Journal of Computer Science

Original Research Paper

Machine Learning Oceanographic Data for Prediction of the Potential of Marine Resources

[1]Denny Arbahri, [1]Oky Dwi Nurhayati and [2]Imam Mudita

[1]Department of Master of Information System, Diponegoro University, Semarang, Indonesia
[2]Department of Oceanography, National Research and Innovation Agency Indonesia, Indonesia

Article history
Received: 16-09-2023
Revised: 06-11-2023
Accepted: 08-11-2023

Corresponding Author:
Denny Arbahri
Department of Master of
Information System,
Diponegoro University,
Semarang, Indonesia
Email: darbahri@gmail.com

Abstract: Marine data and information are very important for human survival, therefore this data and information is attractive to investors because of the potential economic value. This data and information has been difficult to obtain, the solution to overcome this is by analyzing oceanographic data for 2009-2019 collected from the marine database belonging to the Agency for the Study and Application of Technology (BPPT). The data is the result of a collaborative marine survey between Indonesian and foreign researchers from various countries who sailed in various Indonesian waters. Raw oceanographic data is converted and classified into Conductivity, Temperature, and Depth (CTD) data as oceanographic data parameters identified as predictor variables (X) that are correlated with each other. CTD data is processed into numeric data attributes that have been labeled for input and training. The data was modeled using the Machine Learning (ML) type Supervised Learning (SL) method with the Decision Tree (DT), Linear Regression (LR) and Random Forest (RF) algorithms which were interpreted according to the characteristics of the CTD data. ML will learn data models to understand and store. Next, the model is evaluated using accuracy metrics by measuring the difference between the predicted value and the actual value to obtain a good prediction model. The prediction results show a salinity level of 34.0 parts per thousand (ppt), meaning that in this area of marine waters salinity will affect the solubility of Oxygen (O_2) and play a major role in the sustainability and growth of the fertility level of biological resources which is supported by sea surface temperature conditions 29.2°C. So the salinity values obtained using ML techniques and marine resource potential can be assumed to have a strong correlation. The research results show that the RF model has the lowest level of prediction error based on the values: Mean Square Error (MSE) = 0.007; Root Mean Squared Error (RMSE) = 0.082; Mean Absolute Error (MAE) = 0.007 compared to DT model: MSE = 0.008; RMSE = 0.088; MAE = 0.012 and LR model: MSE = 1.008; RMSE = 1.004; MAE = 0.281. The equivalent RF and DT models have a Determination Coefficient (R^2) = 0.999, meaning that a model is created that is good at predicting, compared to the LR model with a value of R^2 = 0.914. The correlation between variables shows that the LR model is very linear with a Correlation Coefficient (r) = 1.000 compared to the DT model (r) = 0.621 and the RF model (r) = 0.379. Therefore the algorithm that has a value of (r) +1 has the best level of accuracy. The use of ML to predict marine resource potential is a relatively new research field, so this research has the potential to contribute data and information as a reference for innovative studies and investment decision material for investors.

Keywords: Machine Learning, Marine Resources, Oceanographic Data, Prediction, Supervised Learning

Science Publications

Introduction

The study of marine science is very important for human survival. Therefore, it is necessary to conserve marine resources in a sustainable manner by mastering marine data and information, which will become the basis for cognition and governance (Zhang *et al*., 2021). Existing marine data and information have not been managed and have not been used to predict marine resources using ML techniques. So mastering marine data and information has become a problem in marine resource conservation because in the past there were still many Indonesian people who used marine resources only as a means of living. However currently, marine scientific research has entered a new era of Artificial Intelligence (AI) and continues to improve marine data (Lin, 2020). AI such as ML can effectively exploit the potential information contained in large amounts of marine data (Jahanbakht *et al*., 2021). Therefore, the prospects for applying ML algorithms in marine scientific research are quite promising, especially for monitoring marine biodiversity, modeling CTD data, and predicting marine resources (Hafeez *et al*., 2019). ML algorithms are currently widely used by researchers in processing marine data and building predictive models (Bahari *et al*., 2023) and can provide large amounts of multi-parameter information to monitor complex marine ecosystems (Jiang and Zhu, 2022). The scientific development of ML to build predictive models has now been widely used by researchers, but research specifications related to the use of ML to predict marine resource potential is a relatively new research field. This research contributes marine data and information for researchers as a reference for innovative research such as to describe certain phenomena and specifications in the ocean, and as a guide for fishermen and especially fisheries investors to find fertile spots for fish in certain marine waters.

Oceanography is a branch of marine science and this study is directed at oceanographic data science in the field of marine ecology with the aim of exploring information on the potential of marine resources in Indonesian marine waters. This study was triggered by factors resulting from previous research which revealed that oceanographic factors play an important role in marine resources (Apriliani *et al*., 2020). However, the inspired idea is that the use of ML techniques to predict marine resource potential is a relatively new research area and can effectively exploit the potential information contained in large amounts of marine data. This is supported by, oceanographic information about the data; temperature, depth, and salinity can provide information about marine resource potential (Wright *et al*., 2016). Likewise, conductivity can provide information about the potential of marine resources (Müller *et al*., 2012). The salinity factor can influence the production, distribution, and lifespan of marine resources (Grilli *et al*., 2020). The acquisition process between CTD and salinity data is different, so the level of correlation must be proven with an ML algorithm in order to obtain empirical facts. This is related to this research, which aims to produce strong correlation values between relevant variables to produce accurate prediction values. Among the CTD data, there are data that influence each other and salinity is influenced by CTD (Ullman and Hebert, 2014). The ML algorithm is used to group CTD data into predictor variables (X) and salinity data into target variables (Y). Data grouping is labeled, input and trained, and modeled with SL using DT, LR, and RF algorithms which are interpreted according to the characteristics of CTD data with patterns; data distribution, data ranking, and data correlation. ML will learn the data model to understand and store it, then the model is evaluated to get a good and accurate prediction model with accuracy metrics, namely; MSE, RMSE, MAE, and R^2 are used to measure the difference between the predicted value and the actual value.

