Classification of Illegal Fishing

Name:	**Sumit Kumar**
Registration No./Roll No.:	2311006
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

1 Introduction

Illegal fishing is a severe global problem that endangers marine ecosystems and threatens the way of life for many people who rely on the waters across the world. This issue, which is characterised by a range of unlawful actions such as overfishing, unreported catches, and unauthorised access into restricted regions, has a detrimental impact on the environment, the economy, and society. Illegal fishing jeopardises the balance of maritime ecosystems, driving species to extinction and weakening the long-term viability of our oceans at a time when marine resources are still limited.

The Automatic Identification System (AIS) vessel tracks that serve as the foundation of our project are an essential resource for studying and understanding ship behaviour at sea. The AIS data contains critical information on the locations, movements, and other pertinent properties of vessels that may be utilised to identify trends associated with illicit fishing activities. Using machine learning, we want to create a system that can automatically recognise and categorise instances of unlawful fishing. This will aid in the protection of our seas and the enforcement of fishing regulations. [1]

Data Description

The training dataset is a comprehensive collection of nautical information, consisting of **838,860 rows** and **8 columns**, with each row representing a separate observation. This dataset is a significant resource for training machine learning models to detect and categorise illicit fishing activity. The data includes a variety of numerical variables as well as three separate class labels that assist to categorise the observed vessel operations. These classification labels are as follows:

- **-1 (No Class Label):** Instances marked with this label indicate the absence of any specific class assignment. These data points do not pertain to either fishing or non-fishing activities.
- **0 (Not Fishing):** Vessels categorized as "0" are indicative of non-fishing activities. This label signifies that the observed vessels are not currently engaged in fishing operations.
- **1 (Fishing):** Instances labeled as "1" correspond to vessels actively involved in fishing activities. This class identifies and differentiates fishing-related behaviors from other maritime actions.

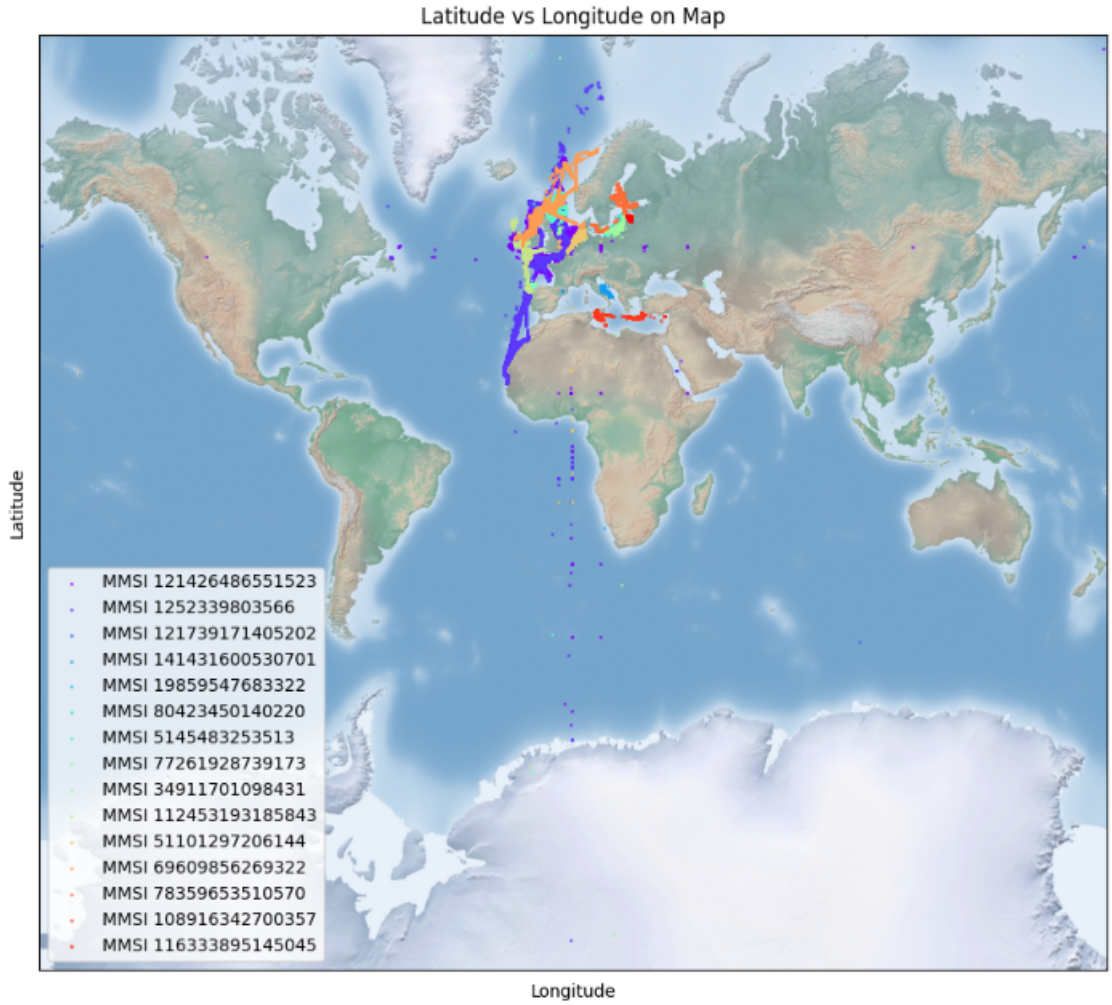

Figure 1: Location of different vessels

2 Methods

2.1 Data Preprocessing

There were six missing values in the training dataset, three in each of the **speed** and **course** features. To deal with the missing numbers, the mean approach was used. The mean method is a fundamental imputation approach that includes replacing missing values in a dataset with the mean value of the remaining data points in the same feature. This strategy is often used when dealing with numerical data and is one of the easiest ways to resolve missing data while keeping the dataset's general statistical properties.

