Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Predictive modeling of marine fish production in Brunei Darussalam's aquaculture sector: A comparative analysis of machine learning and statistical techniques

Haziq Nazmi, Nor Zainah Siau, Arif Bramantoro *, Wida Susanty Suhaili

School of Computing and Informatics, Universiti Teknologi Brunei, Bandar Seri Begawan, Brunei Darussalam

ARTICLE INFO

Article history:
Received 13 December 2022
Received in revised form
18 April 2023
Accepted 19 May 2023

Keywords:
Aquaculture industry
Predictive modeling
Machine learning techniques
Marine fish production
Brunei Darussalam

ABSTRACT

The aquaculture industry has witnessed significant global growth, offering opportunities for sustainable fish production. This research delves into the application of data analytics to develop an appropriate predictive model, utilizing diverse machine learning and statistical techniques, to forecast marine fish production within Brunei Darussalam's aquaculture sector. Employing a machine learning-based algorithm, the study aims to achieve enhanced prediction accuracy, thereby providing novel insights into fish production dynamics. The primary objective of this research is to equip the industry with alternative decision-making tools, leveraging predictive modeling, to identify trends and bolster strategic planning in farm activities, ultimately optimizing marine fish aquaculture production in Brunei. The study employs various time series and machine learning techniques to generate a precise predictive model, effectively capturing the inherent seasonal and trend patterns within the time-series data. To construct the model, the research incorporates notable algorithms, including autoregressive integrated moving average (ARIMA), long short-term memory (LSTM), linear regression, random forest, multilayer perceptron (MLP), and Prophet, in conjunction with correlation analysis. Evaluation of the model's performance and selection of the optimal forecasting model are based on mean absolute percentage error (MAPE) and root mean squared error (RMSE) metrics, ensuring a robust analysis of time series data. Notably, this pioneering research stands as the first-ever attempt to forecast marine fish production in Brunei Darussalam, setting a benchmark unmatched by any existing baseline studies conducted in other countries. The experiment's results reveal that straightforward machine learning and statistical techniques, such as ARIMA, linear regression, and random forest, outperform deep learning methods like MLP and LSTM when forecasting univariate time series datasets.

1. Introduction

Brunei Darussalam stands out as a nation exhibiting one of the most elevated rates of fish and seafood consumption globally, ranging from 19.9 kg to 59.8 kg per capita (Estrebillo and Hiramoto, 2021). Aquaculture represents a viable means of ensuring a consistent supply of fish and seafood products (Arshad et al., 2022). Brunei has witnessed a notable upward trajectory in aquaculture production and revenue generation, surging from 126.46 tons ($1.16 million) in 2000 to 3501.38 tons ($32.35 million) in 2021. The remarkable growth can be primarily attributed to an upswing in shrimp production. Despite the significance of other species such as marine fish, freshwater fish, and crabs in Brunei's aquaculture sector, their production volumes remain comparatively lower than that of shrimp.

This research is dedicated to the application of data analytics techniques, particularly predictive modeling, to enhance marine fish production within Brunei Darussalam's aquaculture industry. Although there has been progress in marine fish production in recent years, the current volume remains insufficient to meet local demand (Okeke-Ogbuafor et al., 2021). The primary impediment to optimizing the fish production process lies in the lack of comprehensive

* Corresponding Author.
Email Address: arif.bramantoro@utb.edu.bn (A. Bramantoro)
https://doi.org/10.21833/ijaas.2023.07.013
Corresponding author's ORCID profile:
https://orcid.org/0000-0003-2772-9427

comprehension regarding data and its capacity to unveil hidden patterns in the available dataset. To bolster fish production optimization, forecasting the production trend for upcoming months or years is instrumental. Time-series data, coupled with machine learning techniques, facilitate learning from historical data to make accurate predictions.

Time series refers to a sequence of values sharing the same statistical indicator arranged chronologically. Predicting a future variable value is attainable when the variable's time-series observation and historical data about the variable are accessible. Past data observations enable the prediction of future variable values, which often rely not only on the variable's past value but also on other temporal variables. Employing precise data analysis and predictive models allows stakeholders to discern how environmental parameters, feed types, feed composition, feeding rates and practices, net changes, and production management practices impact production outcomes (Ouatahar et al., 2021; Petropoulos et al., 2022).

Through the use of suitable machine learning algorithms for time series forecasting and predictive analysis, a deeper understanding of the data is attained, leading to the identification of attribute relationships and consequently enhanced predictions. Machine learning demonstrates a promising approach to achieve forecasts with higher accuracy in comparison to traditional methods.

In Brunei Darussalam, fisheries output still struggles to meet local demand, necessitating reliance on fish supplies from neighboring countries. Effective strategic planning is imperative for the Department of Fisheries to accomplish its production targets. Forecasting future production trends in the months or years ahead empowers the industry to prepare and devise essential strategies to meet these goals. However, challenges arise due to the dearth of proper techniques resulting from limited knowledge and understanding of data, as well as the manual collection of data in the form of Excel files. This hinders decision-makers from identifying hidden patterns within the dataset and deploying appropriate tools and skills for predictive modeling to forecast future production trends effectively. The research endeavors to equip the Department of Fisheries with advanced decision-making tools through predictive modeling. This approach provides valuable insights into the data and enables the accurate forecasting of marine fish aquaculture production in Brunei. By facilitating the analysis and prediction of trends, this research supports strategic planning and action, empowering stakeholders to manage farm activities efficiently and effectively, ultimately leading to increased marine fish production.

2. Background study

The Department of Fisheries was established in 1966 as a constituent of the Ministry of Primary Resources and Tourism in Brunei Darussalam. Its overarching vision entails achieving sustainable growth within the fisheries industry by augmenting productivity and promoting export-oriented strategies. The department's mission focuses on accelerating the growth of the fisheries sector through the application of advanced technologies, fostering increased productivity, and fostering export market opportunities through heightened local and foreign investments (Marsal et al., 2023).

The Department of Fisheries is entrusted with various critical functions and responsibilities, including the management of fisheries resources in adherence to fisheries acts and regulations, conducting fisheries stock assessments, engaging in strategic planning and resource management, conserving fisheries resources, fostering the development of a rational and sustainable aquaculture industry, advancing seafood product development, implementing food safety and quality control measures in the seafood processing sector, and rendering technical and support services to the fisheries industry.

Aquaculture activities were initially introduced to the Department of Fisheries during the early 1970s, with the first facility established at the Sungai Jambu Fish Farm Station in Brunei Muara. Initial research and activities centered around freshwater farming involving species like Carp, Lampam Java, Tilapia, and Gurami, among others. In recent years, the department has successfully implemented several initiatives aimed at boosting production and productivity in the aquaculture industry, harnessing cutting-edge technologies, and expanding the number of fish cages. According to official reports from the Department of Fisheries in Brunei, the total area utilized for aquaculture sites amounts to approximately 19,316.95 hectares. These sites encompass various categories, including marine fish cage culture (18,445 hectares), marine shrimp pond culture (446 hectares), freshwater fishpond culture (55.6 hectares), high-value species aquaculture (47 hectares), hatcheries (12.4 hectares), land-based aquaculture (60.5 hectares), and other areas with potential (285 hectares). Principal locations for marine fish cage culture include Buang Tawar, Pulau Kaingaran, Pulau Pilong-Pilongan, Sungai Dua, Tanjong Pelumpong, Sungai Bunga, Sungai Paku Telisai, Pelong Rock, Nankivell, Victoria, and Littledale, as depicted in Fig. 1.

In Brunei, marine fish aquaculture employs two types of cages: rectangular cages and circular cages. Rectangular cages, as depicted in Fig. 2, are primarily situated in inshore coastal areas and the bay of Brunei. The key advantage of the inshore fish cages lies in their accessibility, as they are conveniently located. Unlike the offshore systems, the inshore setup allows for easy access to saltwater culture without the need for pumping and channeling seawater from the oceans to inland areas. However, these inshore systems have a drawback, as nutrients and wastes are prone to being washed out to the sea, particularly due to their location in shallower waters.

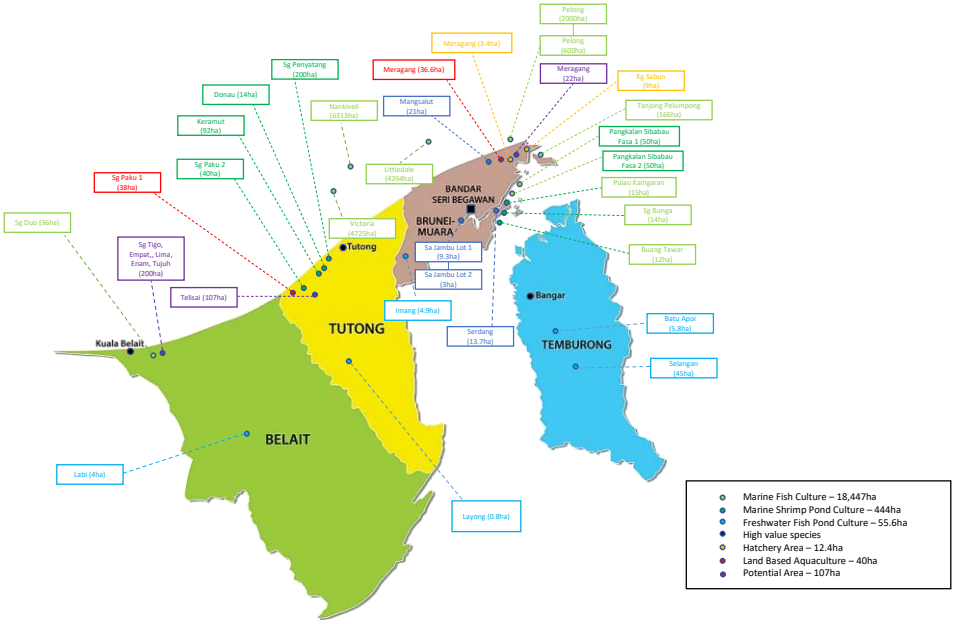

Fig. 1: Brunei aquaculture history

These nutrients and wastes do not readily dissolve, leading to their accumulation on the coastal seabed. This, in turn, heightens the risk of introducing invasive non-native species to the environment and increases the likelihood of diseases spreading among the native fish population. Major inshore fish cage locations include Tanjong Pelumpong, Buang Tawar, Sungai Bunga, and Pulau Kaingaran, encompassing a total area of 292 hectares.

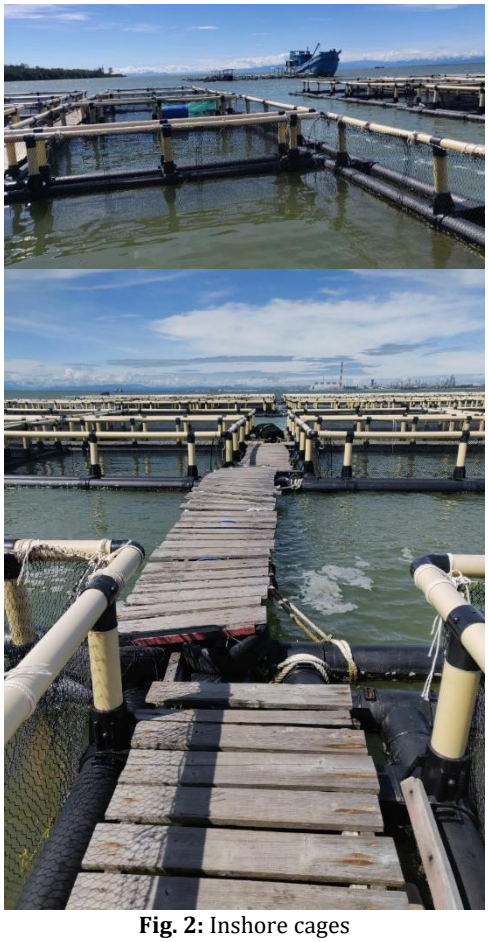

Fig. 2: Inshore cages

The circular cages as illustrated in Fig. 3, are mostly located offshore where the location is excellent for fish culture because this is the natural environment for fish and the water temperature is relatively constant. Each cage has a diameter between 18 and 20 meters, which is ideal for fish culture. One of the disadvantages of using this cage system is that it is highly dependent on weather and climate conditions. Maintaining the cages such as cleaning and monitoring fish health can be challenging due to the size of the cages. It is also prone to outside predators caught in nets and cages. Due to high capital and expensive operation costs, this type of aquaculture system mostly operates by commercial and foreign investors. The main locations for the offshore cages area are Pelong Rock, Nankivell, Victoria, and Littledale with a total area of 18,491 ha.

