

POTENTIAL FISHING ZONE

CAPSTON PROJECT



****

**AHAMMED SAHIL SHAJAHAN**

**ROLL NO:** 213301

DATE: JUNE 11,2022

Contents

[ABSTRACT 2](#_Toc105854400)

[INTRODUCTION 2](#_Toc105854401)

[BACKGROUND 2](#_Toc105854402)

[LITERATURE REVIEW 3](#_Toc105854403)

[DATASET AND DATA PREPROCESSING 5](#_Toc105854404)

[Dataset Description 5](#_Toc105854405)

[Data Standardization 5](#_Toc105854406)

[Feature importance 5](#_Toc105854407)

[Training and Testing Data 6](#_Toc105854408)

[EXPLORATORY DATA ANALYSIS 7](#_Toc105854409)

[ALGORITHMS 8](#_Toc105854410)

[Supervised Learning 8](#_Toc105854411)

[Decision Tree 9](#_Toc105854412)

[SVM 10](#_Toc105854413)

[Naive Bayes 10](#_Toc105854414)

[Assumptions Made by Naïve Bayes 11](#_Toc105854415)

[Random Forest 12](#_Toc105854416)

[BOOSTED TREE 13](#_Toc105854417)

[RESULTS AND EVALUATION 14](#_Toc105854418)

[Confusion Matrix 14](#_Toc105854419)

[SVM 14](#_Toc105854420)

[BOOSTED TREE 14](#_Toc105854421)

[DECISION TREE 14](#_Toc105854422)

[NAIVE BAYES 14](#_Toc105854423)

[Random Forest 14](#_Toc105854424)

[Comparison of Models 14](#_Toc105854425)

[CONCLUSION 15](#_Toc105854426)

[REFERENCE 16](#_Toc105854427)

**POTENTIAL FISHING ZONES**

# **ABSTRACT**

The traditional approach for determining potential fishing zones (PFZs) relies on oceanographic factors (biological, physical, and chemical) and fishermen’s expertise. This approach has disadvantages particularly when it comes to the analysis of combining these factors to find an exact PFZ spatially and temporally. In this study, we proposed a framework for identifying PFZs based on a machine learning approach. We utilized a clustering method to identify clusters of zones with data on the largest number of fish caught, which were then integrated with the sea surface temperature (SST) and the sea surface chlorophyll-a (SSC) data. The results of this data integration method were used as training data in the classification process, which was then used to determine PFZs. During the classification process, we utilized the SVM classification method. The result gave an average accuracy of 70.9%, Naive Bayes which result in an average accuracy of 93.5%, Decision tree, and Random Forest which result in an average accuracy of 100% showed that the proposed framework can be used effectively to determine PFZs.

# **INTRODUCTION**

## **BACKGROUND**

Fish capture is essential for food production in the world. Which contributes about 179 million metric tons of aquatic organisms and out of this, human consumption is about 156 million metric tons. Globally, fisheries and aquaculture provided livelihood to 59.5 million people across the globe.

In India fishing is one of the industries on which most people depend for their livelihood. India has a stretch of 7517 km on the coastline of marine, 3827 villages for fishing, and 1914 landing centers. The annual production of average fish farmers in India is 2 tons per person. Locating and catching fish is always a challenging task. And most probably the search for fish ends up in spending time and oil or resources, thus leading to low profitability. At present, an array of data collection and processing technologies such as Satellite data, machine learning automatic identification systems, and so on. Enable increasing the efficiency of fishing efforts. The satellite data such as thermal-infrared channels of NOAA-AVHRR and Eumetsat series satellites along with optical bands of Oceans-II (India) and MODIS Aqua (USA) satellites are used for the identification of Potential Fishing Zones (PFZ) along the Indian coastline. And we can also use Machine learning algorithms that can predict and quantify fishing efforts with unpredictable spatial and temporal resolutions by tracking fish vessel movements.

Combining remote sensing satellite information with the oceanographic environment and fisheries resource datasets, scientists have developed techniques to identify the location of fish which is known as a potential fishing zone (PFZ). Accurate forecasting of the fishing zone is economically advantageous for fishermen as it significantly reduces the time, effort, and resources required to search for juveniles.

