Updated Project Result and Analysis Submission
Title: Cyber Bullying Detection on Social Media Platforms Using Machine Learning

Course title: Machine Learning Laboratory
Course code: CSE-458
4th Year 2nd Semester Examination 2023

Date of Submission: 03.05.2024

Submitted to-

Sarnali Basak

Associate Professor
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka-1342

Submitted by-

SL	ID	Name
01	343	Tanzila Akter
02	361	Nurun Nahar Fiha
03	378	Md. Parvej Haque Palash
04	387	Md. Sakib Mollah
05	2142	Serajum Monira

Highlighting the Differences

Previous Task:

In previous task, we have worked in binary Classification for Cyber-bullying detection which is- **Bullying, Non-bullying**.

Table 1: Instances Description

Total instances	16074
Cyber Bullying instances	8488
Non Cyber Bullying instances	7585

We used 4 machine learning Algoritms such as-
- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

The Metrics used for determining the performance of models are given below-

Table 2: Supervised Traditional Method

	Decision Tree	Logistic Regression	Support Vector Machine
Accuracy	0.736547434	0.81150855	0.80808709
Precision	0.752954418	0.79290853	0.79053011
Recall	0.761168385	0.88373425	0.87972509

Table 3: Supervised Ensemble Method

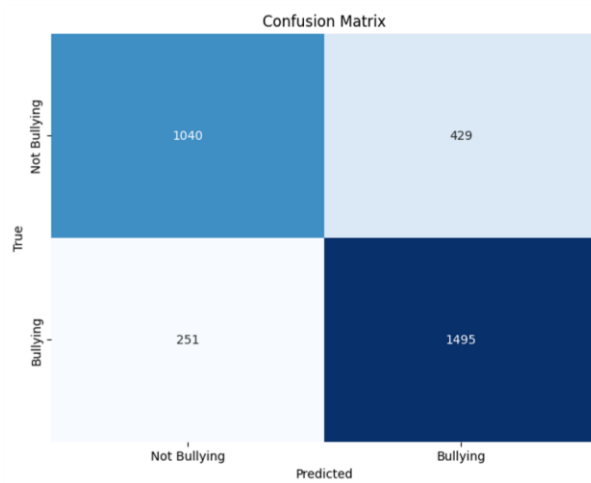|	Random Forest
Accuracy	0.792982549
Precision	0.792982549
Recall	0.791912908

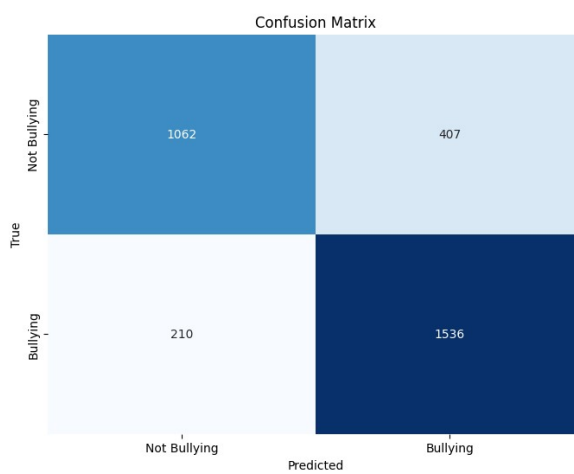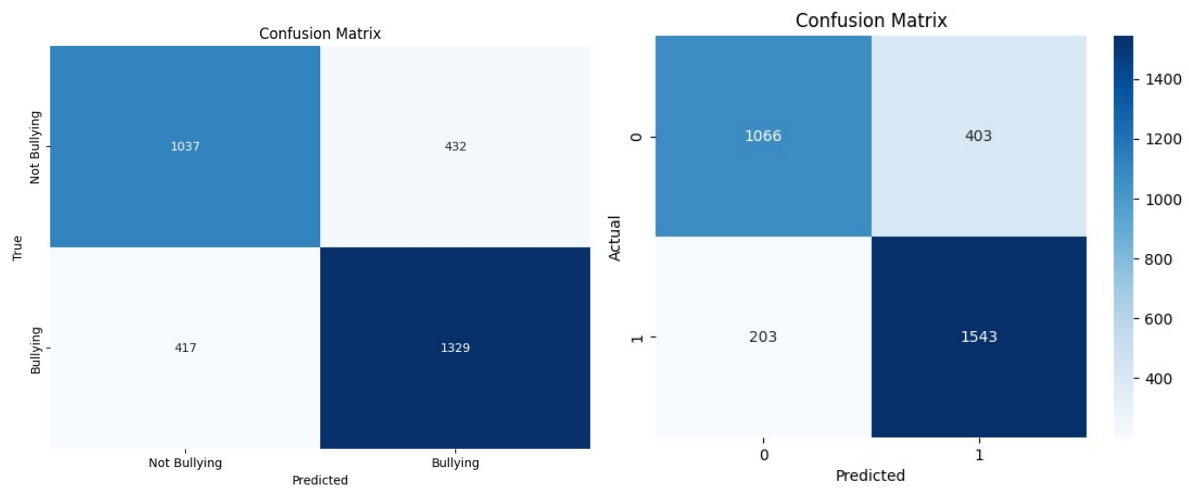Fig 1: Confusion Matrix of Logistic Regression Fig 2: Confusion Matrix of Decision Tree