Fig. 3: Offshore cages

The overall aquaculture production of Brunei in recent years shows a positive trend, increasing from 983 tones (10.2 million) in 2015 to 3501.38 tones (32.35 million) in 2020, as shown in Table 1. The annual production of fish and seafood in the aquaculture sector shows that the production in each year from 2000 to 2014 was not consistent, and the volume of production is still considered low. After a rise in production between 2015 and 2017 and a decline between 2018 and 2019, there was a sharp revival in 2020, when the production started to increase more than twice from the previous year,

from 973.99 metric tons ($10.64 million) to 3501.38 metric ton ($32.35 million).

Table 1: Total production of aquaculture from 2016-2020

Year	Production (Metric ton)	Revenue ($)
2000	126.46	1.16
2001	330.4	3.16
2002	398.84	3.73
2003	617.39	5.06
2004	698.02	5.78
2005	540.24	4.2
2006	551.06	5.49
2007	677.64	6.26
2008	566	4.83
2009	460	4.18
2010	424	4.33
2011	302	3.34
2012	556	5.17
2013	606	5.43
2014	761	7.81
2015	983	10.02
2016	949	9.89
2017	1632.18	16.7
2018	1247.8	13.77
2019	973.99	10.64
2020	3501.38	32.35

Fig. 4 shows the aquaculture production by type of species from 2015 to 2020. The graph clearly indicates that marine shrimp production has been consistently the most significant contributor to the sector, followed by marine fish production. By comparing the value of production for marine shrimp to marine fish production in 2020, there is a huge difference in production between these two species. The production of marine shrimp is about $26.08 million and marine fish production is only about $4.82 million. Looking at the breakdown of aquaculture production, marine shrimp has a five-fold rise in a year, which is one of the main reasons for a vast increase in the aquaculture sector's production value.

The quality of water in which the fish are raised is one of the main factors affecting marine fish growth and production. Grouper production could be maximized by exploiting the physicochemical parameters, such as the temperature, salinity, and carbon dioxide level of the rearing environment (Das et al., 2021). The optimal temperature could be different due to various reasons, such as differences in geographical location. Experimental studies conducted in Indonesia have demonstrated that the optimal temperature for Grouper hatcheries in the Asia-Pacific is around 26°C to 30°C.

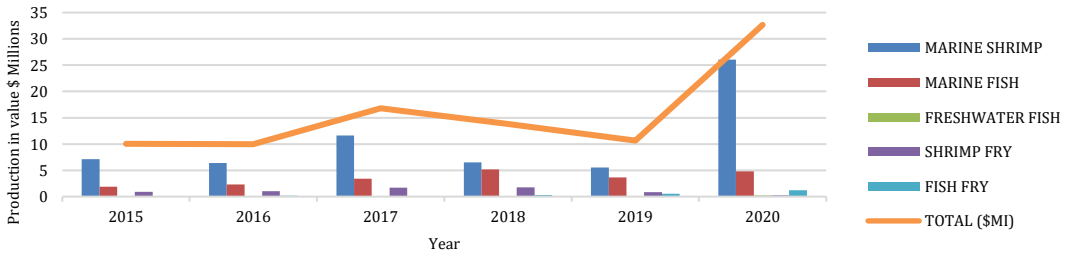

Fig. 4: Aquaculture production by type of species, 2015-2020

Aquaculture organisms are sensitive to both physical and chemical factors such as water temperature, salinity, pH, and dissolved oxygen. Even slight changes outside the optimal conditions may cause stress on organisms, such as reduced food intake, increased energy consumption, and susceptibility to infectious disease, which may lead to the death of aquaculture organisms (Hu et al., 2019). Dissolved oxygen is the most common parameter used to measure water quality. The oxygen needs of fish differ with the type of species and their age. A dissolved oxygen level with a range of 3-6 mg/liter is almost critical for all kinds of fish (Khotimah, 2014). Very low dissolved oxygen levels may affect the survival of aquatic organisms. Hence, there is a requirement to incorporate computerized mechanisms to address the previous issues.

Machine learning is a subset of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. The idea behind machine learning techniques is to teach the computer to find patterns in data using various collections of algorithms, which can be used for future prediction or as a quality check for performance optimization with minimal human intervention (Belyadi and Haghighat, 2021). To the best of our knowledge, machine learning has been replacing traditional statistical methods and mechanistic models and becoming increasingly prevalent in the scientific process in various fields, such as biology (Pratondo and Bramantoro, 2022), social media (Hana et al., 2020), fraud detection (Alraouji and Bramantoro, 2014), and other important sectors that we will explore in the future. The main objective of using machine learning is to practice different algorithms to analyze data, learn from the outcomes, and finally generate prediction accuracy.

Machine learning-based prediction has gained significant traction within the aquaculture sector, demonstrating its applicability in various aspects. Notably, machine learning techniques have been harnessed to forecast coastal algal blooms, a prominent form of marine disaster that poses substantial threats to both marine ecological environments and human well-being (Yu et al., 2021). By employing a gradient-boosting decision tree model, researchers have successfully discerned the influence of environmental factors on phytoplankton concentration, thereby enhancing our understanding of this critical phenomenon.

Moreover, machine learning has been utilized to construct a precise feed intake prediction model, employing sophisticated algorithms such as back-propagation neural networks, genetic algorithms with backpropagation, and mind evolutionary algorithms with backpropagation. This model aims to minimize food waste and optimize profitability in group fish production settings (Chen et al., 2020).

In the realm of marine fish and aquaculture production forecasting, an ensembled machine learning prediction model has been developed to enhance prediction accuracy. Employing vector regression techniques, this ensemble model effectively combines three distinct machine learning approaches: Linear regression, random forest, and gradient boosting. Notably, this ensembled model outperforms individual machine learning models, as demonstrated by Rahman et al. (2021).

Overall, these advancements illustrate the widespread adoption and efficacy of machine learning methodologies in the aquaculture domain, fostering valuable insights and driving improvements in various aspects of aquaculture management and production.

Time series forecasting involves collecting and analyzing past observations to develop a model to extrapolate an observation in the future (Castán-Lascorz et al., 2022). The area of forecasting, which is part of predictive analytics, has been receiving tremendous interest from several researchers around the world. Those researchers use time-series exploration to forecast and predict the future trend of time-series data over the upcoming years. Machine learning methods such as artificial neural network (ANN), recurrent neural network (RNN), long short-term memory (LSTM), convolutional neural network (CNN), and gated recurrent units also can be used to address production forecasting issues. Another use of time series forecasting has been proposed in agriculture to support its precision (Bramantoro et al., 2022). The forecasting has been combined with statistical linear regression and other machine learning techniques, such as decision tree and artificial neural networks to predict the paddy yields in Brunei Darussalam. The result of the analysis is then correlated with weather parameters, such as rainfall, wind speed, and temperature. Based on this experience, it is believed that other machine learning techniques have potential benefits for aquaculture, which is closely related to agriculture.

Shen et al. (2021) proposed LSTM based deep learning model to predict the time series-based stock prices, and a new regression scheme was implemented on LSTM, based on a deep neural network. Dubey et al. (2021) developed a model that is used to estimate the behavior and the trend of wheat production in Haryana using Box-Jenkins, autoregressive integrated moving average (ARIMA), and ANN. During the evaluation, ARIMA and ANN were found good enough for modeling and forecasting the production behavior of wheat in Haryana, the northern part of India. In their recent work, Fan et al. (2021) introduced a pioneering hybrid model designed to forecast well production with precision and efficiency. This model plays a crucial role in extending the well's life cycle and enhancing reservoir recovery, both of which are critical objectives in the domain of oil and gas exploration and production. The model integrates ARIMA and the LSTM, which are considered advantageous for both linearity and nonlinearity data. The testing results of the three actual wells show that the performance of hybrid models is better and more reliable than the individual traditional ones. Another technique to forecast time series is by using generalized regression neural networks (GRNN) proposed by Martínez et al. (2022). This model demonstrates the capability to generate rapid and exceptionally precise forecasts, adeptly capturing both seasonal and trend patterns inherent in the time series data. Particularly, the GRNN is remarkably efficient and requires swift training to achieve accurate predictions, rendering it a favorable candidate for developing rapid time series forecasting models.

Furthermore, the ARIMA model was employed for modeling and forecasting the captured fishery and aquaculture production trend in the Iranian aquaculture sector (Hülya and Abdallah, 2021). The ARIMA model's suitability arises from its aptness for handling time series data in the context of Iran's aquaculture domain.

3. Methodology

Fig. 5 depicts the comprehensive workflow of the research process, commencing from its inception to its culmination. The initial stage involves the collection of datasets from the Department of Fisheries, which were made available in the form of Excel files and hardcopy. To gain a comprehensive understanding of the research's scope, an online meeting was conducted with the Department of Fisheries, facilitating discussions and a question-and-answer session to gather their requirements and ensure that the research outcomes align with their needs. During this meeting, an overview of the aquaculture sites and production in Brunei over recent years was presented, aiding in discerning the primary objectives and anticipated outcomes of the research.

Two officers from the Department of Fisheries, specializing in statistics and international affairs, and the aquaculture industry division, actively participated in the research. Their primary responsibilities involved providing pertinent information concerning Brunei's aquaculture sites and activities. Subsequently, a second meeting was conducted to further gather data and understand unique characteristics within the dataset. During this meeting, the officers provided additional explanations regarding the datasets, thereby assisting in addressing various challenges, such as missing values and incomplete data.

The third meeting was convened to delve into insights derived from data analysis. This meeting

focused on understanding several factors influencing production levels, exploring the reasons behind sudden production fluctuations in specific years, and examining how such dynamics could impact prediction accuracy. Efforts were made to comprehend the data in light of the aforementioned issues, with discussions centered around improving the research's analytical approach.