CTD parameters have their respective roles, including the role of conductivity; providing marine resources, helping regulate sea surface temperatures, helping regulate ocean currents, and helping maintain the health of marine ecosystems. Temperature plays a role; regulates life processes and the distribution of organisms and affects the amount of Oxygen (O_2) dissolved in water. The lower the sea surface temperature, the greater the solubility of oxygen in the water, and vice versa. Sea surface temperatures in Indonesian seas range between 28-31°C. Depth; It is divided into two, namely shallow waters and deep sea waters. Shallow sea waters are the sea zone, starting from the lowest angle line to a depth of around 120-200 m, the rest is the deep sea category. The ocean depth factor is closely related to the vertical temperature. Salinity is part of oceanographic parameters but is a separate parameter from CTD (Mensah *et al*., 2009). The strong correlation between CTD data influences each other and influences salinity, this will be shown by a correlation value that is close to +1 or -1 and will be an indicator in predicting marine resource potential. Therefore, the correlation value is strong and relevant between CTD data and has an effect on salinity as a result of accurate prediction values.

The use of ML techniques to analyze CTD oceanographic data as study material for predicting marine resources is still relatively new. Therefore, the development of ML science from previous research topics experienced more varied developments with increasing ML insights which were studied by combining oceanographic science. The similarity factor with previous research becomes inspirational material as a result of a thought that is worthy of being quoted and accompanied by citation as a scientific development that underlies this research. The following is a literature review presented in (Table 1), which outlines a summary of various areas of previous research.

Table 1: Summary of previous research literature review

Research areas	Method	Research results	Reference
Knowledge-based systems	CNN, LSTM, ConvLSTM	In total, four different methods are used to predict road surface temperature on inefficient roads. One of them is linear regression, which is a classic statistica l regression technique; the other three methods are machine learning techniques, including supporting vector regression, multilayer perceptron artificial neural networks and random forest regression. Graphical and numerical results show that vector regression is the most accurate method	Hatamzad *et al*. (2022)
Computers and geoscience	Cluster method	The new approach based on anomaly detection technique that I propose here greatly impacts the QC of oceanographic data in twofold. First, optimizing the expertise to efficiently handle the ever-increasing number of measurements in the oceans. Second, it combines some characteristics of each measurement for deeper decision-making, resulting in a higher context awareness for more complicated classifications	Castelão (2021)
Environment management	ANN and SVM	This study uses two ML models, namely ANN and SVM. The performance of the SVM model results is better than the ANN model in predicting water quality	Deng *et al*. (2021)

Materials and Methods

The process of collecting oceanographic data for 2009-2019 began with downloading from the BPPT marine database, the results of collaborative marine surveys between Indonesian and foreign researchers from various countries, such as America, Europe, Japan, and China using BPPT's research vessels sailing in various sea waters Indonesia. Oceanographic data that has been downloaded must be converted and analyzed by expert oceanographic analysts because the raw data still lacks information. The process it carries out, groups CTD data in .hex format and is converted into .cnv format so that the converted data contains information headers; sensor acquisition and calibration times, information, and parameter identity instructions. The result is CTD data as primary data for this research which is presented in Excel format containing numerical data. Raw data when processed goes through 3 stages, namely: First, numerical data which does not yet have information headers (Table 2), secondly it is converted into graphic data (Fig. 1), thirdly it is converted into numerical data which already has complete information header information (Table 3).

Conductivity data is recorded digitally based on seawater's level of conductivity. The temperature data analyzed is sea surface temperature and this data is recorded in degrees Celsius. Depth data is analyzed according to the level of seawater depth, which is directly related to the pressure of seawater currents, and this data is recorded in meters.

The flow of this research is presented in a diagram that describes the stages of research methodology using ML methods for oceanographic data (Géron, 2022), this is to make it easier to record traces of the origin of the data collected, the picture is shown in (Fig. 2).

The oceanographic data collection method begins with downloading from the BPPT marine database. Before the raw data becomes primary data used in this research, it must be identified, selected, analyzed, and converted because there is no information header. This process requires an expert oceanographic analyst to produce CTD data as a predictor variable (X), which is an oceanographic data parameter to predict salinity data as a target variable (Y). The following is the process of converting raw data into primary data in numerical form which already has complete information headers:

- Data CTD obtained from mining results from the BJ research vessel is in the format .hex
- Data format CTD .hex is converted to .cnv and then made per meter
- Data CTD is called with MATLAB one by one, then the data is treated by means; data is deleted if there is an error, data is QC against the spiked value and then filled in with interpolation, information on coordinates and station number is entered, finally, the raw data is exported
- After all the stations have been exported, we call these stations with a loop to combine them into 1 file .txt

Exploratory Data Analysis (EDA) stage, CTD data parameters are analyzed to be identified as predictor variables that have a correlation with CTD data in order to obtain quality data.