2.2 Exploratory Data Analysis (EDA)

To gain a comprehensive understanding of the dataset and its features, an exploratory data analysis (EDA) was conducted.

Class	Number of Instances
-1	802828
0	30237
1	5795

This involved a meticulous examination of the dataset created by merging the training data with the corresponding class labels. EDA aids in uncovering insights, patterns, and potential correlations

between features, setting the stage for more informed decision-making in subsequent modeling steps. [2]

2.3 Model Training

Three classification models, namely Random Forest, Decision Tree, and AdaBoost, were trained and applied on datasets to address a three-class classification problem. The dataset comprises instances belonging to one of three distinct classes. For model training, I employed grid search cross-validation (GridSearchCV) to systematically explore a range of hyperparameter combinations, optimizing the model's performance. Additionally, I incorporated Stratified K-Fold cross-validation (StratifiedKFold) to ensure a balanced distribution of classes in both training and testing sets. Find my code here.

3 Evaluation Criteria

Evaluation criteria like precision, recall, accuracy and F1 score were used to evaluate the performance of the models. These metrics provide a comprehensive understanding of the model and the ability to correctly identify fishing, not fishing, and no-label cases, accounting for both false positives and false negatives.

4 Results and Analysis

4.1 Results

Table 1: Results for different classifiers on various datasets

Dataset Description	Evaluation Criteria	Decision Tree Classifier	Random Forest Classifier	Adaboost Classifier
Dataset 1(No. of Attributes = 4, No. of Classes = 2)	Precison	0.8982	0.8939	0.9096
	Recall	0.8579	0.9321	0.8682
	F1-Score	0.8763	**0.9115**	0.8871
Dataset 2(No. of Attributes = 5, No. of Classes = 2)	Precison	0.9646	0.972	0.9539
	Recall	0.9604	0.9673	0.9517
	F1-Score	0.9625	**0.9688**	0.9528
Dataset 3(No. of Attributes =6, No. of Classes = 2)	Precison	0.9569	0.9663	0.9545
	Recall	0.9645	0.9702	0.9596
	F1-Score	0.9607	**0.9682**	0.957
Dataset 4(No. of Attributes = 4, No. of Classes = 3)	Precison	0.6524	0.4668	0.6033
	Recall	0.5128	0.6645	0.5268
	F1-Score	**0.5616**	0.4954	0.5547
Dataset 5(No. of Attributes = 5, No. of Classes = 3)	Precison	0.9465	0.9453	0.9462
	Recall	0.9489	0.9444	0.9439
	F1-Score	**0.9477**	0.9448	0.945
Dataset 6(No. of Attributes = 6, No. of Classes = 3)	Precison	0.9375	0.9327	0.9411
	Recall	0.9354	0.9301	0.9371
	F1-Score	0.9364	0.9294	**0.9391**

4.2 Analysis

- 3-class classification problem was converted to a 2-class classification problem by dropping one class, namely, -1, because this class label indicated the absence of any class assignment so had very little or no significance.
- Random Forest exhibits outstanding performance on Dataset 2, which has 5 features and 2 classes. It attains impressive precision, recall, F1-score, and accuracy, showcasing a balanced performance across these metrics. The model effectively identifies and classifies instances from both classes without bias. The optimal parameters for training the model include using the criterion 'entropy', setting the maximum depth to 50, employing a minimum samples leaf of 1, specifying a minimum samples split of 2, utilizing 100 estimators, and fixing the random state at 0. This configuration underscores the model's ability to capture intricate patterns within the dataset, leading to its exceptional performance in classification tasks with a binary class distribution.

- As the test data has data for all three classes, the model trained with only two classes will probably fail to classify all the data points in test data. So , the models were again trained for three class classification and the test data was classified using these trained models.
- Decision Tree excels on Dataset 5, characterized by 5 features and 3 classes, as it attains remarkable precision, recall, F1-score, and accuracy. The model's robust performance across these metrics indicates its ability to effectively distinguish and categorize instances among the three classes without showing bias towards any particular class. The optimal configuration for training this model includes using the criterion 'entropy', setting the maximum depth to 30, employing a minimum samples leaf of 1, specifying a minimum samples split of 2, and fixing the random state at 0. This configuration demonstrates the model's ability to capture complex relationships within the dataset, leading to its exceptional performance on classification tasks with multiple classes and diverse feature sets.
- It is observed that selecting 5 features is giving the best results in terms of all evaluation criteria, irrespective of the number of classes. This highlights the importance of feature selection in enhancing model performance.

5 Conclusion

In essence, the study demonstrates the efficacy of machine learning models, particularly Random Forest and Decision Tree, in automatically identifying and categorizing instances of illegal fishing. The findings contribute valuable insights into addressing the global challenge of illegal fishing, emphasizing the potential of technology-driven solutions in safeguarding marine ecosystems and promoting sustainable practices in the fishing industry.

References

[1] Abhimanyu Thakur, Anju Dhiman, N. Thakur, Hamid Hamid, Monika Chauhan, and Sunakshi Gautam. An introduction to seafood and recent advances in the processing of seafood products. 10:169–180, 06 2019.

[2] Global Fishing Watch. Revealing the supply chain at sea: A global analysis of transshipment and bunker vessels. *Baltimore: Global Fishing Watch*, 2021.