In India, the Indian national center for ocean information services (INCOIS) actively distributes information on PFZs to fishermen through various communication modes throughout the year.

Oceanographic parameters are important components of fishing, which include water temperatures such as Sea Surface Temperature, Chlorophyll Humidity, Sea Level Pressure, Air Temperature, Total Cloudiness, and Total Fish catch data. In addition, studies show that SSC and SST have a strong correlation with certain aquatic environments. This correlation can be seen from the fact that different environments will have different combinations of SST and SSC.SST is a physical oceanographic component it is widely used to determine fish availability, and SSC the primary production of phytoplankton, the first food chain of the pelagic species, involves strong interrelationships. SST is the physical oceanographic component used to determine fish availability.

Here, we propose an approach for identifying PFZs based on the data-mining approach. It covers includes long-term study observation and fishing data oceanographic Data (SST and SSC). This method could be very useful in designing a complex and generalized model.

Our contribution to this paper suggests a new framework for identifying PFZs based on fisheries data using the machine learning clustering method, which separates the oceanographic features based on the areas identified as PFZs (including a large amount of data, including fish data; Among the features (SST and SSC). PROBLEM STATEMENT

# **LITERATURE REVIEW**

This section reviews the previously implemented research and thesis in this field.

The paper [1] uses The maximum entropy models which were used to evaluate the effects of oceanographic conditions on the potential fishing zones of Albacore tuna and to explore the spatial variability of these features in the SIO using satellite-based oceanographic data of sea surface temperature (SST), sea surface chlorolyll-a (SSC), sea surface salinity (SSS), mixed layer depth (MLD), sea surface height (SSH), and eddy kinetic energy (EKE) (MaxEnt). The findings show that multispectral satellite pictures can be used to describe ecosystems and can be used to predict the geographical distribution of Albacore tuna. The paper[2] they apply a data-mining strategy. and used a spatio-temporal clustering algorithm to find clusters of zones with the most fish catch data, which we then combined with data from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite images for sea surface temperature (SST) and sea surface chlorophyll a (SSC). The data integration method's outputs were used as training data in the classification process, which was then used to decide. In paper[3] Pre-processing, various data mining based methodologies, and some statistical methods were discussed, as well as the various types of validations for those data, and a methodology is provided for further enhancing the research in this area with a model, and this model can be very well applied to identify the potential fishing zones in Tamilnadu because there is no proper methodology for obtaining the details by using image processing technology. In paper [5] This comprises the gradual integration of information technology, data science, and artificial intelligence with fishing and fish farming methods to enable aquaculture production intensification, sustainable exploitation of natural fisheries, and mechanisation and automation of allied activities. To process complicated datasets and execute intelligent activities including analysing cause-effect correlations, forecasting difficulties, and giving smart-precision solutions for farming and capturing fish, unique data mining and machine learning systems are being developed. In this review, we have consolidated basic information on the various practical applications of data mining and machine learning in the aquaculture and fisheries domains from a representative selection of scientific literature, in light of the intensifying research and growing interest of stakeholders.

**Problem Statement**

Using remotely sensed data available from various satellites, the Indian National Center for Ocean Information Services (INCOIS) provides this advice to fishermen daily with specific references to fish landing centers off the Indian coast. However, when these satellites do not work, fishermen do not have PFZ data, and they have to manually pluck the oceans to find a fishing area, which is a waste of time, effort, and resources. Therefore, in such cases, we can use machine learning in the past data for PFZ identification.

# **DATASET AND DATA PREPROCESSING**

## **Dataset Description**

The dataset taken for this project is taken from Kaggle. There are 91 samples and 10 features in this dataset. The 'PFZ' in the label column is the abbreviation for the potential fishing zone, and 'NPFZ' is the abbreviation of NON POTENTIAL FISHING ZONE is the target feature in this dataset. The target column has two values: 0 and 1, with 0 indicating PFZ and 1 indicating that it is NPFZ. The rest of the 9 columns contain float values. There were no missing values in the dataset, but the target feature contained some values called '#NUM!', which were subsequently removed. Before proceeding with any further analysis, the problem of class imbalance must be solved.