In the process and analysis stage, variable X is processed into numeric data attributes which are labeled for the data input process.

The modeling stage is, the process of characterizing data with patterns; data distribution (Fig. 3), data ranking (Table 4), and data correlation (Table 5). The modeling process uses training data that has been labeled so that ML can learn a data model to understand and store.

Table 2: Examples of raw data that have not been analyzed with oceanographic science

295	0	27.81	33.01
298	0	27.53	0.05
299	0	27.40	0.05
300	0	27.27	0.05
301	0	27.16	0.05
302	0	27.07	0.05
303	0	27.01	0.05
304	0	26.94	0.05
305	0	26.89	0.05

Table 3: CTD data which is the primary data in this study

Data_id	Station_sum	Depth_m	Temp_degC	Cond_mS/cm	Sal_psu
Cruise_2009	3	1-1500.00	29.24-3.860	56.10-32.72	34.01-34.56
Cruise_2010	5	1-29.00	29.57-29.24	51.51-53.18	30.69-32.02
Cruise_2011	3	1-22.00	28.12-28.25	52.18-53.44	32.11-32.87
Cruise_2012	4	1-376.00	28.30-9.830	53.15-37.68	32.64-34.57
Cruise_2013	0	0	0	0	0
Cruise_2014	2	1-16.78	28.91-30.39	52.64-56.82	31.88-33.68
Cruise_2015	5	1.73-1010.17	29.48-6.420	56.82-35.02	34.33-34.79
Cruise_2016	1	1-6.31	29.64-29.50	50.21-50.25	29.77-29.89
Cruise_2017	2	5-54.00	28.88-28.16	53.20-52.10	32.28-32.01
Cruise_2018	1	1-12.00	29.78-28.73	53.59-53.08	31.94-32.30
Cruise_2019	3	1.59-69.20	22.69-21.71	50.06-49.57	34.52-34.90

Table 4: Ranking of CTD data

	Univariate regression	RReliefF
Temp_degC	73465,745	0.001
Cond_mS /cm	32393,575	0.001
Depth_m	12966,827	0.001

Table 5: Correlation between variables

Correlation	r
Conductivity and temperature	0.694
Depth and temperature	-0.839
Conductivity and depth	-0.835
Salinity and temperature	-0.682
Conductivity and salinity	-0.652
Depth and salinity	0.321

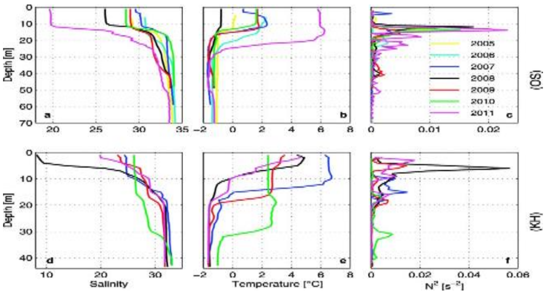

Fig. 1: Graphical data from the results of oceanographic scientific analysis

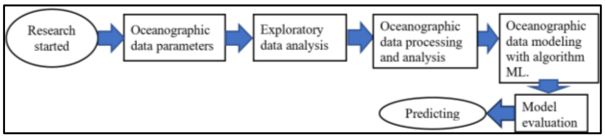
Research started
Oceanographic data parameters
Exploratory data analysis
Oceanographic data processing and analysis
Oceanographic data modeling with algorithm ML.
Model evaluation
Predicting

Fig. 2: Research illustrating workflow diagrams of the stages of the study methodology using the ML oceanographic data method

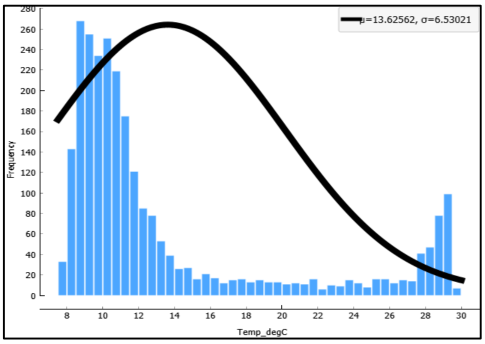

Fig. 3: The distributions of CTD data are well spread out

In the evaluation stage, data modeling is evaluated using the basic concept of accuracy evaluation, namely comparing targets with predictions using accuracy metrics consisting of; MAE, MSE, RMSE, and R^2. The aim of model evaluation is to ensure that the use of ML algorithm methods is in accordance with the accuracy of using CTD data.

The output results is the result of data modeling that has been trained to obtain the right model to be used in predicting the target variable to produce values for the Y variable.

This data distribution display is to see the CTD data distribution pattern with the ML algorithm. The following is shown in, which shows that the distribution of data is well distributed and does not overlap so that the primary data to be used for this research material is as needed.

The feature ranking in CTD data is intended to see which feature is the most significant and the ranking calculation depends on the type of variable being observed, whether it is categorical or numerical. The CTD data in this study are all numerical variables, so the application must use the univariate regression and RReliefF approaches to determine significant feature values. The most significant feature value is the value that has the highest ranking, followed by the values below it. The picture shows that the highest ranking value is temperature, followed by conductivity and depth is the lowest in influencing Salinity.

