Classification of Illegal Fishing.

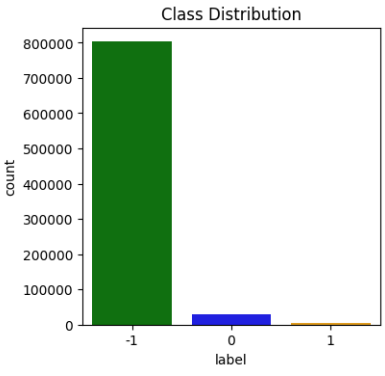
Name:	**Chaudhari Khushi Ganesh**
Registration No./Roll No.:	21084
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August, 2023
Date of Submission:	November 19, 2023

1 Introduction

Problem statement: The objective is to have a supervised machine learning frame work to classify illegal fishing.
Data description: The give dataset have 838860 rows and 8 columns as mmsi, timestamp, distance from shore, distance from port, speed, course, latitude, longitude. There are 3 classes as -1, 0, 1 as no class labels, not fishing and fishing respectively. The data is highly imbalanced with -1 having 802828 datapoints.
The following figure contains graphs plot for the distribution of distance from shore, distribution of distance from port, distance from shore vs distance from port, distribution of speed. Then there is a correlational matrix, bar graphs representing class distribution, distribution of mmsi unique values for longitude vs latitude, a subplot for the dataframe columns and the last is mutual information for feature selection. e.g., Figure 1.

2 Methods

The methods used in the project follows the plan of data analysis, data preprocessing, data balancing, using different models, feature selection, again using models, predicting labels of test data.
Data analysis: Here I got the data size and shape, data description, data information, information on missing values and nan values. The distribution of labels, also missing values in them.
Data preprocessing: It contains imputing missing and nan values, concatenating the labels and training data. Plotting the graphs, visualization of data.
Data balancing: As we have seen the data is very imbalance so I did data undersampling by reducing the majority classes to the size of minority class.



Models: I used Logistic regression, KNN, Naive bayes, Decision tree, Random forest classifier and SVM.
Feature selection: For feature selection I have the mutual information to understand the dependency of the features on the labels. The information I got is the labels depends on longitude, latitude and timestamp the most. So I decided to use these features and dropping the others. Then again using the models.
Models after feature selection: Logistic Regression, KNN, Naive bayes, Decision Tree, Random Forest, SVM.

Table 1: Performance Of Different Classifiers Using All Features

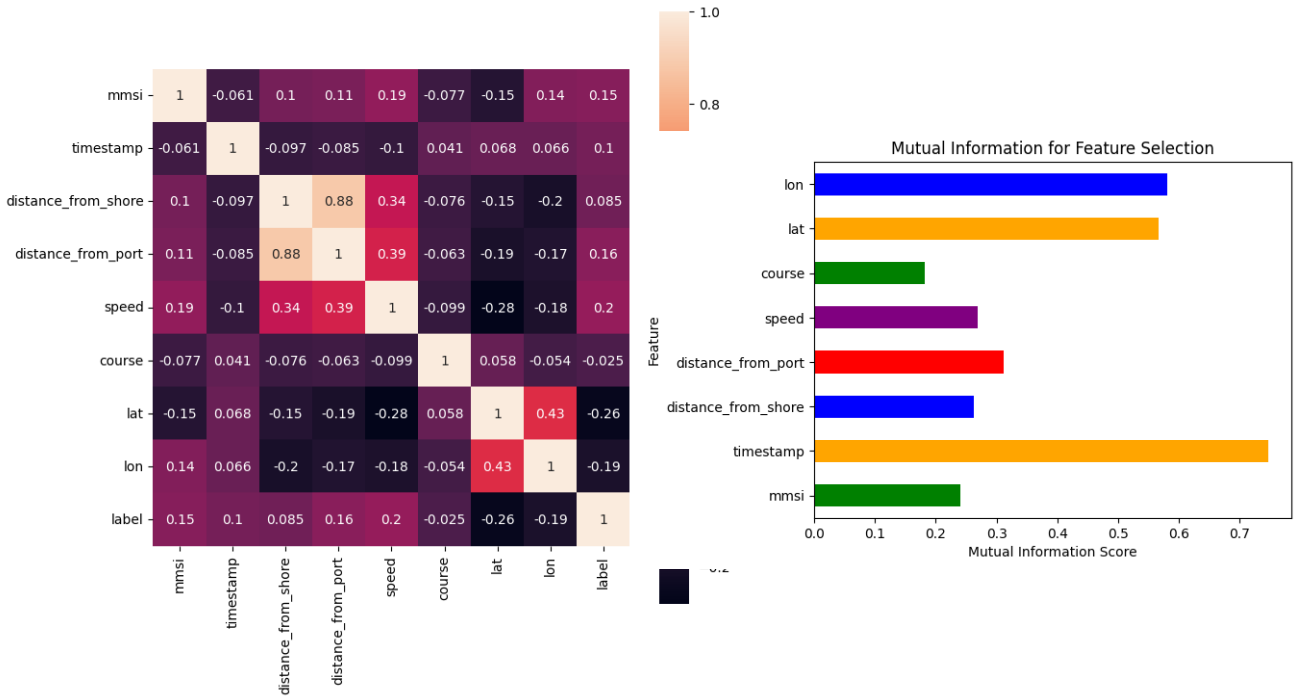
Classifier	Precision	Recall	F-measure
Logistic Regression	0.11	0.33	0.16
K-Nearest Neighbor	0.97	0.97	0.97
Naive bayes	0.38	0.39	0.34
Decision Tree	0.96	0.96	0.96
Random Forest	0.96	0.96	0.96
Support Vector Machine	0.36	0.40	0.36

