Gievn Title of The Project

Name:	**Vishwanaathan B L**
Registration No./Roll No.:	20309
Institute/University Name:	IISER Bhopal
Program/Stream:	e.g., EECS
Problem Release date:	January 12, 2023
Date of Submission:	April 16, 2023

1 Introduction

Illegal, unreported and unregulated (IUU) fishing has become a significant problem for the global fishing industry. It not only harms the marine ecosystem but also threatens the livelihoods of millions of people who depend on fishing for their livelihoods. The increasing use of Automatic Identification System (AIS) by fishing vessels provides an opportunity to detect and prevent IUU fishing activities. This report presents a classification model for illegal fishing using AIS data.

The dataset used in this study contains AIS data from fishing vessels operating in a particular region. The dataset has 3 class labels, **0 for not illegal fishing, 1 for illegal fishing, and -1 for no class label**. The features of the dataset include **distance from shore, distance from port, mmsi, time stamp, speed, course, longitude, and latitude**. The dataset is a mix of categorical and continuous data, and it contains few missing values, which require preprocessing before modeling.

The plan of action involves few steps, starting with data cleaning and preprocessing. Then, several classification models are evaluated, including logistic regression, decision trees, KNN and random forests. The performance of each model is assessed using various metrics such as accuracy, precision, recall, and F1-score. The results of the study provide insights into the characteristics of the data and the effectiveness of different classification models for detecting illegal fishing activities.

In conclusion, this report presents a classification model for illegal fishing using AIS data. The dataset used in this study contains various features that can help detect illegal fishing activities. The performance of several classification models is evaluated, and the results provide insights into the characteristics of the data and the effectiveness of different models. The findings of this study can be used to develop effective strategies for detecting and preventing illegal fishing activities. The below bar plot contains the ratio of class labels in the dataset as we described above, Figure 1.

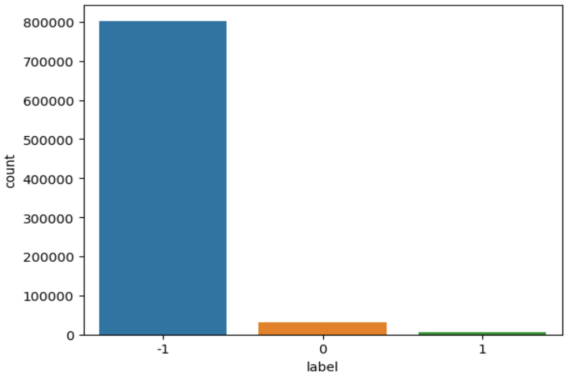
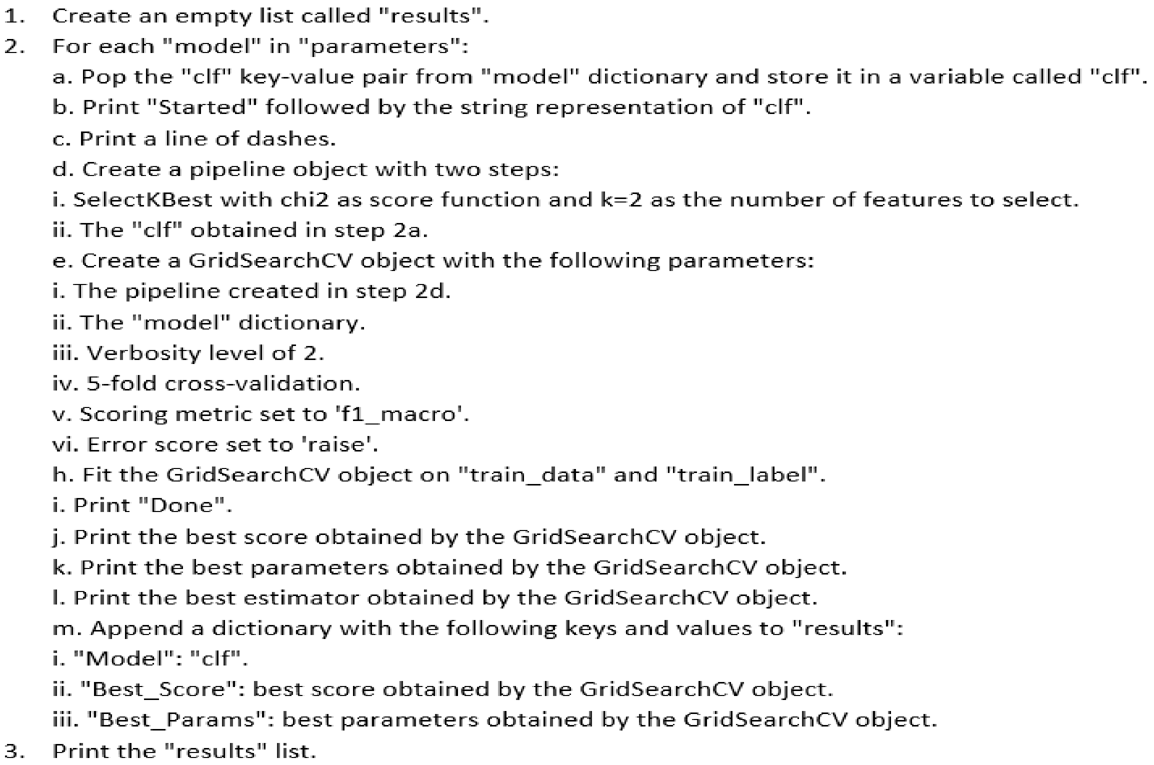