CTD data correlation is measured by comparing variable X with variable Y to determine the linear or nonlinear relationship between these variables. The degree whose value is close to -1 or +1 will indicate the degree of correlation that influences each other in predicting the target (Samudrala, 2019).

Results and Discussion

Oceanographic Data Processing and Analysis

The ML method with DT, LR, and FR algorithms for marine resource prediction is the choice in this study because the aim is to be able to map input to output flows to solve regression problems:

- Decision tree: There are two types of decision tree methods, namely Classification and Regression Trees (CART), but because in this study the types of data used are all numeric data, what is applied is Regression Tree and for its application, it must use the $C_{4.5}$ algorithm, which is an algorithm that generally uses categorical and numeric data to evaluate all attributes. The methodology is to prepare training data to select attributes that are calculated using the Entropy, Split information, and gain ratio formulas (Maimon and Rokach, 2005)

Entropy formula:

$$Entropy(S) = \sum_{i}^{c} - Pi\ log2\ Pi \qquad (1)$$

c = Number of classes
P_i = Object data
i = Number of samples in the data set
S = Data sample set

Splits information formula:

$$Split\ Information = \sum_{i}^{c} - \frac{Si}{S} log_2 \frac{Si}{S} \qquad (2)$$

*S*1-*Sc* = Subsets of data samples that are divided based on the number of variations in the value of attribute A.

Gain ratio formula:

$$Gain\ Ratio\ (S, A) = \frac{Gain(S,\ A)}{Split\ information(,A)} \qquad (3)$$

Application of the $C_{4.5}$ algorithm with the Entropy, *split information*, and *gain ratio* formulas has been calculated. However, when calculating entropy with input data that is entirely numerical, the calculation model cannot accept it. The count pattern will only accept appropriate input data, namely numeric data that includes category data as the class label, so DT images cannot be displayed. Therefore, the alternative solution is to input numerical data into ML modeling to obtain output at the root node (Yaseen *et al*., 2020), as shown in (Fig. 4).

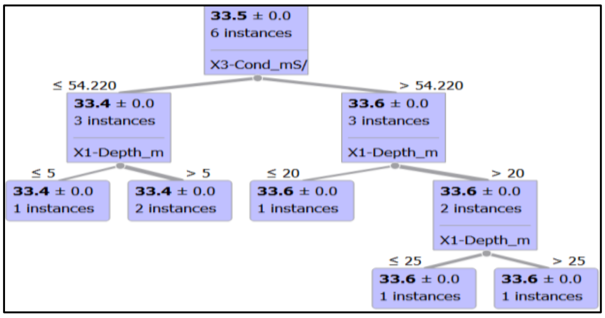

Fig. 4: Roots nodes machine learning modeling of the decision tree algorithms

The root node mapping produced by the calculation of the $C_{4.5}$ algorithm shows that X_3 has a horizontal dividing line with equation = 54.220. That is, data samples below the horizontal line are recommended, while data samples above the horizontal line are not. On the right side of the horizontal dividing line, there are 2 samples with errors, namely, X_1, with equation = 20 and equation = 25, which show a salinity level of 33.6 ppt. On the other hand, on the left, X_1, with equation = 5, shows the level of salinity at 33.4 ppt. Linear regression: The mathematical equation that describes the relationship between the independent and dependent variables is often called the regression equation. In this study, the CTD data is entered into the regression equation, which consists of one dependent variable, Y, and several independent variables, X_1, X_2, and X_3.

The relationship between these variables can be written in the form of an equation:

$$Y = \alpha + \beta 1\ X_2 + \beta 2\ X_2 + \beta n\ Xn + e \qquad (4)$$

Y = Dependents variables or response variable
X = Independent variable or predictor variable
α = Constant
β = Slopes or coefficient estimate

Then a statistical method was chosen, namely Residual Sum of Squares (RSS) to help identify the level of dataset differences that are not predicted by the regression model.

Residual sum of squares formula:

$$RSS = \sum_{i=1}^{n}(y_i - y)^2 = \sum_{i=1}^{n}\left(y_i - (\alpha + \beta x_i)\right)^2 \qquad (5)$$

Y_i = Value of the observed variable
Y = Value estimated by the regression linear
X_i = An independent value
α and β = Constant

RSSesidual Sum of Squares calculations for predictions X_1, X_2, X_3, based on data in (Table 6) and calculation results in (Table 7).

Calculation of predictor X_1-depth index x_0:

$$n = 1 \qquad Y_{R1} = \frac{33.417}{1} = 33.417$$

$$Y_R{}^2 = \frac{33.418+33.418+33.554+33.605+33.637}{5} = 33.5$$

$$RSS_0 = (33{,}417\text{-}33{,}417)^2 + [(33{,}418\text{-}33{,}526)^2 + (33{,}418\text{-}33{,}526)^2 + (33{,}554\text{-}33{,}526)^2 + (33{,}605\text{-}33{,}526)^2 + (33{,}637\text{-}33{,}536) = 0.043$$

Table 6: Predictors: X_1-X_2-X_3

Index	X_1-Depth_m	X_2-Temp_degC	X_3-Cond_mS/cm	Y#Sal_psu
0	5	28,27	54,22	33,41
1	10	28,27	54,21	33,41
2	15	28,26	54,21	33,41
3	20	28,25	54,39	33.55
4	25	28,26	54,48	33,60
5	30	28,23	54.50	33,63

Table 7: RSS $_{0\text{-}5}$ calculations results for predictors X_1, X_2, X_3

	Predictor X_1-depth	Predictor X_2-temperature	Predictor X_3-conductivity
RSS_0	0.043	0.033	0.101
RSS_1	0.028	0.030	0.123
RRS_2	0.004	0.004	0.200
RSS_3	0.015	0.028	0.139
RSS_4	0.033	0.043	0.091
RSS_5	0.053	0.052	0.053

RSS value analysis used statistical standards, significance level (α) = 0.05, the analysis.

