



uOttawa

Faculté de génie  
Faculty of Engineering

Name: Nada Nagah Ahmed Awaad Heider

ID: 300273135

## **Solution:**

First I trained the model on training dataset using 2 different models the first model is random forest and the second one is gradient boosting classifier.

Then build features from the domain

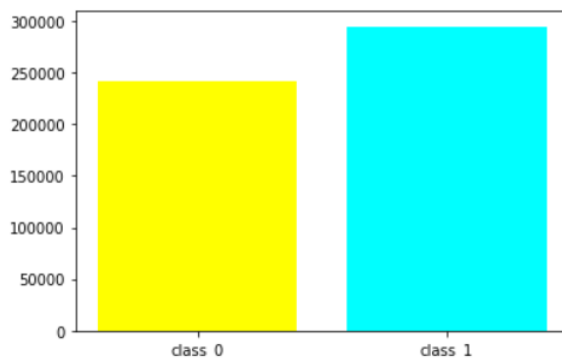
The random forest accuracy is higher than Gradient boosting so I saved the random forest and use it in prediction phase.

Then use the Kafka to ingest data in the model and generate features and convert it to data frame and add predicted label feature and score feature.

## **Dataset preprocessing:**

The training dataset consists of 536139 records. 241785 samples belong to class\_0 and 294353 samples belong to class 1.

And this figure illustrates the percentage between class 0 and class 1.



The data is split into training 70% and testing 30%. Then I dropped 3 features from original dataset `'timestamp', 'longest_word', 'sld'`.

## **Algorithms:**

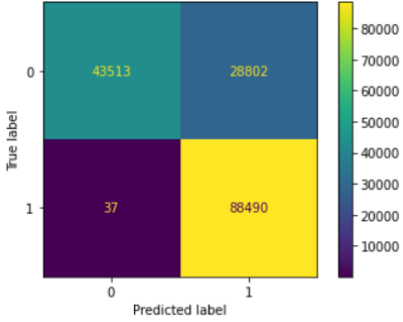
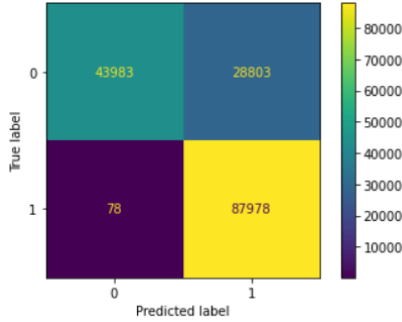
I apply 2 different models to the dataset: RandomForestClassifier and GradientBoostingClassifier.

The RandomForestClassifier is type of trees fit the number of decision tree classifier on sub sample of data and try to reduce the over-fitting and also to enhance the accuracy.

The GradientBoostingClassifier is binary classifier allow optimization in loss function.

## **Experiments:**

	RandomForestClassifier	GradientBoostingClassifier
F1_score=	85.98%	85.9%

Confusion matrix		
Number of misclassification	28839	28881

Because the random-forest f1\_Score is higher than another model so I selected the RandomForestClassifier and save it in Pickle file to use it in the prediction phase.

## Refernces:

- [1]<https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- [2]<https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>