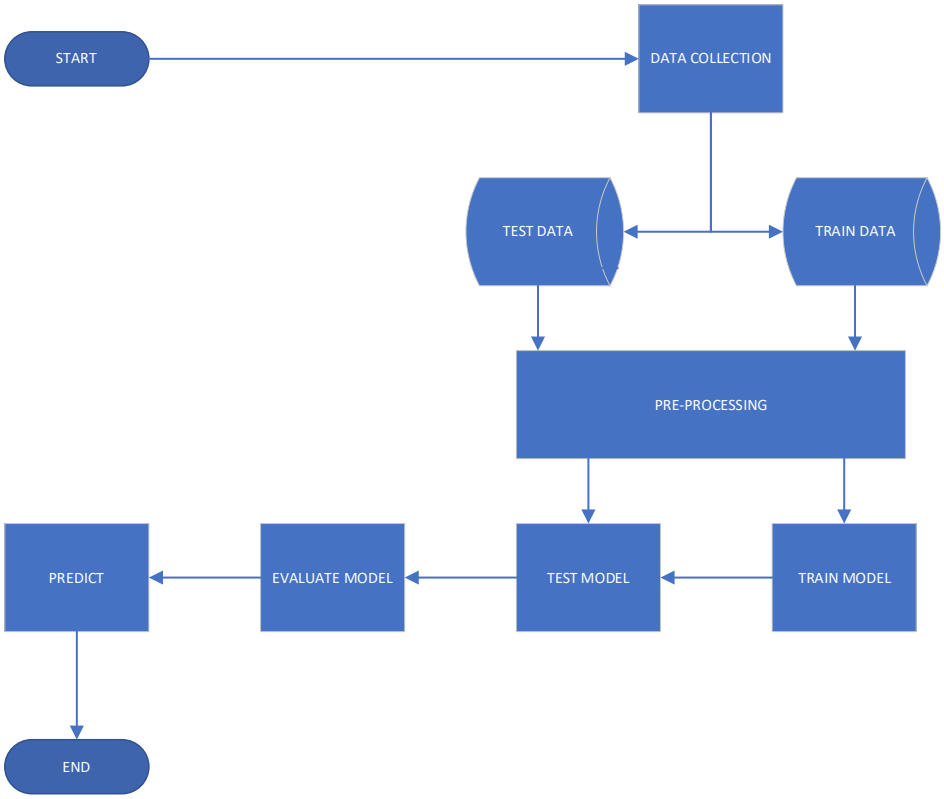

Fig. 5: System workflow

Handling missing values in time-series data is a common challenge attributed to sensor failures, data collection errors, or resource limitations. Imputing missing time-series values assumes paramount significance in time series forecasting endeavors, as it ensures the predictive outcomes are maximally accurate. In the data pre-processing stage, various tasks are undertaken to prepare the dataset before constructing the forecasting model. This preparation phase holds crucial importance, as it significantly influences the model's performance, accuracy, and resultant outcomes. The stage is subdivided into several sub-tasks aimed at organizing the received data from the Department of Fisheries, which predominantly exists in unstructured form across multiple files and datasets. The initial task involves combining all relevant files to structure the entire data cohesively into a singular dataset, thus facilitating a comprehensive understanding and analysis of the data structure, a prerequisite for model development.

Furthermore, data received from the Department of Fisheries may encompass information irrelevant to the current research, warranting the removal of such extraneous data that may obfuscate forecast results. Additionally, the elimination of duplicated data, such as redundant operator and location names, is essential to ensure the presence of only unique instances in the dataset and prevent outliers from impacting prediction results. Addressing missing values is a prevalent issue during the development and evaluation of prediction models, prompting the application of various techniques, including statistical methods such as mean, median, and mode, to tackle these gaps effectively.

The process of predicting future production trends in Brunei's aquaculture sector necessitates the adoption of appropriate modeling techniques to discern relevant patterns and trends that align with the research's objectives. Upon selecting suitable machine learning algorithms to experiment on the dataset, the model undergoes testing and evaluation to ensure realistic performance estimations on unseen data. Ensuring the appropriate segregation of the dataset into training and testing sets is paramount to gauge the model's generalizability and efficacy. It is vital to refrain from utilizing the test set during algorithm selection and model training, as this can lead to over-optimistic evaluation results that do not reflect the model's actual performance (Zheng, 2015).

The data are partitioned into two distinct subsets: the training set utilized to train an initial array of models, and the testing set deployed to determine the algorithm that generates the best model. The testing data also facilitate estimating the generalized performance of the final model on unseen data. Equally important is the careful selection of appropriate evaluation metrics to assess model accuracy during testing. In this research, the root

mean squared error (RMSE) and mean absolute percentage error (MAPE) are employed as performance metrics to guide the selection of the most suitable model for prediction.

4. Analysis and design

Estimating the future trends of production is extremely useful for management purposes. It provides possible directions and the effectiveness of the current plan to reach a goal. This research includes several steps in developing a model to predict and analyze future production trends in upcoming years. This also helps to understand and learn the outcome of the results and make a prediction for better decision-making. It can eventually enhance the productivity and efficiency of marine fish production.

The precision of the method in data gathering is one of the crucial steps that should constantly be scrutinized to ensure the best prediction results. In this research, we use datasets provided by the Department of Fisheries in Brunei that consists of the production of marine fish performed by several operators in recent years, the area, the total number of cages used for marine fish farming, the monthly target production from 2016 to 2021, the production values for the marine fish from 2020 onwards, and the water quality data for several locations of Brunei aquaculture sites in 2020. Fig. 6 shows the statistics for the dataset, such as count, mean, standard deviation, minimum-maximum, upper quartile, lower quartile, and median values for each attribute.

	count	mean	std	min	25%	50%	75%	max
year	3153.0	2017.981922	1.406562	2016.0	2017.000000	2018.00000	2019.00	2021.000000
target_production	3120.0	7385.900699	40973.924726	0.0	509.651369	765.00000	1700.00	321147.000000
no_of_cage	2460.0	111.275610	338.250695	0.0	34.000000	45.00000	86.00	4578.000000
no_of_cage area	2268.0	12.789502	145.115748	0.0	0.200000	1.00000	1.00	2000.000000
no_of_operating_cage	2472.0	61.680320	87.461658	0.0	15.000000	35.00000	72.00	756.000000
no_of_operating_cage area	2152.0	2.264996	12.593660	0.0	0.0240000	0.07965	1.00	150.000000
target_value	609.0	11468.162671	21366.692279	0.0	2040.000000	3060.00000	7650.00	92759.116667
actual_value	609.0	8108.112614	44931.327868	0.0	0.000000	399.70000	4070.00	880690.720000
temp	410.0	29.636976	0.620382	28.6	29.100000	29.60000	30.10	33.000000
ph	410.0	6.996317	0.503587	6.0	6.510000	6.96500	7.49	7.850000
salinity	946.0	20.342283	6.510497	6.0	16.200000	20.23500	26.00	26.00

Fig. 6: Statistics of the attributes

In the preliminary stages of constructing a model, data preparation assumes a pivotal role in rendering the dataset suitable for machine learning algorithms, thereby yielding the desired outcomes. Data cleansing is an essential step, involving the removal of irrelevant information, addressing outliers, duplicated data, and missing values. The quality of the training data significantly impacts the model's performance, making it imperative to employ various pre-processing techniques such as aggregation, dimensionality reduction, sampling, feature creation, and feature transformation to rectify data errors.

The result of the data cleansing process is a refined dataset encompassing all pertinent attributes, features, and input parameters required for our model. Within the dataset, data objects may exhibit duplications or near-duplications, necessitating their detection and elimination. Resolving such duplications ensures that identical operators and sites bearing different names are unified under a single unique name, employing basic natural language processing techniques. As evidenced in Table 2, multiple instances of identical operators and sites with varied names have been identified in the dataset.

In real-world datasets, it is common for objects to have one or more missing attribute values. Handling missing values in datasets used for machine learning is of paramount importance, particularly when utilizing the complete available data is crucial (Abyaneh, 2014). There are various approaches to address missing values. The simplest method involves removing all rows and columns containing missing values or replacing them with zeros. However, such removal significantly reduces the depth of data learning and consequently impairs model accuracy. Elhassan et al. (2022) note that missing value removal is generally considered the least favorable approach in most machine learning techniques. Alternatively, several methods involve calculating the missing or incomplete attribute values to ascertain the results. Fig. 7 illustrates the count and percentage of missing values in the dataset.

Table 2: Data duplicates

Operator	Site
Batriza Ent.	Sungai Bunga
Batriza Enterprise	Sungai Bunga
Ding Dong Farms	Buang Tawar
Ding Dong's Farm Sdn Bhd	Buang Tawar
Sykt Nakoda Emas Aquaculture	Tanjong Pelumpong
Sykt Nakhoda Emas Aquaculture	Tanjong Pelumpong
Syarikat Harvesea	Tanjong Pelumpong
Syarikat Harvesea Enterprise	Tanjong Pelumpong
Sykt Berkat Aquaculture Enterprise (Sykt Dyg Hasimah Binti Haji Mohd)	Tanjong Pelumpong
Sykt Berkat Aquaculture Enterprise	Tanjong Pelumpong
Koperasi Kg Pudak Sdn Bhd	Sungai Bunga
Koperasi Kg Pudak Bhd	Sungai Bunga
Sha-Zan Ent.	Sungai Bunga
Sha-Zan Enterprise	Sungai Bunga
Syarikat Namara Aquaculture (Hazmi H.M. Aquaculture)	Sungai Bunga
Syarikat Namara Aquaculture	Sungai Bunga

In this study, the treatment of missing values involves replacing them with zeros under specific conditions, as encoded in Fig. 8. During the meeting with the Department of Fisheries, it was agreed to

employ this approach due to instances where several operators were unable to initiate production in certain years. Additionally, a statistical approach is utilized to address the missing values by imputing numerical columns with the mean of the remaining values in the same column, as demonstrated in Fig. 9. This method ensures the maintenance of data integrity while enabling a valid estimation of missing values in the dataset.

As coded in Fig. 10, for the rest of the missing values that cannot be replaced with the mean value, we use estimated values calculated with the two equations:

$$NO\ OF\ OPERATING\ CAGE\ AREA = \frac{NO\ OF\ OPERATING\ CAGE}{NO\ OF\ CAGE} \times NO\ OF\ CAGE\ AREA \quad (1)$$

$$NO\ OF\ CAGE\ AREA = \frac{NO\ OF\ CAGE}{NO\ OF\ OPERATING\ CAGE} \times NO\ OF\ OPERATING\ CAGE\ AREA \quad (2)$$

df.isnull().sum()	
operator	0
site	0
year	0
month	0
target_proluction	33
actual_production	59
no_of_cage	693
no_of_cage area	885
no_of operating_cage	681
no_of operating_cage_area	1001
target value	2554
actual value	2544
Cumulative_value	2543
temp	2743
salinity	2207

Fig. 7: Missing value

There are several events that might influence the forecast during model training. This can determine whether the prediction is successful or not. It is important to choose useful data and features for the dataset. Flooding irrelevant data and features can cause difficulty during the model training process, which is why sometimes our model is unable to perform well. One way to handle this is by removing irrelevant data in our dataset that might not influence the forecast. Figs. 11 and 12 show a few line codes to drop several rows and columns that might not be related to the forecast.