Fig 3: Confusion Matrix of Support Vector Machine   Fig 4: Confusion Matrix of Random Forest
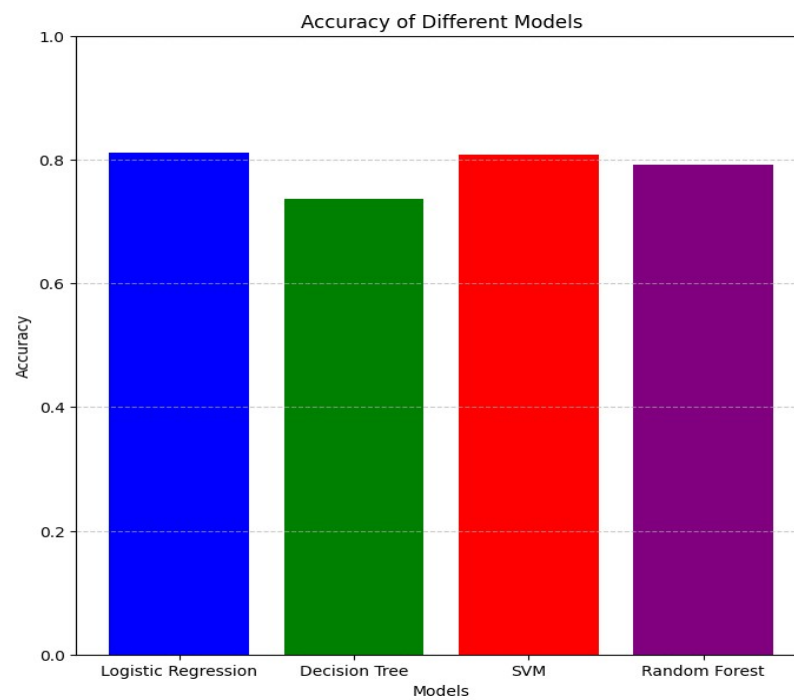

Fig 5: Accuracy

We have taken cv = 10 for cross validation.

- **Logistic Regression:**

Cross-validation scores: [0.8028607, 0.81281095, 0.80721393, 0.81331674, 0.81331674, 0.81144991, 0.80460485, 0.8226509, 0.80211574, 0.81829496]

Mean cross-validation score: 0.8108635416569919

- **Decision tree:**
  Cross-validation scores: [0.75186567, 0.75808458, 0.75248756, 0.75233354, 0.74113255, 0.72806472, 0.73242066, 0.75482265, 0.72868699, 0.7423771]
  Mean cross-validation score: 0.744227601878597

- **SVM:**
  Cross-validation scores: [0.79850746, 0.81654229, 0.81218905, 0.82389546, 0.80958307, 0.81953951, 0.8089608, 0.81953951, 0.81082763, 0.81953951]
  Mean cross-validation score: 0.8139124306903565

- **Random Forest:**
  Cross-validation scores: [0.78420398, 0.8090796, 0.79042289, 0.80647169, 0.80087119, 0.79713752, 0.79838208, 0.80273802, 0.78282514, 0.80149347]

  Mean cross-validation score: 0.797362557158204

SVM has the highest mean cross-validation score (0.8139), followed closely by Logistic Regression (0.8109), Random Forest (0.7974), and Decision Tree (0.7442). SVM and Logistic Regression show relatively consistent performance across folds, with smaller variability compared to Decision Tree and Random Forest. Decision Tree has the highest variability among the models, indicating less stable performance across folds.