The value of the regression coefficient X_1 = 0.004 <0.05, indicates a strong and linear relationship between depth and salinity.

The value of the regression coefficient X_2 = 0.004 <0.05, indicating a linear and strong relationship between Temperature and salinity.

The value of the regression coefficient X_3 = 0.053 = 0.05, indicating a linear and reasonable relationship between conductivity and salinity.

This means that to detect changes in sea depth and changes in sea water temperature, the conductivity will change. Thus, CTD data plays an important role in predicting the location of marine waters:

- Random forest: Ensemble machine learning is a technique that uses multiple decision tree algorithms to create one powerful algorithm (Breiman, 2001). In contrast to decision trees, which depend on a single tree, RF depends on multiple trees, which helps to have the most efficient predictive power with less uncertainty and overfilling (Tilahun and Korus, 2023). RF is used as a comparison because it is not found in other methods (Speiser *et al*., 2019). It can also handle data sets containing continuous variables, such as in the regression case in this study. RF consists of tree-shaped classifications $\{h(x, \theta k), k = 1, .\}$ where θk is an independently distributed random vector and each tree will choose the most popular class at input X. Here are the accuracy characteristics of a random forest: there are classifiers $h_1(x)$, $h_2(x)$, . . . , hk (X) and with training set of random vector distribution Y, X (Han *et al*., 2018). The flow of implementing the RF algorithm is carried out in stages; determine the number of decision trees, take random sample data to form a decision tree, sample data is calculated with the gini index to determine the top node

Gini index formula:

$$Gini = 1\text{-} \sum(pi)2\ n \tag{6}$$

i = 1
Pi = Probability of the object to be classified in a particular feature

RF algorithm calculation results:

- Predicted salinity level of 34.0 ppt
- Weak correlation coefficient of 0.379
- Prediction error rate, for the difference between the predicted value and the real value, MSE = 0.007

This means that the predicted salinity value of 34.0 ppt has the potential to become an area of fertile marine waters which is supported by a low error rate of 0.007 in the predicted salinity value, but the level of correlation between CTD data according to the RF algorithm is only 0.379 (weak standard, 0.2≤ r≤0.39).

Model Evaluation

Data modeling must be evaluated with the basic concept of accuracy evaluation, namely comparing targets with predictions using accuracy metrics; MSE, RMSE, MAE, and R^2 with the aim of ensuring the use of ML algorithm methods in accordance with the accuracy of using CTD data.

Mean Square Errors (MSE); error value of the difference between real value and predicted value.

MSE formulas:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Yi - Yi)^2 \tag{7}$$

Table 8: Test results and scores of modeling performance values

	MSE	RMSE	MAE	R^2
Decision tree	0.008	0.088	0.012	0,999
Random forest	0.007	0.082	0.007	0,999
Linear regression	1.008	1.004	0,281	0,914

Table 9: Correlation coefficient of CTD between algorithm

Model	Decision tree	Random forest	Linear regressions
Decision tree		0.621	0.000
Random forest	0.379		0.000
Linear regressions	1,000	1,000	

Root Means Squared Error (RMSE); squared error value between real and predicted values.

RMSE Formulas:

$$RMSE = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} (Yi - Yi)^2 \quad (8)$$

Mean Absolute Error (MAE); average error value between real and predicted values.

MAE formulas:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} (Yi - Yi) \quad (9)$$

The calculation results show that the RF model has the lowest prediction error rate with accuracy metric values: *MSE* = 0.007; *RMSE* = 0.082; *MAE* = 0.007 compared to DT and LR models, complete results are shown in (Table 8).

The RF and DT models are both equally good models in predicting the target variable with a value of $R^2 = 0.999$ compared to the LR model with a value of $R^2 = 0.914$.

The results of evaluating prediction models with accuracy metrics have shown that modeling with the RF algorithm has the lowest prediction error rate compared to DT and LR. However, because this research wants to produce strong correlation values between variables and is relevant to produce accurate prediction values, it is necessary to measure the correlation coefficient of CTD data by comparing variables X with Y to determine the linear or nonlinear relationship between these variables. The degree whose value is close to +1 or -1 will indicate the degree of correlation that influences each other in predicting the target (Samudrala, 2019), as shown in (Table 9).

Correlation between algorithmic models shows that the LR model is very linear with a Correlation Coefficient value (r) = 1.000 compared to the DT model (r) = 0.621 and the RF model (r) = 0.379, so the LR model is an algorithm that has a better level of accuracy compared to DT and RF.