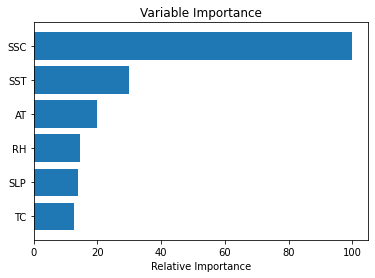
## **Data Standardization**

One of the most crucial phases in data preprocessing is standardization. All of the features are transformed from raw to standard form in this stage. In general, the dataset has a large number of features with a wide range of values. The data preprocessing is done on the training data using the StandardScaler function provided by the Scikit-learn Python library, which transforms data values from different ranges to data values in the same range, making the model more efficient.

## **Feature importance**

Feature significance can be used to improve a prediction model. This can be achieved by using key scores to select delete features (lowest scores) or retention features (highest scores). This is a type of feature selection that simplifies the problem of modeling and speeds up the modeling and, in some cases, improves the performance of the model. It is a techniques that calculate a score for all the input features for a given model — the scores simply represent the “importance” of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable.

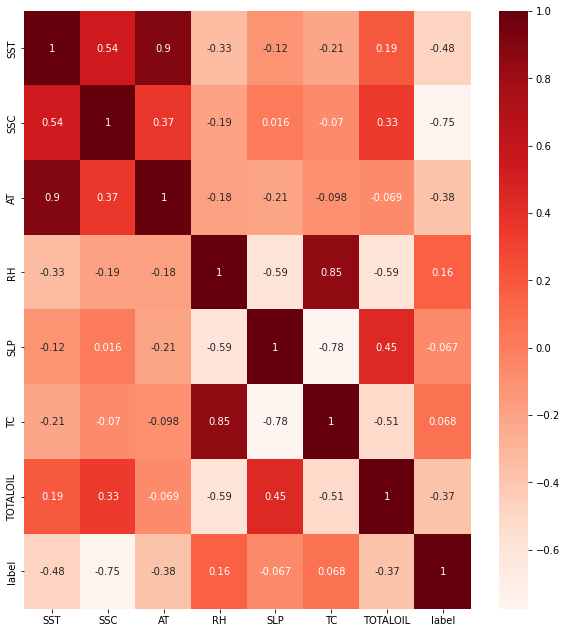
The study shows SST and chlorophyll (SSC) are an important oceanographic parameters for identifying PFZ along the coast.



## **Training and Testing Data**

The splitting of the dataset into training and testing data is the next crucial stage in data preprocessing. A model is built using the training dataset, and the model is tested using the testing dataset. The training and testing datasets are 70 percent and 30 percent of the total dataset, respectively

# **EXPLORATORY DATA ANALYSIS**

**Correlation Plot**

**countplot**



# **ALGORITHMS**

## **Supervised Learning**

Supervised learning is a type of machine learning in which machines are trained using ”labelled” data and then predict the output based on that data.The labelled data indicates that some of the input data has already been tagged with the correct output. The aim is to train the model so that it can predict the correct output when it is given a new data. The supervised learning produces an accurate result. It can be categorized into classification and regression problems. This includes various algorithms like Decision Tree, Support Vector Machine, Logistic Regression ,K-nearest neighbor, Naive Bayes classification etc.

## **Decision Tree**

A decision tree is a supervised learning technique that may be used for both classification and regression and is non-parametric. The hierarchical tree structure is made up of a root node, branches, internal nodes, and leaf nodes.



The decision tree begins with a root node that has no incoming branches, each branch representing a test outcome, as shown in the diagram. The outward branches from the root node are received by the internal nodes, also known as decision nodes. Both node types perform evaluations based on the specified attributes to yield homogenous subsets, which are denoted as leaf nodes and terminal nodes, respectively.

The leaf nodes represent all of the dataset's conceivable outcomes. Determining the attribute of the root node in each level is the main issue in Decision Tree. Attribute selection is the term for this procedure.