Fig. 13 shows the trend of marine fish production from 2010 to 2021. It can be inferred from the graph that the production fluctuated from 2010 to the end of 2018; however, the overall production shows a positive trend during this period. The trend started to decrease dramatically from December 2018 to February 2020. Then it rose back steadily until January 2020, before dropping drastically until April 2020. From July 2020, it increases rapidly until it reaches the highest production in August 2020, which is around 90,000 kg. However, the production slowly decreased in early 2021. The main reason why there is a sudden decrease in marine fish production is that the fish is still in a growing period and has not reached the market size yet. Escaping from the cage could be another factor why the production is very low compared to other months. The COVID-19 pandemic also significantly impacted marine fish production between 2020 to 2021 which explains why the production was relatively low during that period.

```
#fill all empty space in actual columns with 0
df = df.fillna({'actual_production':0})
df = df.replace(to_replace='-', value = 0)

df.loc[(df['actual_production'] == 0) & (df['no_of_operating_cage'] == 0) ,
          'no_of_operating_cage_area'] = 0
df.loc[(df['actual_production'] == 0) & (df['no_of_operating_cage_area'] == 0) ,
          'no_of_operating_cage'] = 0
```

Fig. 8: Replacing missing value with zeros

```
#replace the missing values with the mean values groupby operator
df['no_of_cage'].fillna(df.groupby(['operators'])['no_of_cage'].transform('mean'),
                        inplace=True)
df['no_of_cage_area'].fillna(df.groupby(['operators'])['no_of_cage_area'].transform('mean'),
                             inplace=True)
df['no_of_operating_cage'].fillna(df.groupby(['operators'])['no_of_operating_cage'].transform('mean'),
                                  inplace=True)
df['no_of_operating_cage_area'].fillna(df.groupby(['operators'])['no_of_operating_cage_area'].transform
                                       ('mean'), inplace=True)

#replace the missing values with the mean values groupby site and year
df['temp'].fillna(df.groupby(['site','year'])['temp'].transform('mean'), inplace=True)
df['ph'].fillna(df.groupby(['site','year'])['ph'].transform('mean'), inplace=True)
df['salinity'].fillna(df.groupby(['site','year'])['salinity'].transform('mean'), inplace=True)
df['do'].fillna(df.groupby(['site','year'])['do'].transform('mean'), inplace=True)
```

Fig. 9: Replacing the missing values with the mean

```
df['no_of_operating_cage_area'].fillna(value=(df['no_of_operating_cage']/
                                              df['no_of_cage'])*df['no_of_cage_area'],
                                       inplace=True)
df['no_of_cage_area'].fillna(value=(df['no_of_cage']/
                                    df['no_of_operating_cage'])*df['no_of_operating_cage_area'],
                             inplace=True)

df.loc[(df['operators'] == 'PELONG ROCKS 1 (300 Ha)_HISEATON SDN BHD' ) & (df['year'] == 2018) ,
          'no_of_operating_cage'] = df['no_of_cage']/2
df.loc[(df['operators'] == 'PELONG ROCKS 1 (300 Ha)_HISEATON SDN BHD' ) & (df['year'] == 2018) ,
          'no_of_operating_cage_area'] = df['no_of_cage_area']/2
```

Fig. 10: Replacing the missing values with the estimated values

```
#drop columns
df = df.drop(columns=['cumulative_value'])
df = df.drop(columns=['target_value'])
```

Fig. 11: Column removal

```
# drop rows
df= df[df.operators != 'MY INVESCO']
```

Fig. 12: Row removal

The charts from Figs. 14 to 17 show the correlation between the annual total marine fish production and several hypothesized factors, such as number of the cage, number of cage areas, number of the operating cage, and number of operating cage areas from 2015 to 2019, respectively, to determine whether all these factors could affect the marine fish production or not. This is also important for justifying the prediction result at the end of the analysis. The lines represent the trend of annual marine fish production, and the columns are for the selected factor. In addition, the correlation is not completely analyzed because the factor datasets lack two years compared to the production dataset.

Fig. 14 shows the correlation between the annual marine fish production trend and the total number of cages from 2015 to 2019. From 2015 to 2018, the production and cage numbers have increased steadily. In 2019, the production started to decrease as the number of cages decreased also. A similar trend appears for the number of cages areas as illustrated in Fig. 15. The production and the number of cages area also increased steadily from 2015 to

2018. However, the number of cages areas increased drastically while the production started to decline in 2019. This sudden increase in the number of cages area is because of the new operators, Hiseaton Sdn bhd, located in Pulau Pilong-Pilongan with a total area of 2,000 ha. The operators have just started the operation, which explains why they have not met their target production in that year. The marine fish culture in a cage usually takes from eight to nine months per cycle before achieving its market size.

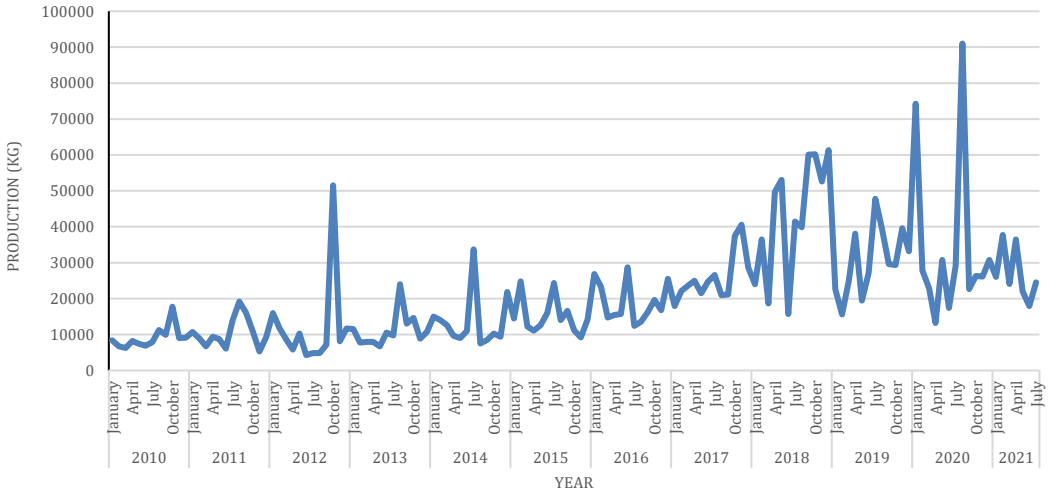

Fig. 13: Total marine fish production 2010-2020

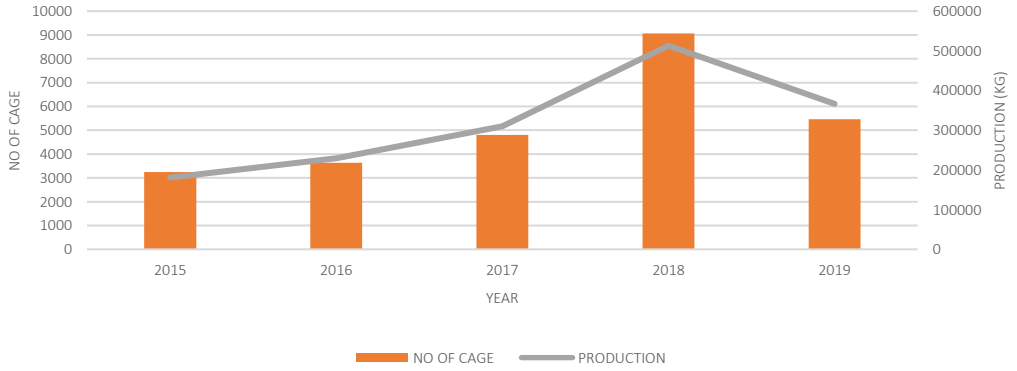

Fig. 14: Number of cages and total production

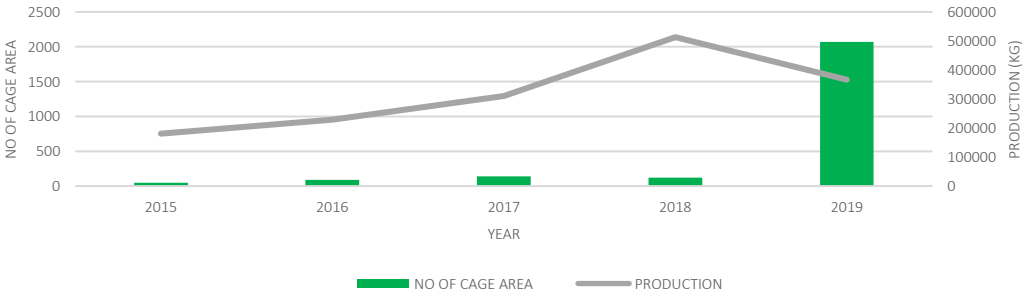

Fig. 15: Number of cage areas and total production

Fig. 16 shows the correlation between the annual marine fish production and the number of operating cages from 2015 to 2019. The total production and the number of operating cages increased steadily from 2016 until 2018. The total number of operating cages increased from 1700 units to 3975 units in 2018. In 2019 the number of operating cages decreased from 3975 units to 1662 units, and the total production also started to decrease. Fig. 17 shows the correlation between the number of operating cage areas and the total production. The number of operating cage areas continued to increase from 13.04 ha in 2015 to 261.64 ha in 2019. The production also increased at a steady rate from 2015 until 2018. However, production started to decline in 2019, even though the number of operating cages was much higher than the previous year. There are several possible factors to explain this anomaly, such as escaping fish from the cage and heavy rainfall that causes a change in the water quality.

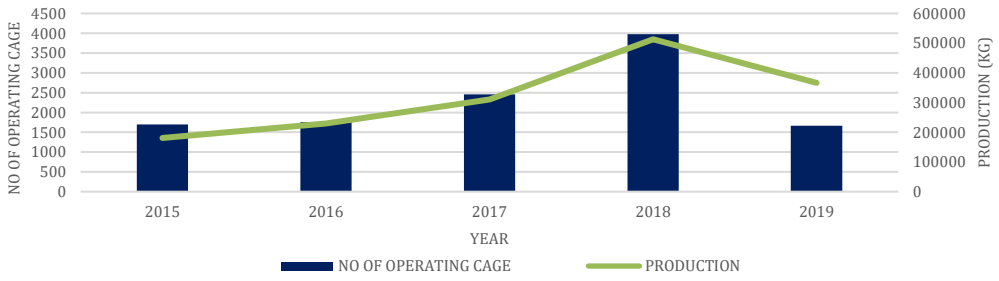

Fig. 16: Number of operating cages and total production

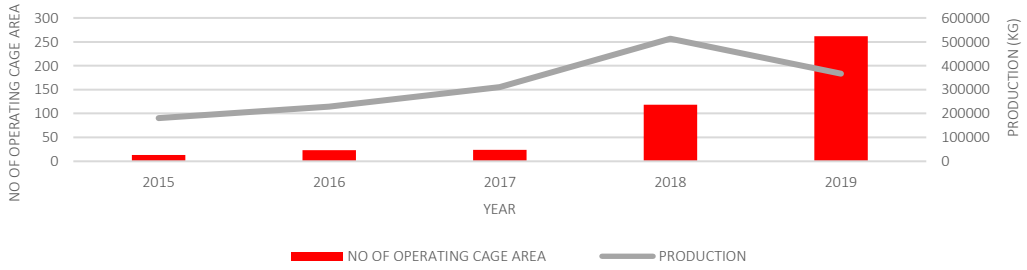

Fig. 17: Number of operating cages area and total production

Water quality represents a consequential determinant affecting marine fish production. Regrettably, the water quality dataset is solely accessible for the year 2020, and it is dispersed across five distinct locations, namely Buang Tawar, Pulau Kaingaran, Sungai Bunga, Sungai Dua, and Tanjong Pelumpong.

Figs. 18 to 22 present charts depicting the monthly average of seawater temperature, salinity, and pH at five distinct locations: Buang Tawar, Pulau Kaingaran, Sungai Bunga, Sungai Dua, and Tanjong Pelumpong. Analyzing these charts enables an assessment of whether water quality influences marine fish production. If a correlation between these two factors is evident, it would warrant further investigation and analysis. Upon overall examination, it becomes apparent that both temperature and pH significantly impact marine fish production. An increase in seawater temperature and pH correlates with a decline in marine fish production, while a decrease in these parameters corresponds to an increase in production. Similarly, water salinity also plays a role, as consistent salinity levels lead to steady production, whereas drastic drops in salinity below the optimum level result in decreased production. Proper water quality management is vital to prevent stress and mortality among marine fish, which may be contributing factors to production declines.

The research papers reviewed corroborate the substantial impact of water quality on marine fish production. However, it is essential to recognize that this study is based solely on monthly water quality data from one year of observations, resulting in a dataset comprising only 12 observations, rendering it less suitable for predictive forecasting. Figs. 18 to 22 specifically depict water quality and total production at Buang Tawar, Sungai Bunga, Sungai Dua, and Tanjong Pelumpong in the year 2020.