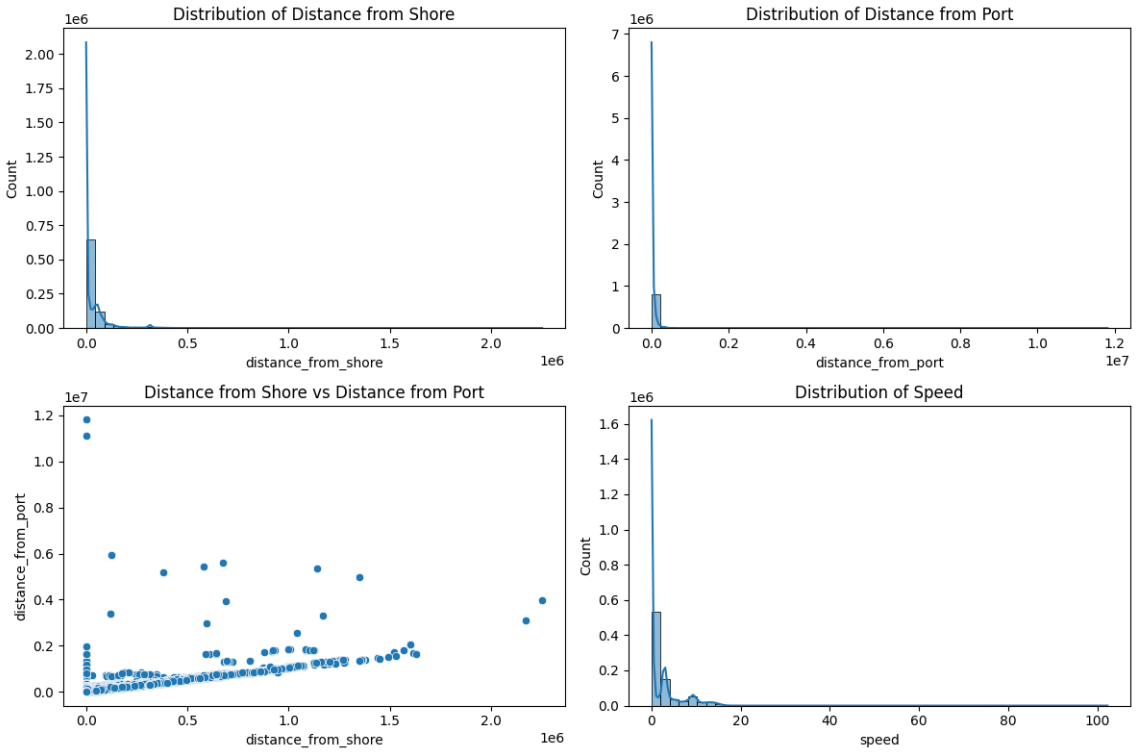
Table 2: Performance Of Different Classifiers after feature selection

Classifier	Precision	Recall	F-measure
Logistic Regression	0.11	0.33	0.16
K-Nearest Neighbor	0.84	0.84	0.84
Naive bayes	0.70	0.65	0.66
Decision Tree	0.97	0.97	0.97
Random Forest	0.97	0.97	0.97
Support Vector Machine	0.50	0.43	0.35

3 Experimental Setup

Different criteria used for evaluation are recall, precision and F- measure. The selection of model is done based upon the grid search result. Hyperparameters used are criteria, splitter, maximum depth, maximum sample split, maximum sample leaf and max features. There are lots of libraries used. Some of the general libraries are numpy, pandas, matplotlib, scikit-learn, seaborn, cartopy.





