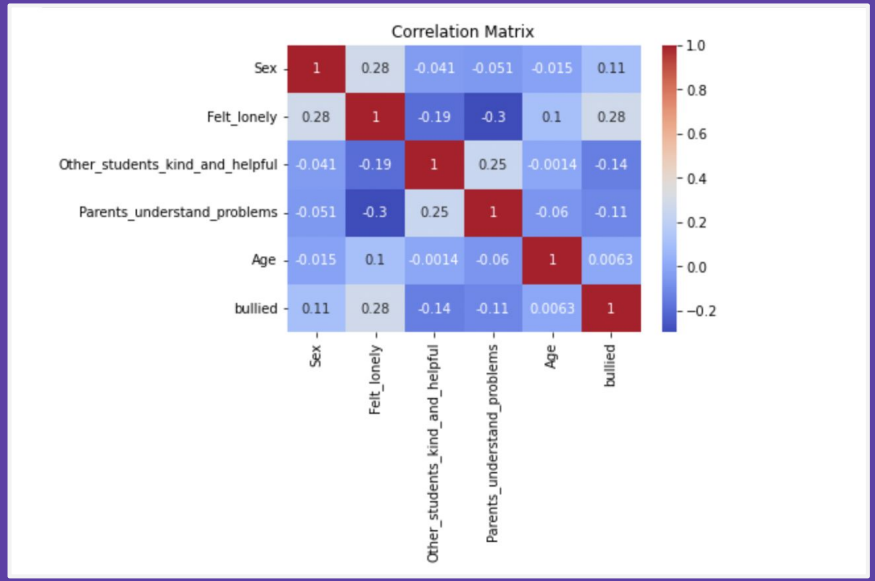




- **40%** of the studied students were bullied
- Females were more likely to be bullied than males
- ⅔ of the students who felt lonely most of the time also experienced bullying

Exploratory Data Analysis

- Age is not a good predictor for getting bullied
- Felt_lonely is the best predictor compared to the other features
- No feature has a high correlation with getting bullied as each victim and bullying case are unique



Most Correlated Features With Bullying	Correlation Value
Felt_lonely	0.28
Other_students_kind_and_helpful	-0.14
Sex	0.11

Scikit-Learn Modeling

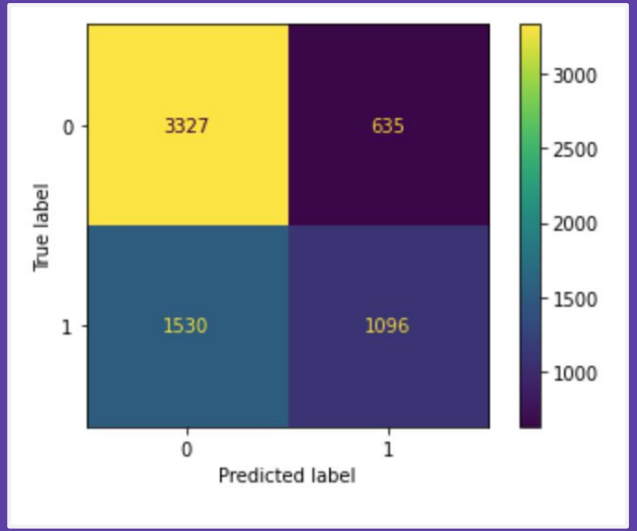
Models evaluated:

1. **Logistic Regression**
2. **Decision Tree**
3. Support Vector Machine
4. Random Forest Model
5. XGBoost Classifier

Target	Model	Accuracy	AUC
bullied	LogisticRegression	0.671	0.629
	DecisionTreeClassifier	0.625	0.593
	SVC	0.667	0.614
	RandomForestClassifier	0.631	0.606
	XGBClassifier	0.658	0.619

Bullying Predictions evaluated:

1. Bullying at school
2. Bullying outside school
3. Cyber bullying
4. Any bullying



Logistic Regression from Scratch

- Line profiler identified gradient descent as taking up the most time (by a significant margin)
- Limited impact from normalizing inputs and threading
- AUC = 0.618, Accuracy = 66.8% for fastest model

```
                # Gradient descent
 0.1            for i in range(self.num_iter):
15.3                z = np.dot(X, self.theta)
42.1                h = self.__sigmoid(z)
41.3                gradient = np.dot(X.T, (h - y)) / y.size
 1.1                self.theta -= self.lr * gradient
```

Numba + SGD improves speed by a factor of ~**300X**!

Model Variants	**Time (seconds)**
Standard Logistic Regression	50.81950
Stochastic Gradient Descent	1.82958
Normalized Inputs + SGD	1.87024
Threading + SGD	38.38546
Numba	41.81044
Numba + SGD (optimized)	**0.16999**

Decision Trees from Scratch

- Line profiler identified evaluate_split() function as taking up the most time (96.5%)
- Avoiding calculating info gain for all possible results in ~5X speed improvement
- Threading did not improve results
- AUC = 0.593, Accuracy = 64.2% for fastest model

```
96.5            gain, left, right = evaluate_split(split, data)
 0.0            if gain > best_gain:
 0.0                best_gain = gain
 0.0                best_split = split
 0.0                best_left = left
 0.0                best_right = right
```

Using both a random subset of features and adding adjustable numbers of features and variables resulted in a **~76X speed improvement!**

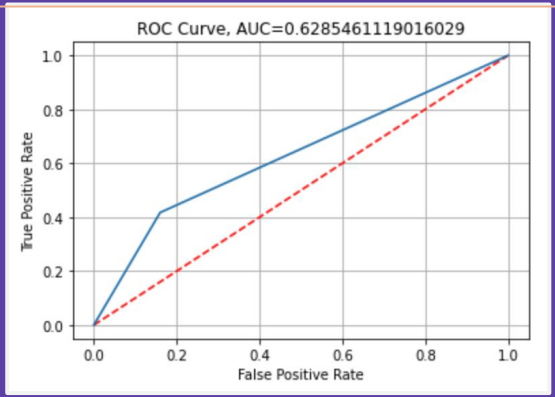
Model Variants	Time (seconds)
Standard Decision Tree	267.23535
Evaluate on Random Subset of Features	49.65542
Threading	370.37799
Threading + Random Subset	69.14372
Include Variables for Number of Features and Values	**3.51326**

Conclusions

1

Predicting bullying is hard!

- Top models **AUC ~0.63**



2

Making use of Numba + SGD improved our logistic regression from scratch model by **~300X**

3

Using random subsets of features + including adjustable numbers of features and variables improved speed by **~76X**

THANK YOU!