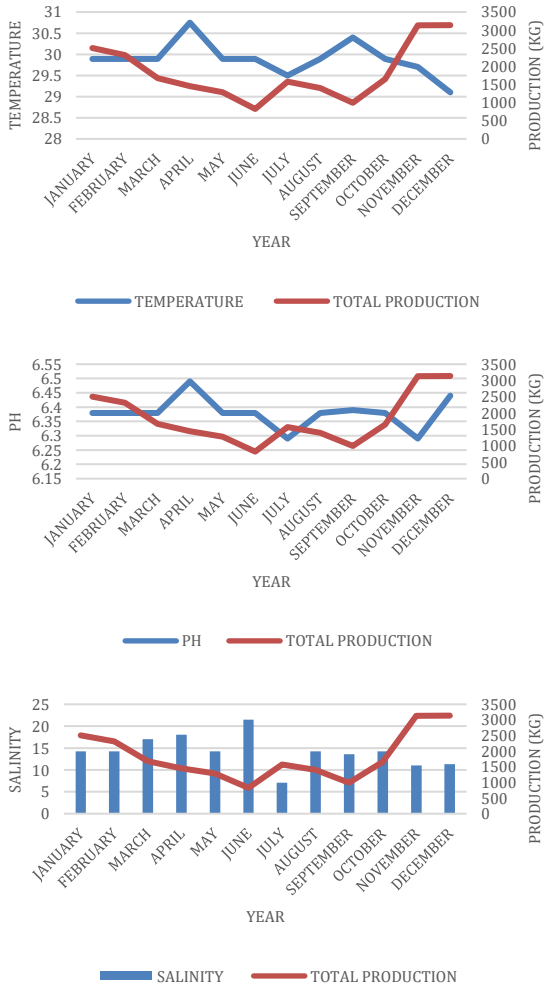

Fig. 18: Buang Tawar water quality

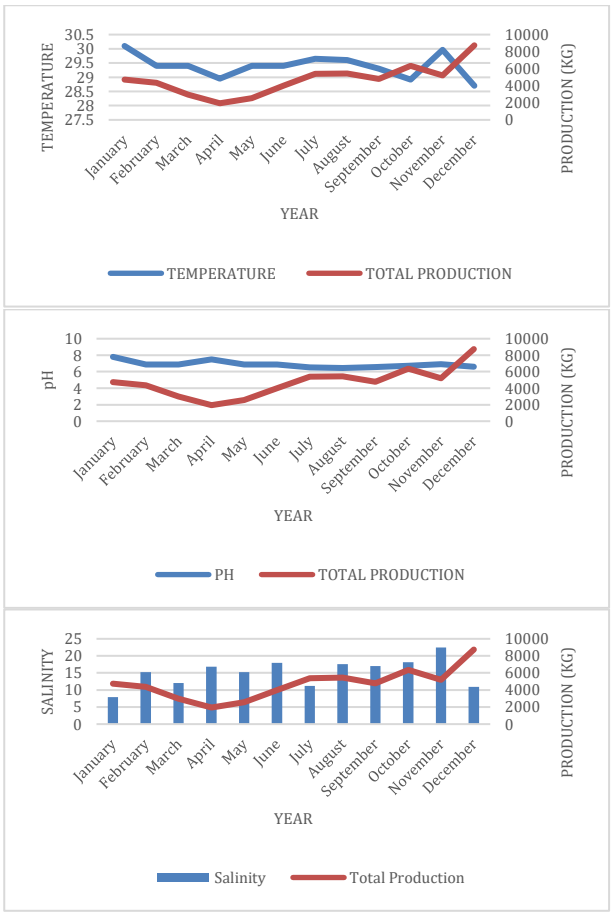

Fig. 19: Pulau Kaingaran water quality

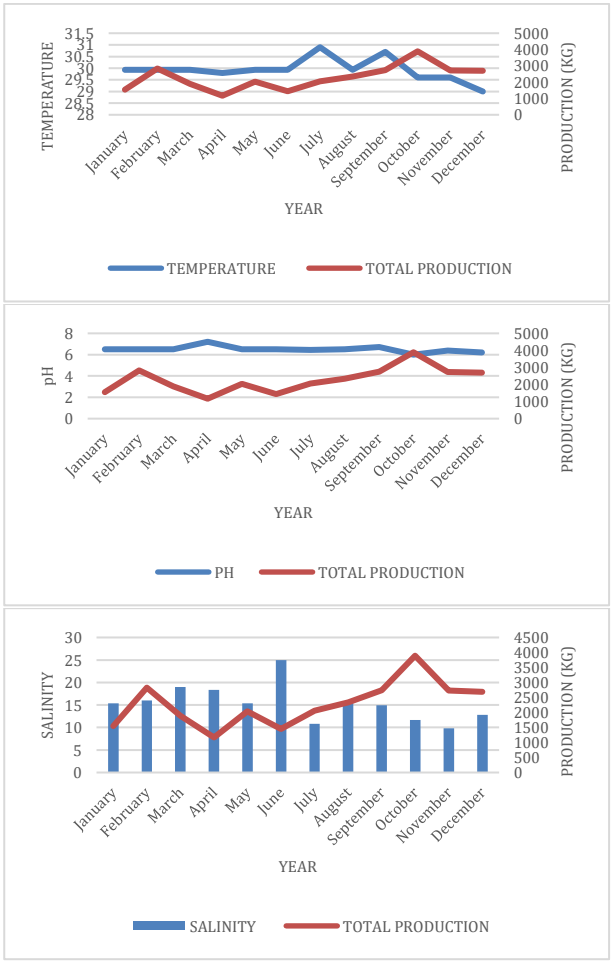

Fig. 20: Sungai Bunga water quality

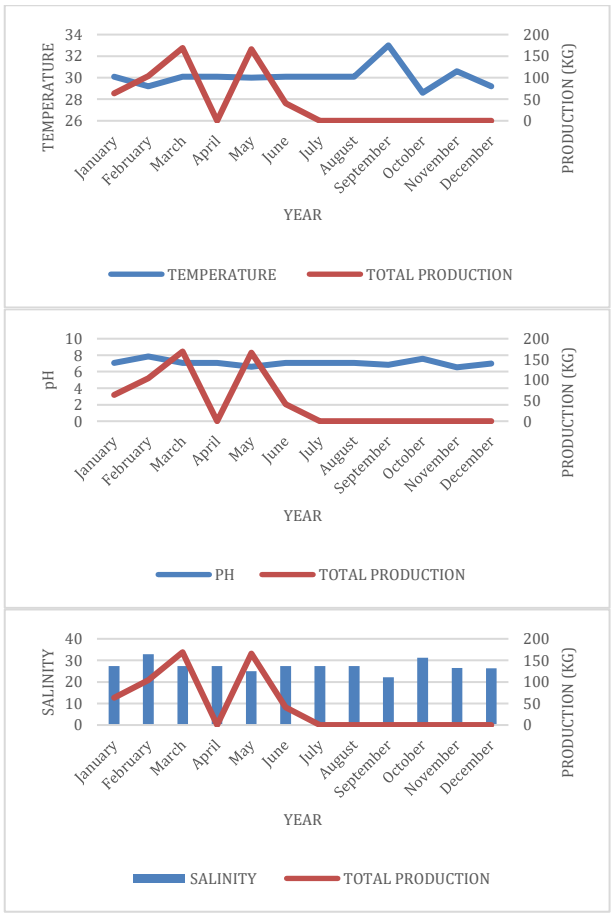

Fig. 21: Sungai Dua water quality

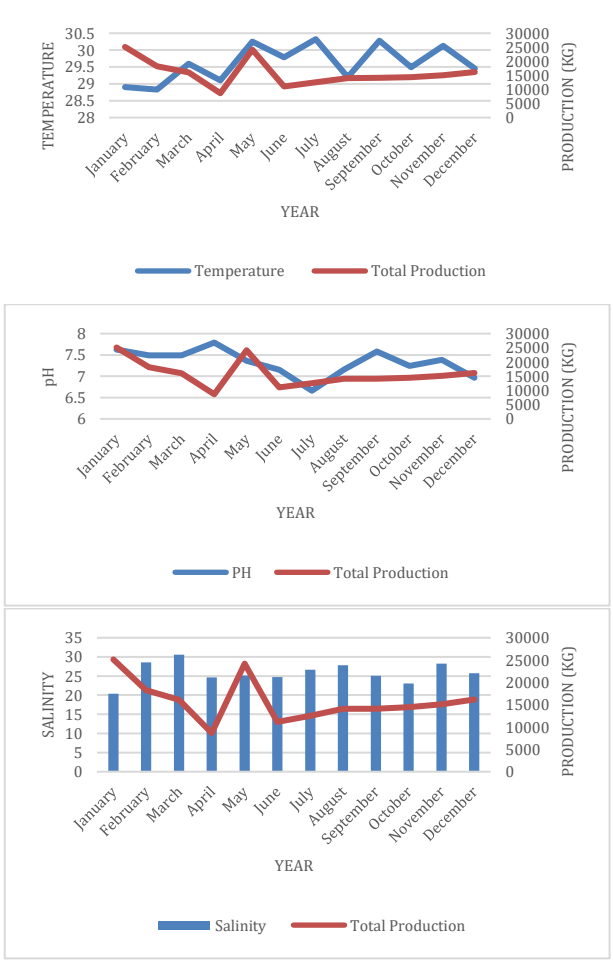

Fig. 22: Tanjong Pelumpong water quality

Univariate time series forecasting focuses exclusively on predicting a single target variable based on historical input of said target variable (Elsayed et al., 2021). In the present study, only actual production data serve as the target variable for time series forecasting. The dataset spans from January 2010 to July 2021, encompassing 139 observations. These data are partitioned into separate training and testing sets, with approximately 85 percent utilized for model training and the remaining 15 percent reserved for model testing. This ratio is consistently maintained across all locations and models.

To ascertain data stationarity, the Augmented Dickey-Fuller (ADF) test is employed as a statistical evaluation method. This test determines the extent to which a null hypothesis should be either rejected or accepted, indicating the stationarity status of the data. A threshold of 0.05 is used to determine whether the null hypothesis is rejected or accepted. If the p-value resulting from the ADF test is less than the threshold (0.05), the data are deemed stationary, and the null hypothesis can be rejected. Conversely, if the p-value exceeds the threshold, the data are considered non-stationary, and the null hypothesis cannot be rejected (Satrio et al., 2021). The code and results of the ADF test are presented in Fig. 23.

```
from statsmodels.tsa.stattools import adfuller
def ad_test(dataset):
    dftest = adfuller(dataset, autolag='AIC')
    print("1. ADF: ",dftest[0])
    print("2. P-Value: ",dftest[1])
    print("3. Num of Lags: ",dftest[2])
    print("4. No of observation for ADF and Critical Values calculation :",dftest[3])
    print("5. Critical Values :")
    for key, val in dftest[4].items():
        print("\t", key, ": ",val)
```

```
1. ADF:  -1.6098702384368089
2. P-Value:  0.4786336854895145
3. Num of Lags:  6
4. No of observation for ADF and Critical Values calculation : 132
5. Critical Values :
        1% :  -3.4808880719210005
        5% :  -2.8836966192225284
        10% :  -2.5785857598714417
```

Fig. 23: ADF test

The log scale transformation serves as a method to convert non-stationary data into stationary form. Its application aids in stabilizing data series characterized by non-constant variance, consequently leading to the attainment of a normal distribution within the series. This transformation involves calculating the logarithm of the numerical values present in the dataset. However, due to the presence of null values in the dataset, log-scaling can yield infinity values. To address this concern, the log (x+1) transformation is employed, ensuring the avoidance of null values during the log transformation process.