To find the optimal split from a root node and subsequent splits, the Gini Index was employed. The Gini Index is calculated as follows:



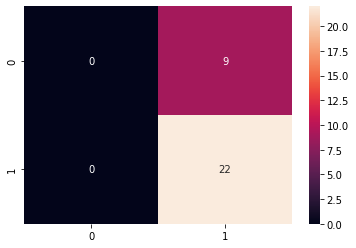
where pi is the probability of an object being classif ied to a specif ic class.

## **SVM**

Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. All of these are common tasks in machine learning.SVMs differ from other classification algorithms in that they choose a decision boundary that maximises the distance between all classes' nearest data points. The maximum margin classifier or maximum margin hyper plane is the decision boundary established by SVMs.

Both dense (numpy.ndarray and numpy.asarray) and sparse (any scipy.sparse) sample vectors are supported as input by scikit-support learn's vector machines. To use an SVM to create predictions for sparse data, however, it must have been fitted on sparse data. Use C-ordered numpy.ndarray (dense) or scipy.sparse.csr matrix (sparse) with dtype=float64 for best performance.

CONFUSION MATRIX of svm result as :



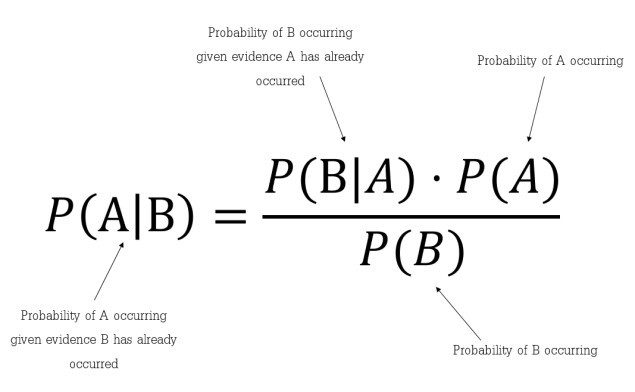
Svm confusion matrix

## **Naive Bayes**

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. In this article, we will understand the Naïve Bayes algorithm and all essential concepts so that there is no room for doubts in understanding*.*

***Bayes' Theorem***

The Bayes' Theorem is a straightforward formula for calculating conditional probabilities.Conditional probability is a measure of the likelihood of an event happening if another event has already happened (by assumption, presumption, statement, or fact).



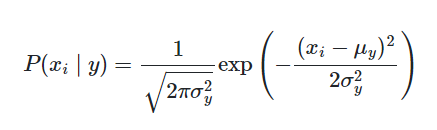
It tells us: how many times A occurs when B occurs, P (A | B) is also known as the probability of the latter, when we know: how many times B occurs when A occurs, P (B | A), and A is the probability. Own, written P (A) and how much more likely B is on its own, written P (B).

## Assumptions Made by Naïve Bayes

The fundamental Naïve Bayes assumption is that each feature makes an:

\*Independent \*equal contribution to the outcome.

[**GaussianNB**](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html#sklearn.naive_bayes.GaussianNB) implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

