Classification of Illegal Fishing

Name:	**Shubham Meena**
Registration No./Roll No.:	20263
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	January 12, 2023
Date of Submission:	April 16, 2023

1 Introduction

The objective is to develop supervised machine learning frameworks to identify illegal fishing. And it is multiclass classification problem. First we see Exploratory data analysis(EDA) and then at last we will use some classification models on the given training dataset and after that we will choose best model for prediction of test data classes.

1.1 Overview of dataset

The given training dataset has 838860 rows $\times$ 8 columns. And it has 3 class labels which are -1,0,1 which tells that for -1 there is no class label, 0 not fishing, and 1 there is fishing. All values are numerical. And it has 8 features i.e: mmsi, timestamp, distance from shore, distance from port, speed, course, lat, lon.

2 Methods

- Data cleaning and Pre-processing
- Exploratory data analysis(EDA)
- Training on models

2.1 Data cleaning and Pre-processing

First we check for any null/missing values in which case i had 3 null values so after that i dropped that 3 null values rows.Then, i start with checking for imbalanced data using the value counts function and countplots and i observe that the training data is imbalanced in which case -1 class was dominating other classes with huge ratio.

2.2 Exploratory data analysis(EDA)

I perform EDA on the merged dataset in which i merge the class label with training data to get insights of the data and its features. I start with heatmap to see correlation between all features and class. and the plot the histogram for features to see distribution of each feature. And then at last i plot the pair plot to see correlation.

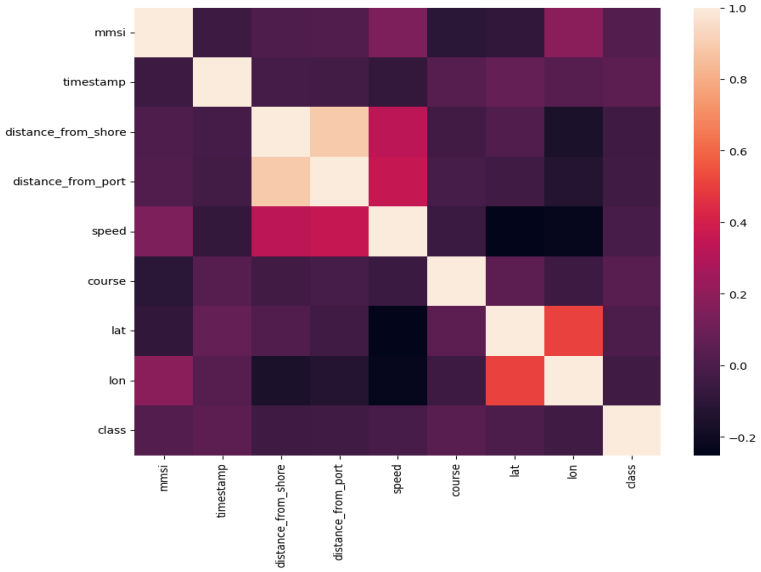

Figure 1: Correlation Heatmap of features

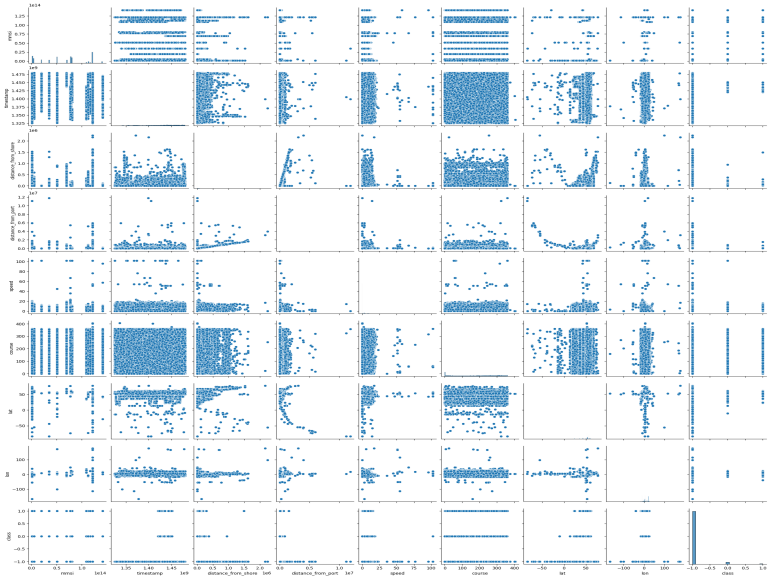

Figure 2: Pair plot

2.3 Training on models

I am going to try different models for classification of training dataset. In which i used KNeighbors Classifier, Decision Tree Classifier, Random Forest Classifier.
For hyperparameter tuning i am using GridSearchCV to try out different parameters and pick the one that gives the best score when using k-fold cross validation. It trains classifiers with all the specific values for different parameters and gives the best performance values for those parameters.

3 Evaluation Criteria

Precision is the ratio of the number of true positives to the total number of true positives and false positives,
Precision = True Positives/True Positives + False Positives

Recall on the other hand is the ratio of the number of True positives to the total number of true

positives and false negatives,
Recall = True Positives/True Positives + False Negatives

F-Measure combines both of these as it is the Harmonic mean of precision and recall,
FMeasure = 2×Precision×Recall/Precision + Recall

In my case i mainly focused on precision and recall values because i was getting high accuracy due to imbalanced data so i choose my best model on the basis of accuracy as well as precision and recall.

4 Analysis of Results

I finally get the results of the different classifiers with hyperparameter tuning on the training dataset by using different parameters. I have applied three classifiers KNN , Decision tree and Random forest.

I used the F1 score, Precision, and Recall of each classifier to evaluate how well it performed on the training set of data. I select the best model for the project based on these three criteria.We can see the scores in Table 1.

Table 1: Performance Of Different Classifiers

Classifier	Precision	Recall	F-measure
KNN	0.96	0.96	0.96
DT	0.93	0.92	0.93
RFC	0.95	0.92	0.93

5 Discussions and Conclusion

KNeighbors Classifier (KNN) is so far the best performing classifier on my dataset.

In my dataset there are 802828 unlabelled datapoints out of 838860. So, due this imbalanced dataset it was causing problems like high biasness. And only 36032 are labelled out of them 30237 showing not fishing and only 5795 showing fishing.Due to which i got high accuracy and other metrics also.
So, to overcome this we can use sampling.

And my feature values are also big in numbers due to which my device was taking too much computational time .
So, to overcome this issue we can do scaling.

6 Github Link

Link

7 References

- Tanmay sir class notes
- https://scikit-learn.org/stable/
- Tutorials of machine learning on youtube