One straightforward technique to eliminate data trends involves differencing. This entails subtracting the current observation (t) from the observation recorded at the previous time step (t-1). Such differencing operations effectively contribute to detrending the data, resulting in a more stationary data series (Ramazan, 2019). Differencing the data makes the data easy to analyze. Fig. 24 shows the dataset before differencing that creates non-stationary time series and Fig. 25 shows the dataset after differencing that creates stationary time series.

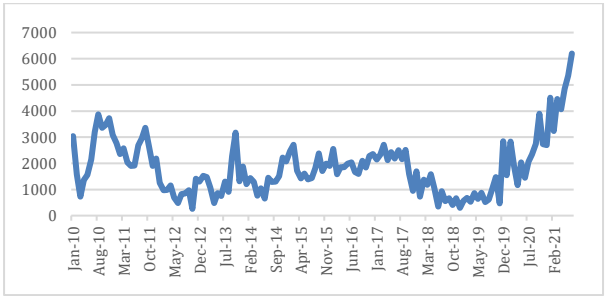

Fig. 24: Non-stationary Time series

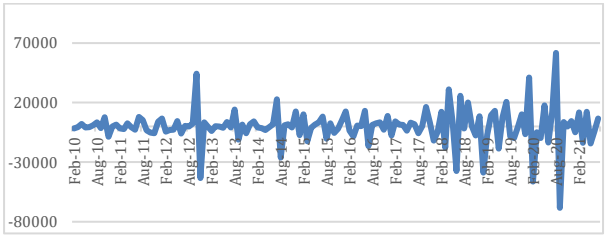

Fig. 25: Stationary Time series

The window or lag method is often used to convert time series problems into supervised learning problems. The number of previous time series is called window width or lag size. The rolling window is used to predict the next time steps using window width or lag size (Zhou et al., 2020). Before applying a machine learning algorithm, the time-series data need to be reconstructed. This is because the sequence of observations can highly affect the result of the forecast (Ramazan, 2019). To reconstruct the dataset, window width or lag size is used as the input data for the selected model.

Several prospective algorithms have been identified to construct a predictive model for forecasting the trend of marine fish production in Brunei. In this research, a comparative analysis of the algorithms' performances is conducted, focusing on specific locations. The chosen algorithms include ARIMA, Prophet, Linear Regression, Random Forest, LSTM, and Multi-Layer Perceptron (MLP).

These algorithms have been selected based on their effectiveness in addressing time series forecasting problems, each offering unique approaches to enhance model performance and achieve desired outcomes, often by parameter tuning to attain optimal model configurations.

ARIMA, a widely used method for time series dataset forecasting, is distinguished by its capacity to yield exceptional forecasting results with high accuracy. It adeptly handles small datasets and is an amalgamation of two constituent models: autoregressive (AR) and moving average (MA). The integration aspect, represented by 'I' in the ARIMA model, unites these two models. ARIMA can be effectively employed to model and forecast both seasonal and non-seasonal data. For successful application of the ARIMA model, the time series dataset must exhibit stationarity, with constant

mean, variance, and autocorrelation across the time series.

The model has three parameters: *p* for the autoregressive lags, *q* for the moving average, and *d* for the order of differentiation. To predict the dataset using the ARIMA model, we have to find the rolling statistics of the dataset. Then we perform the Dickey-Fuller test, which is used to test the null hypothesis that a unit root is present in an autoregressive time series model. The Dickey-Fuller test is associated with the trends of the dataset. Then the auto correlation (ACF) and partial autocorrelation (PACF) as illustrated in Figs. 26 and 27 are used to determine the values *p* and *q*. The best performance of the model in the prediction of test data is chosen based on searching all possible values of the parameters that have the best predictability for the production time series. A simpler estimation parameter method is AUTO ARIMA. It is a library that automatically searches for the best value of *p*, *d*, and *q* based on Akaike's information criteria and Bayesian information criteria.

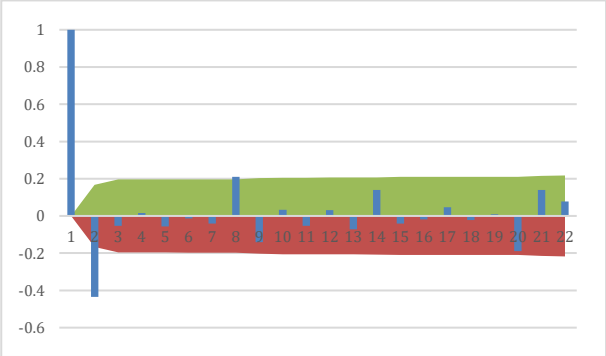

Fig. 26: ACF plot

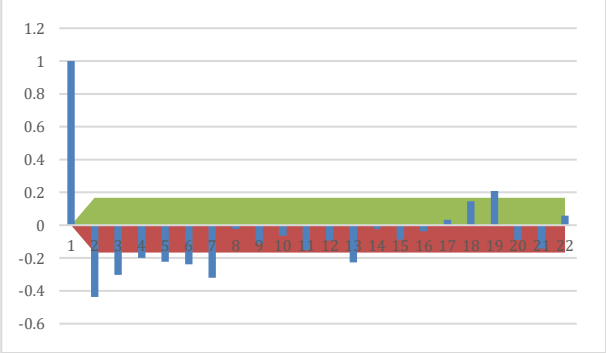

Fig. 27: PACF plot

Prophet is one of the simplest time series forecasting methods because it is easy to implement by any analyst, including people with no prior knowledge of forecasting. Prophet is an open-source library created by Facebook that is used to build a model to forecast time-series data. According to Facebook, Prophet works effectively on time series data with strong seasonal effects and historical data, in addition to being robust to the outliers and shifts in trends (Satrio et al., 2021). The model is implemented by creating a data frame using the dataset with two columns: Data stamp in a DateTime format, and the forecasting measurement in numerical values. The hyperparameters are tuned automatically by using default parameters. After implementing the model, the forecasting result will provide a data stamp and a column that contains the forecasted results in the historical data frame.

Linear regression is one of the most popular machine learning models widely used in various fields, including in time series forecasting. One of the main reasons for using the linear regression model is that it is easy to interpret the relationship between inputs and output in a linear model, making it more robust for prediction. To use the linear regression model, the time series data need to be converted first into supervised learning, where there must be an input and output. This can be done by shifting the dataset because the supervised machine learning methods use historical time series data for the input of the model (Pavlyshenko, 2019).

The random forest can be referred to as an ensemble learning method that belongs to the supervised learning technique. This machine-learning algorithm can be used in time-series forecasting by creating lag and seasonal component variables from the respective dataset. Random forest performed well for most time-series data, especially for intermittent data because it can handle the probability of zero production or demand (Kane et al., 2014). The random forest model can be generated using the random forest package available for the Python environment. Similar to linear regression, the time-series dataset needs to be converted into supervised learning by shifting the dataset to make it suitable for the machine learning format.

LSTM is a kind of recurrent neural network that has the capability of remembering the value at the earlier stages when the values can be used for the future. The memory scheme of LSTM provides three types of gates: Forget gates, input gates, and output gates (Dubey et al., 2021). LSTM is an efficient algorithm for building a time series sequential model. Another advantage of using Recurrent LSTM networks is that they can address the limitations of traditional time series forecasting techniques by adapting nonlinear time series data. LSTM model usually contains four neural networks: The first layer is the sigmoid neural net layer (0 and 1), the second layer is the layer that shows cell state, the third layer is the language layer which is used to drop the information, and the output layer (Dubey et al., 2021). MLP is a feed-forward artificial neural network that uses a supervised learning technique called back-propagation. It consists of an input layer, at least one hidden layer, and an output layer; in which each layer is made up of multiple nodes (Taud and Mas, 2018). MLP has been successfully used to handle missing time-series values and make accurate predictions across different applications that have various complexity. It is crucial to select the best hyperparameters as model architecture to ensure the performance of MLP during the development of the model. There are several ways to optimize the predictive performance of MLP, such as adjusting the

weight and biases associated with each node which helps to minimize the loss function. The size of the batch, the number of epochs, the number of hidden layers to control the complexity of the model, and the number of nodes in each layer are other essential hyperparameters that can influence the performance of MLP. The best hyperparameters are usually selected by trial and error for obtaining the MLP model with the least and minimum errors.

To compare the trained model predictions with the actual data from the testing dataset. performance metrics are widely accepted in machine learning. There are various metrics that can be used in machine learning techniques, especially regression and forecasting. The use of performance metrics in forecasting is to measure how much the forecast deviates from observations in order to assess the quality and choose the best forecasting methods (Siami-Namini et al., 2018). The evaluation metrics usually proposed to evaluate the univariate time series models are RMSE and MAPE. RMSE is used mostly to access the accuracy of prediction obtained by a model. RMSE measures the differences between actual and predicted values by comparing the prediction errors of different models for particular data but not between datasets. The RMSE formula is represented in Eq. 3.

$$RMSE = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(Z_I - \hat{Z}_I)} \quad (3)$$

where, $\hat{Z}_I$ is predicted value, Z_I is actual value, and N is total number of observations.

The main advantage of using the metric is that it penalizes large errors and scales the scores in the same units as the forecast values. Mean absolute percentage error is a performance metric that can be used to calculate margin error from the predicted least square data. It is a method to perform a calculation for time series data with seasonal trends. The mean absolute percentage is calculated by using the absolute error in each period divided by the observed values that are evident for that period, and finally averaging those fixed percentages (Khair et al., 2017). The MAPE formula is represented in Eq. 4.

$$MAPE = \sum \frac{\frac{|y1-yt|}{y1}}{n} \times 100 \quad (4)$$

where, y1 is actual value, yt is forecast value, and n is sample size.

5. Results and discussion

This research endeavors to identify the most accurate model for time series forecasting by evaluating several models. For the sake of simplicity and direct comparison, the RMSE and MAPE are adopted as performance metrics for each model. The model exhibiting the lowest RMSE and MAPE values is chosen as the optimal model for prediction. Forecasts solely based on total monthly production prove inadequate, given the substantial variations in production across different locations. Such variations have the potential to introduce biases in the prediction, affecting the accuracy and precision of the results. To extract more informative insights from the datasets and enhance their utility, it is advantageous to predict monthly production for each specific location. Considering the limited availability of production data, this study focuses on five distinct aquaculture locations: Tanjong Pelumpong, Pulau Kaingaran, Sungai Dua, Sungai Bunga, and Buang Tawar. The initial experiment centers on the total marine fish production between January 2016 and July 2021, aiming to identify the optimal model for predictions. The univariate dataset employed in this experiment encompasses total marine fish production data from all locations and selected locations, including Pulau Kaingaran, Tanjong Pelumpong, Sungai Bunga, Sungai Dua, and Buang Tawar. Tables 3 to 8 present a comparative analysis of the performance of different univariate models applied to marine fish production data from 2016 to 2021, utilizing RMSE, MAPE, and RMSPE as performance metrics. The datasets for this initial experiment span from January 2016 to July 2021, encompassing total marine fish production for all locations, as well as individual locations: Pulau Kaingaran, Tanjong Pelumpong, Sungai Dua, Buang Tawar, and Sungai Bunga.