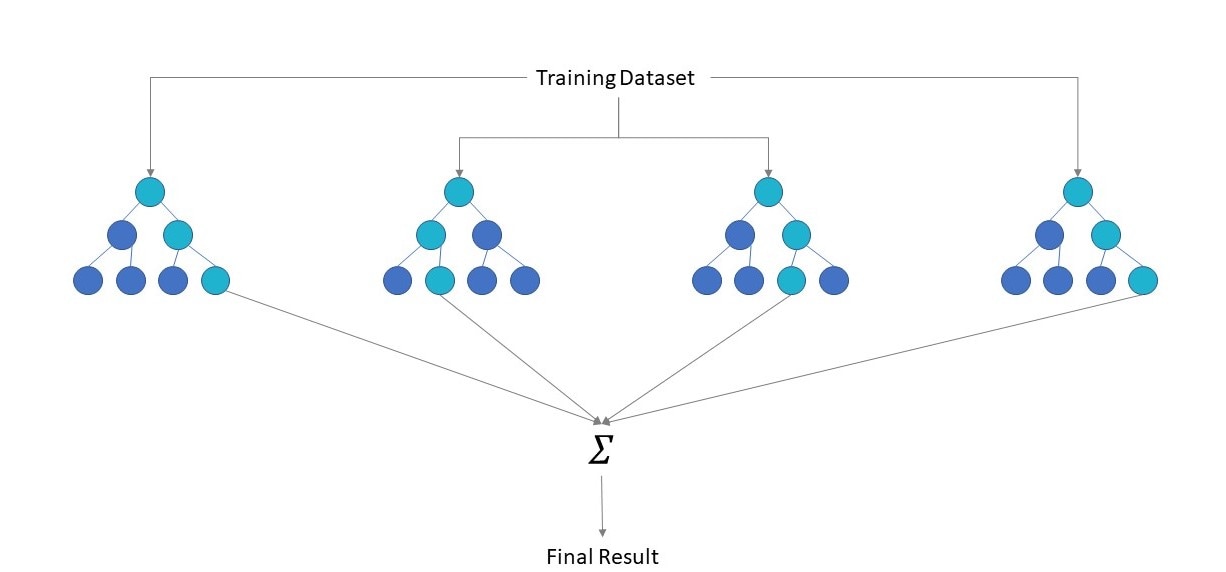


The parameters σy and μy are estimated using maximum likelihood.

# **Random Forest**

Random forest is a machine learning technique that mixes the output of numerous decision trees to produce a single outcome. Its popularity is due to its ease of use and adaptability, since it can handle both classification and regression problems. The three primary hyperparameters of random forest algorithms must be established before training. The size of the nodes, the number of trees, and the number of characteristics sampled are all factors to consider. The random forest classifier can then be used to tackle problems involving regression or classification.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is made up of a bootstrap sample, which is a data sample obtained from a training set with replacement.

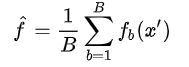


The random forest training algorithm uses the common technique of bootstrap aggregation, or bagging, to train tree learners. Bagging repeatedly (B times) takes a random sample with replacement of the training set and fits trees to these samples given a training set X = x1,..., xn and responses Y = y1,..., yn:

For *b* = 1, ..., *B*:

1. Sample, with replacement, *n* training examples from *X*, *Y*; call these *Xb*, *Yb*.
2. Train a classification or regression tree *fb* on *Xb*, *Yb*.

After training, the missing samples can be used to make predictions of x, averaging the predictions from all the individual regression trees in x ':

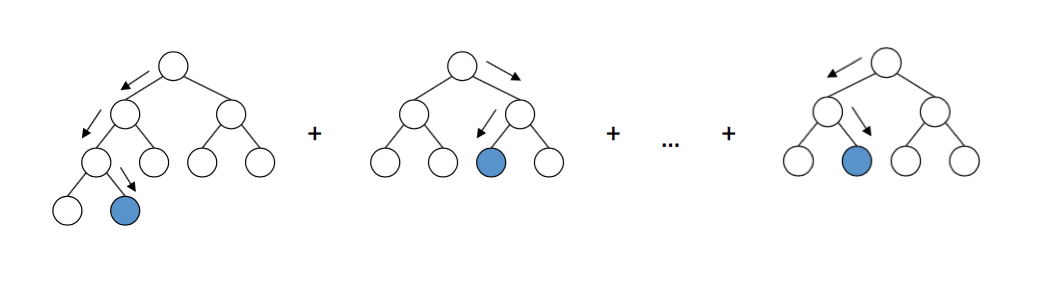


or by taking the majority vote in the case of classification trees. This bootstrapping procedure leads to better model performance because it decreases the [variance](https://en.wikipedia.org/wiki/Bias%E2%80%93variance_dilemma) of the model, without increasing the bias.

## **BOOSTED TREE**

The name "Gradient Boosting" comes from Friedman's paper Greedy Function Approximation: A Gradient Boosting Machine. XGBoost stands for "Extreme Gradient Boosting."

Gradient boosted trees have been around for a long time, and there is a wealth of information available on the subject. Using the basics of supervised learning, this lesson will explain boosted trees in a self-contained and principled manner. This explanation, we believe, is clearer, more formal, and better justifies the model formulation employed in XGBoost.