Predicting Marine Resources

The prediction results of the three algorithms show that the target results tend to be the same, namely 34.0 as seen in (Table 10), meaning that a sea water salinity level of 34.0 ppt has the potential for renewable marine resources. So it affects the solubility level of Oxygen (O_2) in water and has a big role in the sustainability and growth of the fertility level of biological resources in marine waters. This is also caused by the influence of upwelling in the east monsoon which is supported by sea surface temperature conditions of 29.2°C and salinity levels of 34.0 ppt (Kuswardani and Qiao, 2014).

This predicted value is very closely related to the results of previous research (Tangke *et al*., 2011), that sea surface temperatures of 29.1-29.5°C and salinity levels ranging from 32.7-34.2 ppt were obtained from fish catches. The highest reached 669,930 kg.

The temperature characteristics of fishing potential in the sea surface temperature range of 29.1-29.5°C are also related to data showing the pattern and value of the distribution of fishery potential in Indonesian territory according to the reference of the Minister of Maritime Affairs and Fisheries Regulation number 18 of 2014, which stipulates that there are 11 fishery management areas of the Republic of Indonesia, shown in (Fig. 5) and (Table 11).

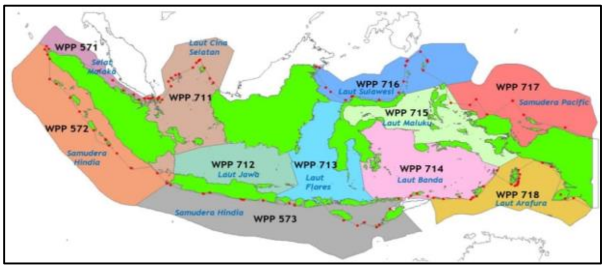

Fig. 5: Data on the distribution and potential value of fisheries in Indonesia

Table 10: Predictions results

Linear regression	Decision tree _	Random forest	Depth_m	Temp_degC	Cond_mS /cm
34.0496	34.0077	34.0069	1	29.2459	56.1052
34.0425	34.0077	34.0048	2	29.2561	56.1013
34.0525	34.0077	34.0160	3	29.2616	56.1296
34.0623	34.0333	34.0345	4	29.2632	56.1533
34.0667	34.0333	34.0381	5	29.2630	56.1635
34.0748	34.0512	34.0521	6	29.2579	56.1767
34.1125	34.0172	34.1108	7	29.2542	56.2548
34.0507	34.0075	34.0133	8	29.5276	56.4050
33.8608	33.7334	33.7319	9	29.6709	56.1459
34.0526	34.0161	34.0171	10	29.2447	56.1193

Table 11: Data on the distribution and potential value of fisheries in Indonesia

Fishery potential in Indonesian waters territories		Fisheries management areas (WPP)
Waters	Malacca straits	WPP 571
Waters	Indian Ocean west of Sumatra and the Sunda Strait	WPP 572
Waters	Indian Ocean to the south of Jawa, to the south of Nusa Tenggara, the Sawu Sea, the western part of the Timor Sea	WPP 573
Waters	Karimata Sea, Natuna Sea, South China sea	WPP 711
Waters	Jawa Sea	WPP 712
Waters	Makassar Strait, bone bay, Flores Sea, Bali sea	WPP 713
Waters	Tolo bay, Banda sea	WPP 714
Waters	Tomini bay, Maluku sea, Halmahera sea, Seram sea, Berau bay	WPP 715
Waters	Celebes Sea, north of Halmahera island	WPP 716
Waters	Cendrawasih Bay, Pacific Ocean	WPP 717
Waters	Aru sea, Arafuru sea, East Timor sea	WPP 718

Table 12: Description of pearson correlation coefficient

Correlation coefficient	Correlation level
0.0≤ r≤0.19	Very weak
0.2≤ r≤0.39	Weak
0.4≤ r≤0.59	Moderate
0.6≤ r≤0.79	Very strong
0.8≤ r≤0.10	Strong

Table 13: Correlation of features between this research and previous research

This research		Previous research	
Correlation	r	Correlation	r
Conductivity_temperature	0,999	Surface_temp1	0,960
Depth_temperature	-0.839	Base_temp	0,740
Conductivity_depth	-0.835	Dew_point	0,722
Salinity_temperature	-0.682	Surface_state1	0,601
Conductivity_salinity	-0.652	Snow_L	0,615
Depth_salinity	0,462	Water_t2	0,625

Comparative Analysis of this Research with Other Research

This research has similarities with previous research in the use of ML techniques for prediction but is different in the use of algorithms. Likewise, there are differences in the use of CTD oceanographic data as primary research data for marine resource predictions, so this research has relatively new ideas.

The uniqueness of this research compared to previous research (Hatamzad *et al*., 2022) lies in the large level of correlation between the independent and dependent variables. A high level of correlation between variables will make the output more efficient and have a good level of accuracy. In this regard, ML algorithms require a good level of accuracy so that predictive machine learning can be supervised (Yilmazer and Kocaman, 2020). The following shows the level of correlation values in (Table 12) and a comparison of the correlation values of this research with previous research in (Table 13).