Figure 1: Amount of class -labels in given dataset

2 Methods

This project explored the use of sklearn ML models like KNNs, SVMs, Decision Trees and Random Forest via a pipeline based on the 'Divide and Conquer Algorithm' juxtaposed to the generally used pipeline which is defined to run everything from the pre-processing to hyperparameter tuning to the evaluation and prediction at one go. The grid search was ran with the parameters which was spearately specified in the function: **parameters**.As the grid search won't accept negative values we **change the class label -1 to 2** which also has been mentioned as 2 in Fig.1. The pseudo-code of the whole grid search is given below



```
1.  Create an empty list called "results".
2.  For each "model" in "parameters":
    a. Pop the "clf" key-value pair from "model" dictionary and store it in a variable called "clf".
    b. Print "Started" followed by the string representation of "clf".
    c. Print a line of dashes.
    d. Create a pipeline object with two steps:
    i. SelectKBest with chi2 as score function and k=2 as the number of features to select.
    ii. The "clf" obtained in step 2a.
    e. Create a GridSearchCV object with the following parameters:
    i. The pipeline created in step 2d.
    ii. The "model" dictionary.
    iii. Verbosity level of 2.
    iv. 5-fold cross-validation.
    v. Scoring metric set to 'f1_macro'.
    vi. Error score set to 'raise'.
    h. Fit the GridSearchCV object on "train_data" and "train_label".
    i. Print "Done".
    j. Print the best score obtained by the GridSearchCV object.
    k. Print the best parameters obtained by the GridSearchCV object.
    l. Print the best estimator obtained by the GridSearchCV object.
    m. Append a dictionary with the following keys and values to "results":
    i. "Model": "clf".
    ii. "Best_Score": best score obtained by the GridSearchCV object.
    iii. "Best_Params": best parameters obtained by the GridSearchCV object.
3.  Print the "results" list.
```

Figure 2: Pseudo Code of the Grid Search

We split the whole data into sets of train, valid and test in ratio of (0.6:0.2:0.2). We perform cross validation in folds of 5 with train data and train label with a list of parameters defined in separate function.This means that the training data will be split into 5 equal parts, and the model will be trained and evaluated 5 times, with each of the 5 parts used as the validation set once and the remaining 4 parts used for training.

The cross-validation score will be calculated as the average score across these 5 folds. This is done to get an estimate of the model's performance on unseen data and to ensure that the model is not overfitting the training data. Which reduced the overall time complexity and made flexible to identify the best model in a single run. We run the ML models before and after removing -1 class label as it mentioned earlier it depicts the unlabeled part of the data.

.

3 Evaluation Criteria

During each grid search run in each model we average out all the F measure value and fitting the best model at the end with outputting the best parameters for each model. It also selects the model with best average score as the final model. We get the best score by fitting the train data and test data in the grid model which ends up having the best parameters

The below table has the models and its corresponding score with and without -1 label

Classifier	Best-score with -1	Best-score without-1
LR	0.023	0.456
RFC	0.958	0.867
GaussianNB	0.326	0.456
SVM	0.225	0.468
KNN	0.958	0.864
DTC	0.957	0.858

Table 1: Grid search results for each model.