Table 3: Evaluation of total marine fish production in all locations from 2016 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	31257	7560	0.221	0.241866
Prophet	26980	14106	0.3	0.522832
Random forest regression	28465	6280	0.153	0.220622
LSTM	26980	8643	0.229	0.320348
ARIMA	26980	6240	0.147	0.231282
MLP	26980	8078	0.244	0.299407

Table 4: Evaluation of total marine fish production in Pulau Kaingaran from 2016 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	5674	946	0.123	0.166725
Prophet	5058	686	0.124	0.135627
Random forest regression	5815	737	0.118	0.126741
LSTM	5058	645	0.097	0.127521
ARIMA	5058	705	0.089	0.139383
MLP	5058	759	0.0994	0.150059

Table 5: Evaluation of total marine fish production in Tanjong Pelumpong from 2016 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	16121	6549	0.28	0.40624
Prophet	15418	9290	0.361	0.602542
Random forest regression	16655	6588	0.279	0.395557
LSTM	15418	4993	0.223	0.323842
ARIMA	15418	6360	0.275	0.412505
MLP	15418	5784	0.265	0.375146

Table 6: Evaluation of total marine fish production in Sungai Dua from 2016 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	86	109	0.559	1.267442
Prophet	77.64	276	17.43	3.554869
Random forest regression	70	114	0.601	1.628571
LSTM	77.64	91	0.51	1.172076
ARIMA	78	113	0.539	1.448718
MLP	77	122	0.283	1.584416

Table 7: Evaluation of total marine fish production in Buang Tawar from 2016 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	1794	1108	0.422	0.617614
Prophet	1650	1428	2.15	0.865455
Random forest regression	1619	1131	0.365	0.698579
LSTM	1651	1214	0.355	0.735312
ARIMA	1651	1133	0.411	0.686251
MLP	1651	1578	0.429	0.955784

Table 8: Evaluation of total marine fish production in Sungai Bunga from 2016 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	3872	1126	0.271	0.290806
Prophet	4468	3325	2.31	0.744181
Random forest regression	3872	2152	0.73	0.555785
LSTM	4668	558	0.106	0.119537
ARIMA	4668	595	0.112	0.127464
MLP	4668	2829	1.353	0.606041

The final experiment is similar to the first experiment; however, the marine fish production data used in this experiment is from January 2010 to July 2021. From Tables 9 to 14, the evaluation of time series forecasting techniques on the univariate time-series dataset is presented. As in the previous analysis, each table represents different datasets based on marine fish production in different locations.

Table 9: Evaluation of total marine fish production in all locations from 2010 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	33661	20275	0.449	0.602329
Prophet	31657	19165	0.386	0.605395
Random forest regression	33661	19756	0.393	0.586911
LSTM	31657	27929	0.53	0.882238
ARIMA	31657	11372	0.232	0.359225
MLP	31657	20693	0.54	0.653663

Table 10: Evaluation of total marine fish production in Pulau Kaingaran from 2010 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	5053	1392	0.216	0.27548
Prophet	4836	1738	0.2605	0.359388
Random forest regression	5069	1183	0.179	0.233379
LSTM	4836	1577	0.236	0.326096
ARIMA	4836	1150	0.185	0.2378
MLP	4836	1577	0.214	0.326096

Table 11: Evaluation of total marine fish production in Tanjong Pelumpong from 2010 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	17570	6094	0.269	0.346841
Prophet	15675	8218	0.348	0.524274
Random forest regression	17570	7852	0.291	0.446898
LSTM	15675	7636	0.382	0.487145
ARIMA	15675	4420	0.227	0.281978
MLP	15675	5107	0.247	0.325805

Table 12: Evaluation of total marine fish production in Sungai Dua from 2010 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	86	109	0.559	1.267442
Prophet	88	193	0.645	2.193182
Random forest regression	70	114	0.601	1.628571
LSTM	88	100	0.557	1.136364
ARIMA	89	84	0.61	0.94382
MLP	88	89	0.572	1.011364

Based on the evaluation result, the best model for each location is visualized to have further insight. The best fit model for the total marine fish for all locations from 2010 to 2021 is ARIMA (21, 1, 19) with ACF and PACF plot methods for the parameter estimation. The MAPE value of this model is 0.232 and the RMSE value is 11372. Fig. 28 shows the predicted and test result from this model for 18 months up to July 2021. This period is chosen because it gives the best prediction result. Based on this result, the total marine fish production in Brunei shows a positive trend over the next few years, where the highest production could reach up to 50,000 kg per month.

Table 13: Evaluation of total marine fish production in Buang Tawar from 2010 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	1465	833	0.412	0.568601
Prophet	1496	1669	0.449	1.115642
Random forest regression	1465	797	0.477	0.544027
LSTM	1496	1110	0.522	0.741979
ARIMA	1497	807	0.419	0.539078
MLP	1496	986	0.456	0.659091

Table 14: Evaluation of total marine fish production in Sungai Bunga from 2010 to 2020

Model	Mean	RMSE	MAPE	RMSPE
Linear regression	2662	925	0.296	0.347483
Prophet	3160	2407	1.711	0.761709
Random forest regression	2567	1275	0.418	0.496689
LSTM	3160	883	0.25	0.27943
ARIMA	3160	606	0.166	0.191772
MLP	3160	2017	0.891	0.638291

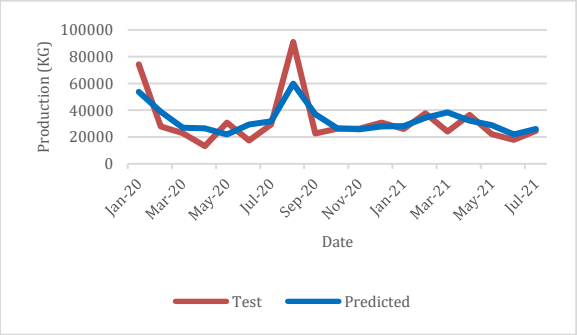

Fig. 28: Forecast from ARIMA in all locations

In Pulau Kaingaran, the most suitable model for the total marine fish production is random forest regression with a MAPE value of 0.179. Fig. 29 shows the predicted and test result using this model and the forecasting results for Pulau Kaingaran within the same period. The forecast shows that the trend of marine fish production slightly decreased for the few months ahead of the dataset period. Overall, the forecast shows the rate of production remains constant, which is from 4,000 to 5,000 kg per month.

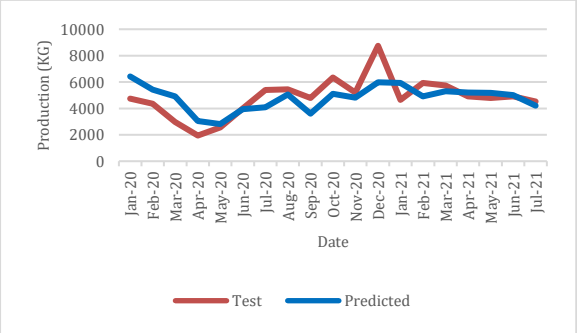

Fig. 29: Forecast from Random Forest in Pulau Kaingaran

The best model for the Tanjong Pelumpong production dataset is ARIMA (12, 1, 7) with a MAPE value of 0.227. The predicted and test result using this model for Tanjong Pelumpong with the same period is illustrated in Fig. 30. The trend shows long-term fluctuations of marine fish production in Tanjong Pelumpong with the highest can reach up to 16,000 kg and the lowest being 12,000 kg per month.

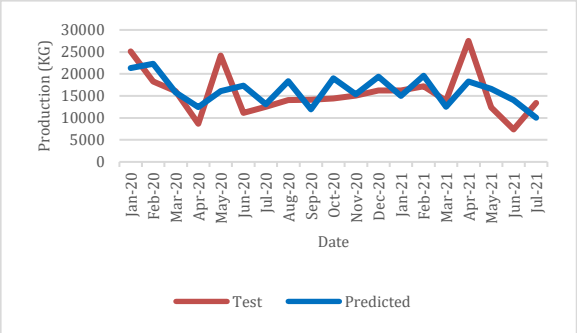

Fig. 30: Forecasts from ARIMA (Tanjong Pelumpong)

The best model for Sungai Bunga is ARIMA (18, 1, 7) because it has the lowest MAPE value of 0.166. Fig. 31 shows the predicted and test result of this model as a monthly forecast for total marine fish production in Sungai Bunga from August 2021 until August 2022 are visualized to analyze the trend. Fig. 32 shows the positive trend in marine fish production, where the production may reach up to almost 8,000 kg per month.

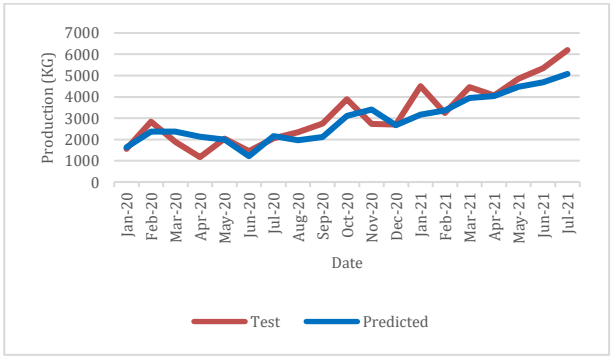

Fig. 31: Forecasts from ARIMA (Sungai Bunga)

The linear regression model performs better than other models, with a MAPE value of 0.559. The model is chosen to forecast the total marine fish production in Sungai Dua from January 2019 to July 2020. The period is different from other locations due to the best prediction result during this period. Fig. 32 illustrates a gradual decline in marine fish production in Sungai Dua over the upcoming months, with the lowest production reaching only 127 kg per month.

ARIMA has the lowest value of MAPE compared to other models, which is around 0.419, making it the best fit model for the Buang Tawar dataset. Fig. 33 shows the predicted and test result of this model from August 2020 to January 2021. Based on this forecasting result, the marine fish production in Sungai Bunga will remain constant over the few months after the dataset period.

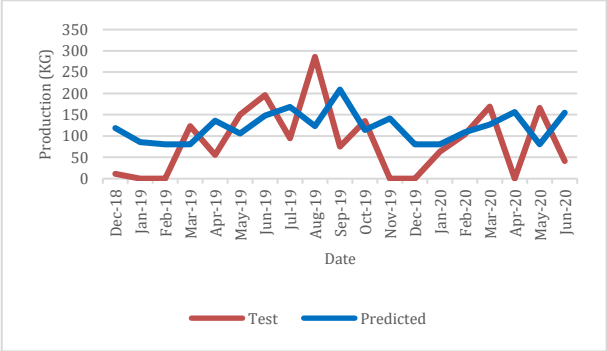

Fig. 32: Forecasts from linear regression in Sungai Dua

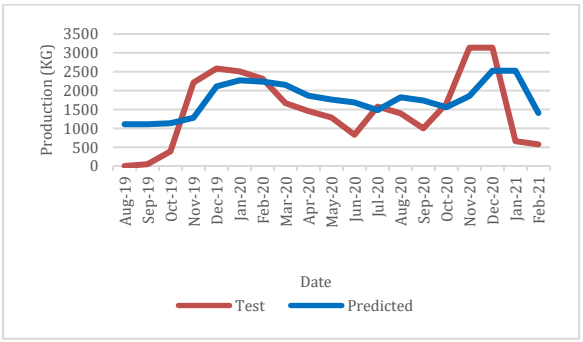

Fig. 33: Forecasts from ARIMA in Buang Tawar

In this study, various univariate models for time series analysis were employed to forecast marine fish production data in five different locations. The objective was to identify the best-performing model to predict production for the twelve months following the dataset period. The assumption made for this prediction was that sudden changes in operational activities would not significantly impact production levels. However, based on the results obtained in the previous section, no single model demonstrated 100% accuracy, with some exhibiting either over-forecasting or under-forecasting tendencies in relation to the actual values. Nonetheless, the model indicated a continuous increase in total marine fish production from August 2021 to August 2022, with the highest monthly production projected to reach 50,000 kg.