Conclusion

This research uses ML techniques for marine resource prediction which is relatively new, SL with DT, LR, and RF algorithms as part of the type of ML technique is the choice for modeling. The process of determining the algorithm used must undergo measurements with the data criteria used. CTD oceanographic data criteria measurements are carried out in 3 ways: First, with data distribution to see that the data distribution pattern is well distributed and does not overlap with each other, so that the data that will be used is as needed and can be analyzed, meaning that to determine the amount of data that will be used, it must be tested first using data distribution techniques; secondly, by ranking the data to see which features are the most significant and calculating the ranking depending on the type of variable observed, whether it is categorical or numerical, the data used in this study are all numerical variables, meaning that to determine the significance of the data that will be used, it has an effect on the value prediction results. numerical results such as predicted salinity levels of 34.0 ppt; third, with data correlation to determine the linear or nonlinear relationship between variables *X* and *Y*, the degree whose value is close to -1 or +1 will indicate the degree of correlation that influences each other in predicting the target.

The results of this research show a relationship between conductivity and salinity (r) = -0.652, and temperature and salinity (r) = -0.682. Meanwhile, the correlation between depth and salinity (r) = 0.462 is weaker in influencing the target. The values r = -0.652 and r = -0.682, meaning to detect how the conductivity and temperature of water change at each depth, can be analyzed on the physical properties of water. Thus, CTD data plays an important role in predicting the location of marine waters for further exploration in the future.

The use of the DT algorithm with C4.5 in the calculation process experienced problems because the data used in this research were all numerical data, while the C4.5 algorithm calculation model could not accept it. The calculation pattern will only accept appropriate input data, namely numeric data that include categorical data as its class label, so DT images cannot be displayed. Therefore, an alternative solution is to enter numerical data into ML modeling to obtain output at the root node. The results of the root node image have a horizontal dividing line and those below the horizontal line are recommended, while those above the horizontal line are not recommended with equation = 5 indicating a salinity level of 33.4 ppt, meaning that at the beginning of using the DT algorithm it can estimate predictions salinity value. The results of the prediction model evaluation show that the three models used all have superior levels of accuracy. The RF model was evaluated with accuracy metrics (MSE, RMSE, MAE), having the lowest prediction error rate compared to the DT and LR models. However, the RF and DT models are equivalent to having a better model in predicting the target variable compared to the LR model, evaluated by the level of the coefficient of Determination (R^2). The LR model has the advantage of strong correlation values between relevant variables to produce accurate prediction values compared to the DT and RF models, evaluated by the level of correlation coefficient (r). Because this research wants to produce strong correlation values between relevant variables to produce accurate prediction values, the LR model is an algorithm that has a superior level of accuracy compared to the DT and RF algorithms, meaning that with the modeling results of the three algorithms, the focus of the analysis is on Predictive value is aimed at modeling with the LR algorithm.

The prediction results of the three algorithm models show an average salinity level of 34.0 ppt at an average sea surface temperature of 29.2°C, which is an area of fertile marine waters and has the potential for prosperous biological resources. These predicted values are very closely related strongly with the results of previous research (Tangke *et al.*, 2011). The research results are in sync with the results of accurate ML predictions in this study, that at sea surface temperatures of 29.1-29.5°C and salinity levels ranging from 32.7-34.2 ppt the highest fish catch was obtained reaching 669,930 kg. This means that fishermen and fisheries investors before taking action should pay attention to this valuable marine data and information so that their decisions to fish or invest are well planned.

Paying attention to the R^2 value in the algorithm model LR = 0.914 means that the ability to predict the target variable is 91.4%, thus there is still the influence of other factors of 8.6%, namely apart from the CTD data factor. Seeing the large values of other factors, this means that this research still does not contribute much in terms of predicting more elements of marine resources other than fisheries. So that in the future, the research can be further improved by adding oceanographic data parameters other than CTD data.

Acknowledgment

We would like to thank Arifiyana (center for marine survey technology-BPPT) as an expert oceanographic analyst who has helped analyze oceanographic data starting from raw data downloaded from BPPT's database to CTD data in complete information format as primary data in this research. This research is supported in terms of data by BRIN and the postgraduate school of Diponegoro University-Indonesia in terms of data science. However, this research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Funding Information

This research received no external funding.

Author's Contributions

Denny Arbahri: Conducted data preparation, analysis, and interpretation; drafted and edited the manuscript.

Oky Dwi Nurhayati: Supervised the research on data science aspects; reviewed and revised the manuscript.

Imam Mudita: Contributed to the oceanographic interpretation; suggested and implemented revisions to the manuscript.

Ethics

The author states that the research manuscript in this publication is original and has never been published in any journal, because we, the authors, comply with the applicable regulations in the Journal of Computer Science. If it is proven that this manuscript has been published in another journal, we are ready to accept sanctions. There are no ethical issues because the paper has been read, validated, and approved by all authors. There are no ethical issues related to data collection, because the process of obtaining and retrieving data from the database belonging to the agency for the assessment and application of technology (BPPT, now renamed BRIN), is a department in our office. There are no issues regarding conflicts of interest.

References

Apriliani, I. M., Putra, P. K., Dewanti, L. P., & Akbarsyah, N. (2020). Geographic information systems an analysis instrument for oceanographic parameters of catches: A review of research. *International Journal of all Research Writings*, *2*(8), 1-7.

Bahari, N. A. A. B. S., Ahmed, A. N., Chong, K. L., Lai, V., Huang, Y. F., Koo, C. H., ... & El-Shafie, A. (2023). Predicting Sea Level Rise Using Artificial Intelligence: A Review. *Archives of Computational Methods in Engineering*, 1-18. https://doi.org/10.1007/s11831-023-09934-9

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Castelão, G. P. (2021). A machine learning approach to quality control oceanographic data. *Computers and Geosciences*, *155*, 104803. https://doi.org/10.1016/j.cageo.2021.104803

Deng, T., Chau, K. W., & Duan, H. F. (2021). Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management*, *284*, 112051. https://doi.org/10.1016/j.jenvman.2021.112051

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc. ISBN: 10-9781492032649.