Updated Task

Now, we have detected cyber-bullying into 3 classes – **Non-bullying, Bullying, Crime.** From the dataset, we categorized comments labeled as religious or threat or hate as Crime class. The remaining categories such as vulgar, troll, and insult were grouped as Bullying. Anything left was categorized as Non-Bullying. In our processed dataset we add a new category named class and marked them 0 for non-bullying, 1 for bullying and 2 for crime. This allowed us to train our models to detect cyber-bullying effectively within these three classes.

Total instances	16073
Cyber Bullying instances	4007
Crime instances	4481
Non Cyber Bullying instances	7585

Table 1: Instances Description

We used 4 machine learning Algoritms such as-
- Logistic Regression
- Decision Tree
- Random Forest
- **Extreme Gradient Boosting (XGBoost)**

The Metrics used for determining the performance of models are given below-

	Decision Tree	Logistic Regression
Accuracy	0.6186625194	0.6846034214
Precision	0.7157258064	0.7596153846
Recall	0.7249829816	0.8066712049
F1 Score	0.7203246533	0.7824364476

Table 3: Supervised Traditional Method

	Random Forest	XGBoost
Accuracy	0.6594090202	0.7539657853
Precision	0.7323135755	0.7435928883
Recall	0.7821647379	0.7190216932
F1 Score	0.7564186965	0.7264522034

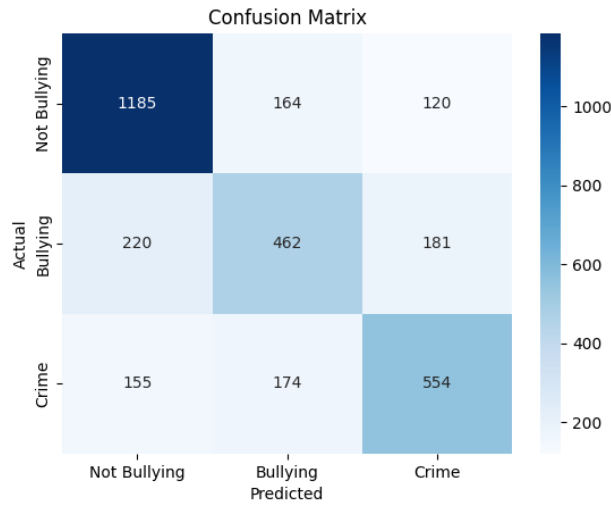Table 4: Supervised Ensemble Method

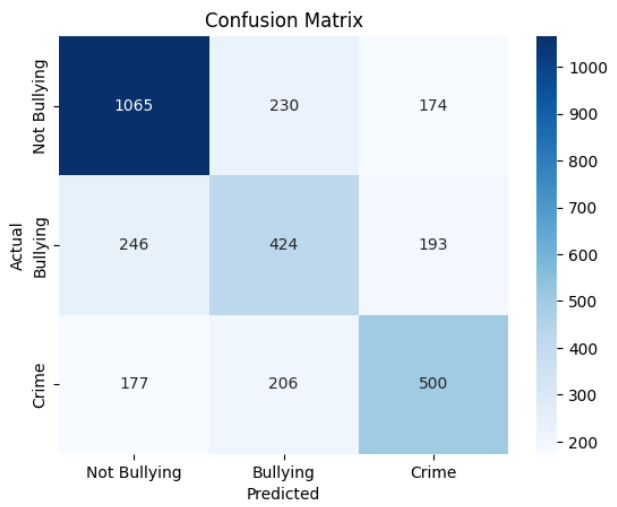Fig 1: Confusion Matrix of Logistic Regression
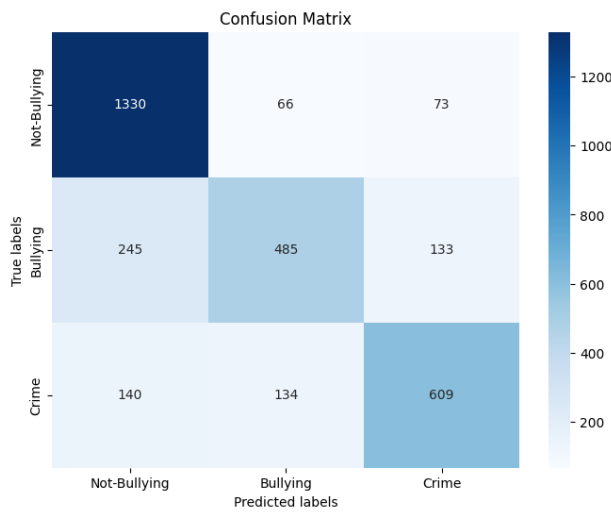