Throughout the experiment, the performance of both marine fish datasets, spanning from January 2016 and January 2010, was assessed. It was observed that the dataset with a duration of five years exhibited superior performance compared to the dataset spanning ten years, but it was posited that the dataset from January 2010 might provide more extended forecasting capabilities.

Notably, the forecasting performance of the models tended to decline with longer forecasting periods. To enhance the accuracy of time series forecasting, it is crucial to update the monthly production data regularly, enabling the utilization of more refined models and achieving better accuracy. From the model testing, it was evident that simple statistical algorithms like ARIMA and traditional machine learning methods such as linear regression and random forest outperformed deep learning methods like LSTM and MLP across all datasets.

Despite the forecasting results falling short of expectations, they can still provide valuable insights for the Department of Fisheries, aiding in anticipating potential adverse situations and enabling timely precautionary measures. The research also highlights the need to consider the impact of the COVID-19 pandemic on the aquaculture industry in Brunei, which may have significantly affected production due to movement restrictions and importation disruptions.

To enhance the model's performance and facilitate future analyses, several recommendations are put forth. First, incorporating additional relevant factors such as feed intake, water quality, annual rainfall, and marine fish mortality rate can significantly improve predictive accuracy. Second, implementing appropriate database systems for data collection, storage, retrieval, and analysis is crucial to streamline the data gathering process and mitigate missing data issues. Third, regularly updating the models with new data is essential for improving forecast accuracy, particularly in response to significant events that can impact production. Finally, the integration of an Internet of Things (IoT) system in aquaculture can automate data collection, reducing human errors, and enabling real-time data capture under varying conditions, making it particularly suitable for time series forecasting applications.

6. Conclusion

This research proposes six forecast methods for developing predictive models of marine fish production in Brunei Darussalam. Among these methods, ARIMA, linear regression, and random forest are found to be suitable for forecasting marine fish production. The forecasted values indicate an increasing trend in marine fish production over the upcoming months following the dataset period. The projections suggest that production could potentially reach up to 50,000 kg per month, assuming no sudden changes in operational activities that may impact monthly marine fish production. Forecast accuracy is assessed using MAPE and RMSE metrics, and based on these measures, ARIMA, linear regression, and random forest are selected as the best models, outperforming other methods. Consequently, these models are deemed optimal for forecasting univariate time-series data of marine fish in Brunei.

The implications of these findings are significant for operators and the Department of Fisheries in Brunei, providing valuable insights for planning and improving marine fish production in the country. Notably, to the best of our knowledge, such predictive models have not been previously developed in Brunei Darussalam. However, it is acknowledged that this study focuses on a limited set of univariate time series models due to the unique characteristics of the dataset in Brunei Darussalam. Therefore, in future research, we recommend exploring multivariate time series models and incorporating additional datasets from various stakeholders to further enhance forecasting capabilities.

Acknowledgment

The authors would like to acknowledge the Department of Fisheries from the Ministry of Primary Resources and Tourism for providing the data to be used in this research. The authors would also like to thank UTB for the Centre grant to fund the whole aquaculture project under the Centre of Innovative Engineering.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abyaneh HZ (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science and Engineering, 12: 40. https://doi.org/10.1186/2052-336X-12-40 **PMid:24456676 PMCid:PMC3906747**

Alraouji Y and Bramantoro A (2014). International call fraud detection systems and techniques. In the 6th International Conference on Management of Emergent Digital EcoSystems, Association for Computing Machinery, Buraidah, Saudi Arabia: 159-166. https://doi.org/10.1145/2668260.2668272

Arshad N, Samat N, and Lee LK (2022). Insight into the relation between nutritional benefits of aquaculture products and its consumption hazards: A global viewpoint. Frontiers in Marine Science, 9: 925463. https://doi.org/10.3389/fmars.2022.925463

Belyadi H and Haghighat A (2021). Introduction to machine learning and Python. In: Belyadi H and Haghighat A (Eds.), Machine learning guide for oil and gas using Python: 1-55. Gulf Professional Publishing, Houston, USA. https://doi.org/10.1016/B978-0-12-821929-4.00006-8

Bramantoro A, Suhaili WS, and Siau NZ (2022). Precision agriculture through weather forecasting. In the International Conference on Digital Transformation and Intelligence, IEEE, Kuching, Malaysia: 203-208. https://doi.org/10.1109/ICDI57181.2022.10007299

Castán-Lascorz MA, Jiménez-Herrera P, Troncoso A, and Asencio-Cortés G (2022). A new hybrid method for predicting univariate and multivariate time series based on pattern forecasting. Information Sciences, 586: 611-627. https://doi.org/10.1016/j.ins.2021.12.001

Chen L, Yang X, Sun C, Wang Y, Xu D, and Zhou C (2020). Feed intake prediction model for group fish using the MEA-BP neural network in intensive aquaculture. Information Processing in Agriculture, 7(2): 261-271. https://doi.org/10.1016/j.inpa.2019.09.001

Das SK, Xiang TW, Noor NM, De M, Mazumder SK, and Goutham-Bharathi MP (2021). Temperature physiology in grouper (Epinephelinae: Serranidae) aquaculture: A brief review. Aquaculture Reports, 20: 100682. https://doi.org/10.1016/j.aqrep.2021.100682

Dubey AK, Kumar A, García-Díaz V, Sharma AK, and Kanhaiya K (2021). Study and analysis of SARIMA and LSTM in

forecasting time series data. Sustainable Energy Technologies and Assessments, 47: 101474. https://doi.org/10.1016/j.seta.2021.101474

Elhassan A, Abu-Soud SM, Alghanim F, and Salameh W (2022). ILA4: Overcoming missing values in machine learning datasets: An inductive learning approach. Journal of King Saud University-Computer and Information Sciences, 34(7): 4284-4295. https://doi.org/10.1016/j.jksuci.2021.02.011

Elsayed S, Thyssens D, Rashed A, Jomaa HS, and Schmidt-Thieme L (2021). Do we really need deep learning models for time series forecasting? ArXiv Preprint ArXiv:2101.02118. https://doi.org/10.48550/arXiv.2101.02118

Estrebillo RA and Hiramoto H (2021). Brunei Darussalam aquaculture feasibility study for investment. ASEAN-Japan Centre: ASEAN Promotion Centre on Trade, Investment and Tourism, Tokyo, Japan.

Fan D, Sun H, Yao J, Zhang K, Yan X, and Sun Z (2021). Well production forecasting based on ARIMA-LSTM model considering manual operations. Energy, 220: 119708. https://doi.org/10.1016/j.energy.2020.119708

Hana KM, Al Faraby S, and Bramantoro A (2020). Multi-label classification of Indonesian hate speech on Twitter using support vector machines. In the International Conference on Data Science and Its Applications, IEEE, Bandung, Indonesia: 1-7. https://doi.org/10.1109/ICoDSA50139.2020.9212992

Hu Z, Zhang Y, Zhao Y, Xie M, Zhong J, Tu Z, and Liu J (2019). A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. Sensors, 19(6): 1420. https://doi.org/10.3390/s19061420 **PMid:30909468 PMCid:PMC6470961**

Hülya S and Abdallah TSY (2021). Aquaculture production of North African countries in the year 2030. Survey in Fisheries Sciences, 8(1): 107-118. https://doi.org/10.18331/SFS2021.8.1.8

Kane MJ, Price N, Scotch M, and Rabinowitz P (2014). Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics, 15: 276. https://doi.org/10.1186/1471-2105-15-276 **PMid:25123979 PMCid:PMC4152592**

Khair U, Fahmi H, Al Hakim S, and Rahim R (2017). Forecasting error calculation with mean absolute deviation and mean absolute percentage error. Journal of Physics: Conference Series, 930: 012002. https://doi.org/10.1088/1742-6596/930/1/012002

Khotimah WN (2014). Aquaculture water quality prediction using smooth SVM. IPTEK Journal of Proceedings Series, 1: 342-345.

Marsal CJ, Jamaludin MH, Anwari AS, and Chowdhury AJK (2023). The potential of aquaculture development in Brunei Darussalam. Agriculture Reports, 2(1): 12-21.

Martínez F, Charte F, Frías MP, and Martínez-Rodríguez AM (2022). Strategies for time series forecasting with generalized regression neural networks. Neurocomputing, 491: 509-521. https://doi.org/10.1016/j.neucom.2021.12.028

Okeke-Ogbuafor N, Stead S, and Gray T (2021). Is inland aquaculture the panacea for Sierra Leone's decline in marine fish stocks? Marine Policy, 132: 104663. https://doi.org/10.1016/j.marpol.2021.104663

Ouatahar L, Bannink A, Lanigan G, and Amon B (2021). Modelling the effect of feeding management on greenhouse gas and nitrogen emissions in cattle farming systems. Science of the Total Environment, 776: 145932. https://doi.org/10.1016/j.scitotenv.2021.145932

Pavlyshenko BM (2019). Machine-learning models for sales time series forecasting. Data, 4(1): 15. https://doi.org/10.3390/data4010015

Petropoulos F, Apiletti D, Assimakopoulos V, Babai MZ, Barrow DK, Taieb SB, Bergmeir C, Bessa RJ, Bijak J, Boylan JE, and Browell J (2022). Forecasting: Theory and practice. International Journal of Forecasting, 38(3): 705-871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Pratondo A and Bramantoro A (2022). Classification of *Zophobas morio* and *Tenebrio molitor* using transfer learning. PeerJ Computer Science, 8: e884. https://doi.org/10.7717/peerj-cs.884 **PMid:35494845 PMCid:PMC9044276**

Rahman LF, Marufuzzaman M, Alam L, Bari MA, Sumaila UR, and Sidek LM (2021). Developing an ensembled machine learning prediction model for marine fish and aquaculture production. Sustainability, 13(16): 9124. https://doi.org/10.3390/su13169124

Ramazan ÜNLÜ (2019). A comparative study of machine learning and deep learning for time series forecasting: A case study of choosing the best prediction model for turkey electricity production. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 23(2): 635-646. https://doi.org/10.19113/sdufenbed.494396

Satrio CBA, Darmawan W, Nadia BU, and Hanafiah N (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. Procedia Computer Science, 179: 524-532. https://doi.org/10.1016/j.procs.2021.01.036

Shen ML, Lee CF, Liu HH, Chang PY, and Yang CH (2021). Effective multinational trade forecasting using LSTM recurrent neural network. Expert Systems with Applications, 182: 115199. https://doi.org/10.1016/j.eswa.2021.115199

Siami-Namini S, Tavakoli N, and Namin AS (2018). A comparison of ARIMA and LSTM in forecasting time series. In the 17th IEEE International Conference on Machine Learning and Applications, IEEE, Orlando, USA: 1394-1401. https://doi.org/10.1109/ICMLA.2018.00227

Taud H and Mas JF (2018). Multilayer perceptron (MLP). In: Camacho Olmedo M, Paegelow M, Mas JF, and Escobar F (Eds.), Geomatic approaches for modeling land change scenarios: 451-455. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-60801-3_27

Yu P, Gao R, Zhang D, and Liu ZP (2021). Predicting coastal algal blooms with environmental factors by machine learning methods. Ecological Indicators, 123: 107334. https://doi.org/10.1016/j.ecolind.2020.107334

Zheng A (2015). Evaluating machine learning models. O'Reilly Media Inc., Sebastopol, USA.

Zhou T, Jiang Z, Liu X, and Tan K (2020). Research on the long-term and short-term forecasts of navigable river's water-level fluctuation based on the adaptive multilayer perceptron. Journal of Hydrology, 591: 125285. https://doi.org/10.1016/j.jhydrol.2020.125285