Grilli, F., Accoroni, S., Acri, F., Bernardi Aubry, F., Bergami, C., Cabrini, M., ... & Cozzi, S. (2020). Seasonal and interannual trends of oceanographic parameters over 40 years in the northern Adriatic Sea in relation to nutrient loadings using the EMODnet chemistry data portal. *Water*, *12*(8), 2280. https://doi.org/10.3390/w12082280

Hafeez, S., Wong, M. S., Ho, H. C., Nazeer, M., Nichol, J., Abbas, S., ... & Pun, L. (2019). Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: A case study of Hong Kong. *Remote Sensing*, *11*(6), 617. https://doi.org/10.3390/rs11060617

Han, T., Jiang, D., Zhao, Q., Wang, L., & Yin, K. (2018). Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control*, *40*(8), 2681-2693. https://doi.org/10.1177/0142331217708242

Hatamzad, M., Pinerez, G. C. P., & Casselgren, J. (2022). Intelligent cost-effective winter road maintenance by predicting road surface temperature using machine learning techniques. *Knowledge-Based Systems*, *247*, 108682. https://doi.org/10.1016/j.knosys.2022.108682

Jahanbakht, M., Xiang, W., Hanzo, L., & Azghadi, M. R. (2021). Internet of underwater things and big marine data analytics-a comprehensive survey. *IEEE Communications Surveys and Tutorials*, *23*(2), 904-956. https://doi.org/10.1109/COMST.2021.3053118

Jiang, M., & Zhu, Z. (2022). The Role of Artificial Intelligence Algorithms in Marine Scientific Research. *Frontiers in Marine Science*, *9*, 920994. https://doi.org/10.3389/fmars.2022.920994

Kuswardani, R. T. D., & Qiao, F. (2014). Influence of the Indonesian Throughflow on the upwelling off the east coast of South Java. *Chinese Science Bulletin*, *59*, 4516-4523. https://doi.org/10.1007/s11434-014-0549-2

Lin, B. (2020, April). Overview of High Performance Computing Power Building for the Big Data of Marine Forecasting. In *2020 International Conference on Big Data and Informatization Education (ICBDIE)* (pp. 79-82). IEEE. https://doi.org/10.1109/ICBDIE50010.2020.00025

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook* (Vol. 2, No. 2005). New York: Springer. https://doi.org/10.1007/978-0-387-09823-4_8

Mensah, V., Le Menn, M., & Morel, Y. (2009). Thermal mass correction for the evaluation of salinity. *Journal of Atmospheric and Oceanic Technology*, *26*(3), 665-672. https://doi.org/10.1175/2008JTECHO612.1

Müller, H., von Dobeneck, T., Hilgenfeldt, C., SanFilipo, B., Rey, D., & Rubio, B. (2012). Mapping the magnetic susceptibility and electric conductivity of marine surficial sediments by benthic EM profiling. *Geophysics*, *77*(1), E43-E56. https://doi.org/10.1190/geo2010-0129.1

Samudrala, S. (2019). *Machine Intelligence: Demystifying machine learning, neural networks and deep learning*. Notion Press. ISBN: 10-1684660831.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93-101. https://doi.org/10.1016/j.eswa.2019.05.028

Tangke, U., Mallawa, A., & Zainuddin, M. (2011). Analysis of the relationship between oceanographic characteristics and catches of yellowfin tuna (*Thunnus albacares*) in the Banda Sea waters. *Agrikan: Journal of Fisheries Agribusiness*, *4*(2), 1-14. https://doi.org/10.29239/j.agrikan.4.2.1-14

Tilahun, T., & Korus, J. (2023). 3D hydrostratigraphic and hydraulic conductivity modelling using supervised machine learning. *Applied Computing and Geosciences*, 100122. https://doi.org/10.1016/j.acags.2023.100122

Ullman, D. S., & Hebert, D. (2014). Processing of underway CTD data. *Journal of Atmospheric and Oceanic Technology*, *31*(4), 984-998. https://doi.org/10.13155/33951

Wright, S., Hull, T., Sivyer, D. B., Pearce, D., Pinnegar, J. K., Sayer, M. D., ... & Hyder, K. (2016). SCUBA divers as oceanographic samplers: The potential of dive computers to augment aquatic temperature monitoring. *Scientific Reports*, *6*(1), 30164. https://doi.org/10.1038/srep30164

Yaseen, Z. M., Al-Juboori, A. M., Beyaztas, U., Al-Ansari, N., Chau, K. W., Qi, C., ... & Shahid, S. (2020). Prediction of evaporation in arid and semi-arid regions: A comparative study using different machine learning models. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 70-89. https://doi.org/10.1080/19942060.2019.1680576

Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy*, *99*, 104889. https://doi.org/10.1016/j.landusepol.2020.104889

Zhang, B., Li, F., Zheng, G., Wang, Y., Tan, Z., & Li, X. (2021). Developing big ocean system in support of Sustainable Development Goals: Challenges and countermeasures. *Big Earth Data*, *5*(4), 557-575. https://doi.org/10.1080/20964471.2021.1965371