Fig 2: Confusion Matrix of Decision Tree

Fig 3: Confusion Matrix of XGBoost

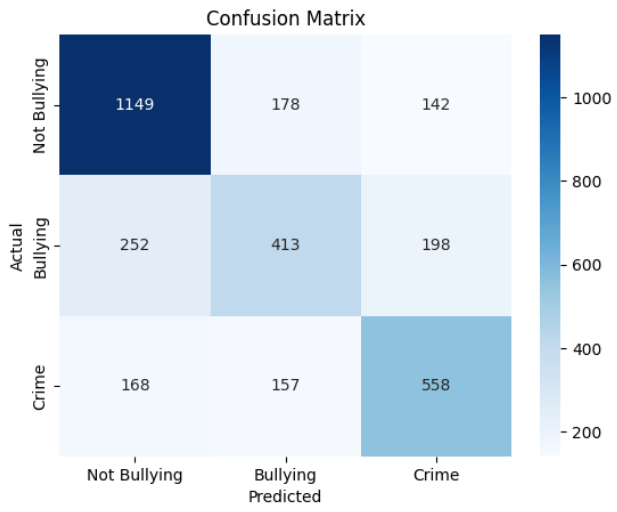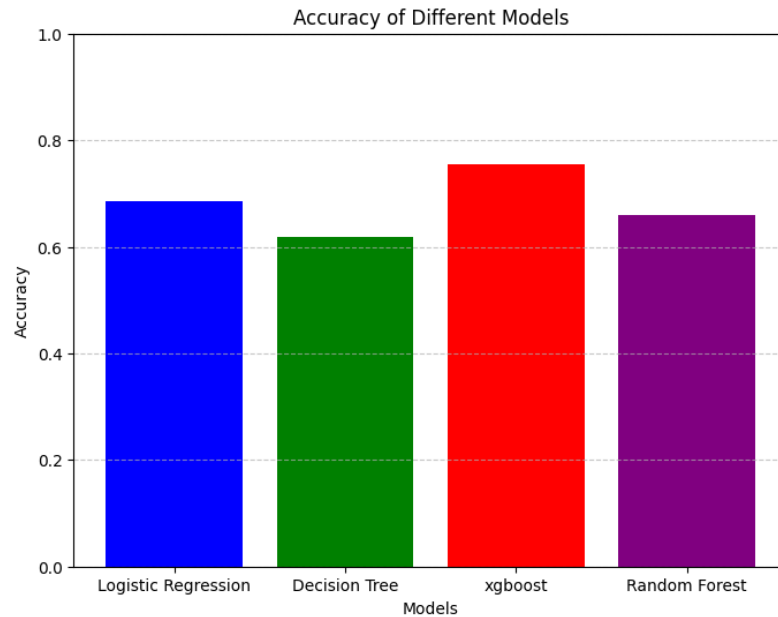Fig 4: Confusion Matrix of Random Forest

Accuracy of Different Models

We chose CV = 10 for cross-validation to ensure robust evaluation of the models.

• Logistic Regression:

Cross-validation scores: [0.68159204, 0.69216418, 0.70273632, 0.69197262, 0.69570629, 0.69570629, 0.68512757, 0.69446173, 0.6925949, 0.69570629]

Mean cross-validation score: 0.6927768206261784

• Decision tree:

Cross-validation scores: [0.6050995, 0.62562189, 0.62686567, 0.62538892, 0.61169882, 0.62725576, 0.61543248, 0.63410081,0.60236465, 0.62352209]

Mean cross-validation score: 0.619735059921302

• Random Forest:

Cross-validation scores: [0.66604478, 0.67412935, 0.68594527, 0.69632856, 0.66334785, 0.6813939, 0.66645924, 0.68388301, 0.65899191, 0.68388301]

Mean cross-validation score: 0.6760406895206604

• Extreme Gradient Boosting (XGBoost):

 Cross-validation scores: [0.76360809, 0.7651633, 0.77138414, 0.74727838, 0.75583204, 0.76205288, 0.76360809, 0.74416796, 0.76575875, 0.7688716]

Mean cross-validation score: 0.7607725218001706

Extreme Gradient Boosting (XGBoost) has the highest mean cross-validation score (0.76077), followed closely by Logistic Regression (0.6927), Random Forest (0.6760), and Decision Tree (0.6197). XGBoost performs the best across most metrics, with Logistic Regression being a close competitor, followed by Random Forest. Decision Tree seems to be the weakest performer among the four models.