As from the table above we can observe that the best results in both cases are obtained from Random Forest Classifiers, K-Nearest Neighbors and Distance Tree Classifier.

4 Analysis of Results

As mentioned in previous section, now lets verify the precision and recall in selected models for both the cases with test data and best parameters from the result of grid search.

Classifier	Label	Precision	Recall	F1-score	Accuracy of the model
RFC	-1	1	1	1	0.9968
	0	0.97	0.98	0.97	
	1	0.88	0.88	0.88	
KNN	-1	1	1	1	0.9976
	0	0.98	0.98	0.98	
	1	0.92	0.91	0.92	
DTC	-1	1	1	1	0.9968
	0	0.97	0.98	0.97	
	1	0.88	0.88	0.88	

Table 2: Including -1 Class label.

Classifier	Label	Precision	Recall	F1-score	Accuracy of the model
RFC	0	0.98	0.98	0.98	0.964
	1	0.88	0.91	0.89	
KNN	0	0.98	0.98	0.98	0.963
	1	0.88	0.89	0.88	
DTC	0.	0.98	0.98	0.98	0.964
	1	0.88	0.91	0.89	

Table 3: Without including -1 Class label.

From the above tables we can clearly see that removing -1 class label decreases overall accuracy as it consisted the huge amount of data. Also we can see KNN in the table 2 has higher precision(0.92), recall(0.91) and F1-score(0.91) with overall accuracy of 0.9976 for the class label 1 which was found out to be in the lowest amount from the whole data set. Thus this result is much better compared to other evaluations. Thus there involves a certain trade off in accuracy while removing -1 class label.

5 Discussions and Conclusion

The purpose of this study was to explore the use of machine learning techniques for the classification of illegal fishing using AIS data. Illegal fishing is a significant problem that can result in economic and environmental damage. AIS data provides valuable information about the movements of vessels and can be used to detect potential illegal fishing activity. Machine learning can be used to analyze this data and develop models that can identify patterns and anomalies associated with illegal fishing.

The use of machine learning in this context has several benefits. First, it can help to automate the classification process, reducing the time and resources required for manual analysis. This can enable more rapid and efficient monitoring of fishing activity, which is particularly important in areas where illegal fishing is prevalent. Additionally, machine learning can help to improve the accuracy of classification models by identifying patterns that may not be immediately apparent to human analysts.

Overall, the application of machine learning in the classification of illegal fishing using AIS data has the potential to improve monitoring and enforcement efforts, thereby reducing the incidence of illegal fishing and its associated impacts. However, it is important to carefully evaluate the performance of these models and to ensure that they are not biased or subject to other limitations that could impact their effectiveness. Further research is needed to explore the potential of machine learning in this context and to develop more advanced models that can address the complex challenges associated with illegal fishing detection and prevention

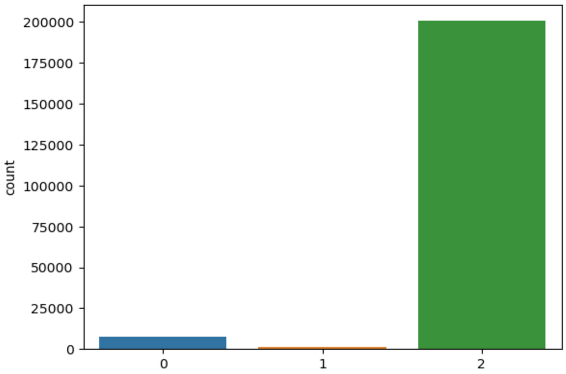


Figure 3: Labels of final test data without removing -1 (-1 is mentioned as 2 over here)

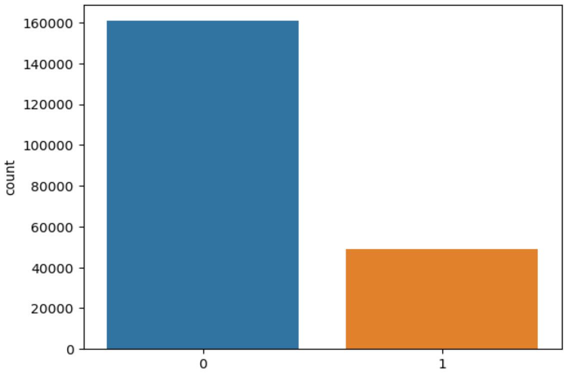


Figure 4: Labels of final test data with removing -1