Gradient Boosting

In supervised learning, the model usually refers to the mathematical structure of the prediction from the input *xi*. A linear model, in which the prediction is presented as, a linear combination of weighted input features, is a popular example. Depending on the goal, such as regression or classification, the prediction value can be interpreted in a variety of ways. It can, for example, be logistically transformed to provide the likelihood of a positive class in logistic regression, and it can also be used as a ranking score when rating outputs.

The parameters are the parts that are unknown and must be learned from data. The coefficients are the parameters in linear regression issues. In most cases, we'll use to denote the parameters.

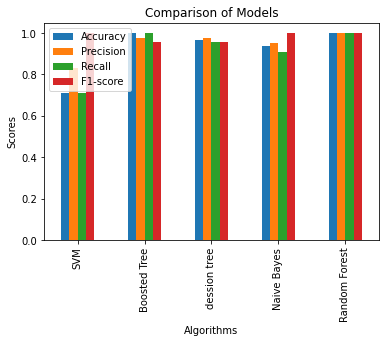
# **RESULTS AND EVALUATION**

## **Confusion Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithms** | **TN** | **FP** | **FN** | **TP** |
| **SVM** | 0 | 9 | 0 | 22 |
| **BOOSTED TREE** | 9 | 0 | 0 | 22 |
| **DECISION TREE** | 8 | 1 | 0 | 22 |
| **NAIVE BAYES** | 9 | 0 | 2 | 20 |
| **Random Forest** | 9 | 0 | 0 | 22 |

## **Comparison of Models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithms | Accuracy | Precision | Recall | F1-score |
| **SVM** | 0.709 | 0.709 | 1.0 | 0.830 |
| **BOOSTED TREE** | 1.0 | 1.0 | 1.0 | 1.0 |
| **DECISION TREE** | 0.967 | 0.956 | 1.0 | 0.977 |
| **NAIVE BAYES** | 0.935 | 0.909 | 1.0 | 0.952 |
| **Random Forest** | 1.0 | 1.0 | 1.0 | 1.0 |



# **CONCLUSION**

This study showed that the proposed framework identified areas that have high potential as fishing or PFZs based on daily fish catch data. The data integration framework is an alternative to the current data-processing procedure for large remote sensing data in extracting oceanographic characteristics. The prediction model yielded 27.06 ◦C and 0.24 mg/m3 for the SST and the SSC, respectively. The model also demonstrated that it reached 87.11% average accuracy within the predicted area. . Finally, the framework proposed in this study delivered a prediction model of the potential area based on the spatio-temporal approach based on the oceanographic characteristics SST and SSC.

# **REFERENCE**

* A Spatio-Temporal Data-Mining Approach for Identification of Potential Fishing Zones Based on Oceanographic Characteristics in the Eastern Indian Ocean Devi Fitrianah, Achmad Nizar Hidayanto, Jonson Lumban Gaol, Hisyam Fahmi, and Aniati Murni Arymurthy
* Interlacing Ocean Model Simulations and Remotely Sensed Biophysical Parameters to Identify Integrated Potential Fishing Zones P. R. C. Rahul, Sobhan Kumar Sahu, and P. S. Salvekar
* PREDICTING WINTER POTENTIAL FISHING ZONES OF ALBACORE TUNA (THUNNUS ALALUNGA) USING MAXIMUM ENTROPY MODELS AND REMOTELY SENSED DATA IN THE SOUTH INDIAN OCEAN Ming-an Lee1,2 \*, Wan-Chen Yang1 , I-Cheng Hung1 and Sheng-Yuan Teng1 1 Department of Environmental Biology and Fisheries Science, National Taiwan Ocean University, 2 Pei-Ning Rd., Keelung, 20224, Taiwan. 2 Taiwan Group on Earth Observations, 13F-3, No.32, Gaotie 2nd Road, Hsinchu 30274, Taiwan (corresponding author: [malee@mail.ntou.edu.tw](mailto:malee@mail.ntou.edu.tw))
* A remote sensing approach to monitor potential fishing zone associated with sea surface temperature and chlorophyll concentration S Karuppasamy\*,a, T P Ashithab , R Padmanabanc , M Shamsudeend & J M N Silvac