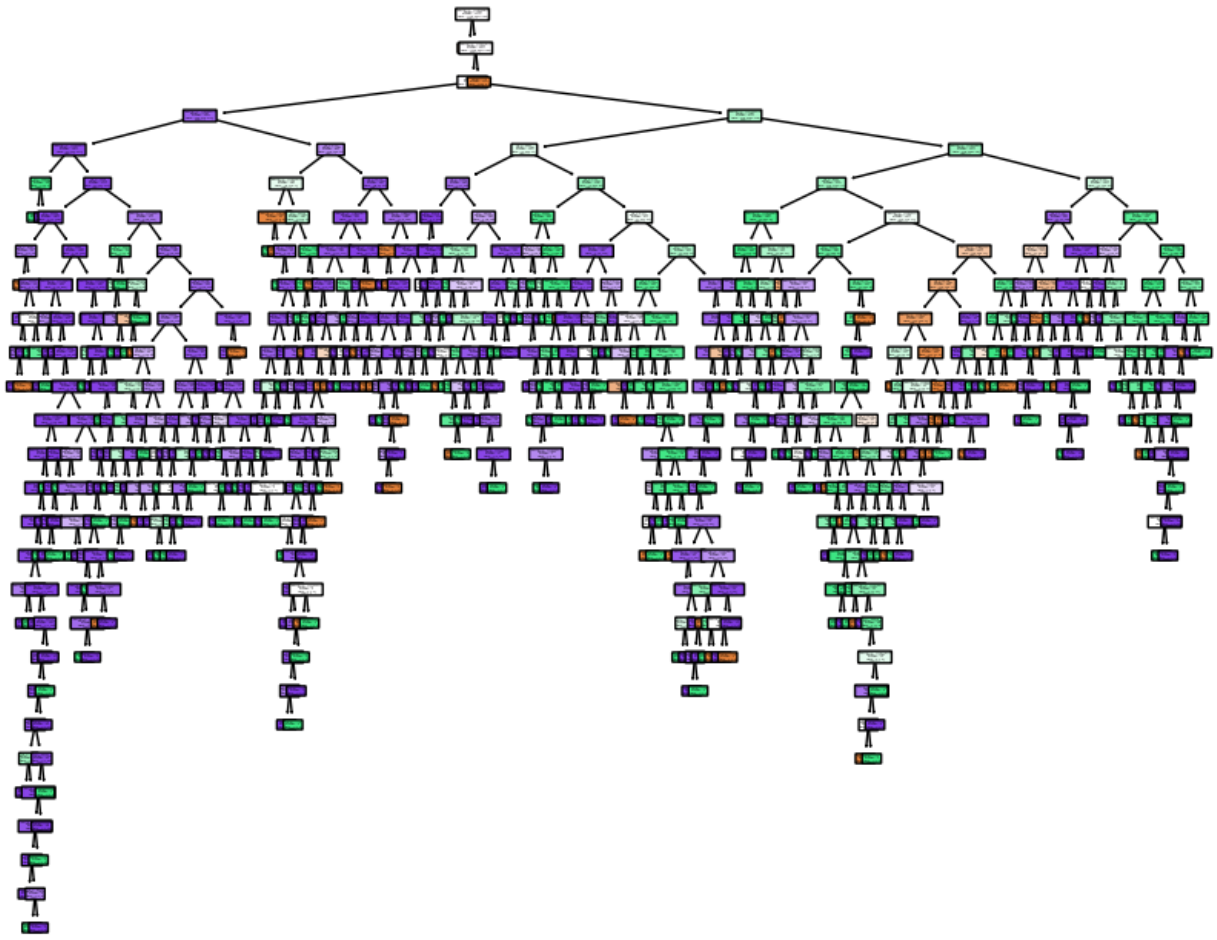


Figure 1: Decision tree

4 Results and Discussion

The accuracy, recall, precision and F-Measure of different models is different also it is based upon feature selection. So from the given tables we can infer that the best score was given by KNN before feature selection. After feature selection the best score is given by Decision tree and Random forest classifier. For reference here are the confusion matrix of different models after feature selection.

Confusion Matrix of SVM

	Predicted Class		
Actual Class	-1	0	1
-1	252	293	377
0	0	26	935
1	0	6	895

Confusion Matrix of Logistic Regression

	Predicted Class		
Actual Class	-1	0	1
-1	0	0	922
0	0	0	961
1	0	0	901

Confusion Matrix of K-Nearest Neighbor

	Predicted Class		
Actual Class	-1	0	1
-1	0	0	922
0	0	0	961
1	0	0	901

Confusion Matrix of Naive Bayes

	Predicted Class		
Actual Class	-1	0	1
-1	593	184	145
0	0	567	394
1	0	262	639

Confusion Matrix of Random Forest

	Predicted Class		
Actual Class	-1	0	1
-1	910	4	8
0	10	906	45
1	8	18	875

Confusion Matrix of Decision Tree

	Predicted Class		
Actual Class	-1	0	1
-1	909	5	8
0	5	913	43
1	5	19	877

5 Conclusion

In conclusion, the project have included a lot of models and our study have comprehensively evaluated different machine learning models. The models used such as Logistic regression, KNN, Naive Bayes, Decision Tree, Random Forest and SVM gave better results after feature selection.

GITHUB: https://github.com/KhushiGC5/Classification-of-illegal-fishing

References

https://iuuriskintelligence.com/illegal-fishing-detection-using-machine-learning-ml/
https://hfmandell.github.io/Detecting-Illegal-Fishing/