International Journal of Frontier Technology and Engineering (IJFTE)
Vol. 02, No. 02, June 2024, pp. 65-78
ISSN: 2986-9064

Data-Driven Predictions of Fish Production: Applying Regression Methods in Aquaculture

Alfiandi Aulia Rahmadani[1], Yan Watequlis Syaifudin[2*], Triana Fatmawati[2], Nobuo Funabiki[3], Pramana Yoga Saputra[2], Rokhimatul Wakhidah[2], and Mustika Mentari[3]

[1] *Department of Electrical Engineering, Politeknik Negeri Malang, Malang, Indonesia*
[2] *Department of Information Technology, Politeknik Negeri Malang, Malang, Indonesia*
[3] *Department of Electrical and Communication Engineering, Okayama University, Okayama, Japan*

Corresponding Author: Yan Watequlis Syaifudin, qulis@polinema.ac.id

Received Date: 10-04-2024 Revised Date: 05-05-2024 Accepted Date: 21-06-2024

Abstract

The Industrial Revolution 4.0 has significantly advanced the fisheries sector, improving the efficiency and effectiveness of aquaculture, which is vital for food security in Indonesia. Aquaculture, defined as the cultivation of fish in various environments such as rice paddies and ponds, has been recognized as a key area of economic growth by the Ministry of Marine Affairs and Fisheries. The ministry's innovative strategies include the development of export-based fishery cultivation and the establishment of fishery villages that incorporate local wisdom, offering promising business opportunities for fishermen. Optimizing the use of available aquaculture land is crucial to preventing ecosystem imbalances and maximizing the potential of this sector. To promote sustainable fish production, the study proposes a predictive harvest system aimed at improving the profitability of harvest partners. This system uses data collected from aquaculture practices, considering essential factors such as the number of seeds, average weights, pond volumes, and survival rates. The research employs multiple regression techniques, including polynomial regression, support vector regression, random forest regression, and linear regression, to construct accurate predictive models. By analyzing these variables, the study seeks to improve the precision of forecasting fish harvest results, which is essential for informed decision-making in aquaculture management. The evaluation of the regression models demonstrated that polynomial regression was the most accurate method, showing the lowest root mean square error (RMSE) and the mean absolute percentage error (MAPE). This suggests that selecting the right model and optimizing hyperparameters are critical for reliable predictions in fish harvest forecasting. The findings of this research underscore the importance of predictive analytics in the aquaculture sector, contributing not only to increased profitability for harvest partners, but also to the sustainability of fish production practices. By implementing such predictive systems, stakeholders can ensure better management and resource allocation, ultimately leading to a more resilient fisheries industry.

Keywords : prediction system, aquaculture, fishery, regression model, Python

1 Introduction

In today's era, the Industrial Revolution 4.0 marks the advancement of technology in the fisheries sector, which makes the fisheries cultivation industry more effective and efficient [1]. In addition, aquaculture in Indonesia is also a source of food security. Aquaculture involves the cultivation of fish in various environments, such as fields, rice paddies and other land areas, with definitions varying by source [2]. According to the Central Statistics Agency of Malang Regency (2019), it includes marine cultivation, ponds, pools, karamba, floating nets, and rice fields. The Ministry of Marine Affairs and Fisheries (KKP) optimistically states that the aquaculture sector is a potential sector to accelerate economic growth in 2022. This is because the Ministry of

Marine Affairs and Fisheries is developing two new breakthroughs to increase aquaculture productivity. These breakthroughs are the development of export-based fisheries cultivation with superior commodities and the development of fisheries villages based on local wisdom [3]. The KKP statement shows that the aquaculture sector can be a business opportunity for harvest partners (fishing entrepreneurs).

The aquaculture sector is considered a sector that can still grow and develop, because the existing potential has not been optimally used compared to the area of land available for aquaculture. The characteristics of aquaculture, namely that it can be carried out at all levels of society, from villages to cities, provides fast and high profit margins, is the basis for increasing industrial development and can apply various technologies [4]. Indonesia has a very large potential for aquaculture land [5, 6]. From this potential, it is necessary to optimize the cultivation of fisheries by harvest partners to prevent ecosystem imbalance by studying the nature of life and the original habitat of each organism so that the development of organisms or the maintenance techniques carried out can be manipulated in the cultivation environment.

Increasing fisheries production through fish cultivation is one way to increase sustainable production and pay attention to environmental sustainability. The existing fish population is decreasing due to continuous fishing activities, especially in public waters, so restocking is needed [7]. To carry out restocking, harvest partners need to calculate the estimated consumer demand and the harvest results of aquaculture production in the future [8]. The harvest season factor has an impact on aquaculture's selling price, which is currently fluctuating and related to the amount of stock on the market where the law of supply and demand is in effect. This does not happen with the price of food, which tends to continue to increase every year, resulting in increased operational costs that harvest partners must incur [9].

Predictions of events related to future fish harvest production can be used as a consideration in planning [10, 11]. A prediction system is an interconnected assembly of components designed to facilitate the transfer of material, energy, or information, encompassing goals, inputs, processes, and outputs, along with factors such as limitations and feedback mechanisms. Objects are used to predict future variables more intuitively than to rely solely on historical data, often integrating quantitative data for price forecasting. Multiple Linear Regression (MLR) is a technique that analyzes the linear relationships between a dependent variable and multiple independent variables to construct predictive models [12, 13, 14]. Support Vector Regression (SVR) applies this concept using a hyperplane to minimize errors in regression tasks, effectively addressing overfitting issues [15]. Polynomial regression extends MLR by modeling curvilinear relationships through predictor variables raised to an nth power, while Random Forest Regression utilizes an ensemble of decision trees to enhance prediction performance, with the number of trees impacting the results.

Thus, this study proposes a fisheries cultivation harvest prediction system to increase the profitability of harvest partners by comparing four regression methods, namely linear regression, polynomial regression, random forest regression, and support vector regression to obtain accurate predictions. The prediction method for fisheries harvests is based on data collected from ponds cultivated by harvest partners, including seeding information, feeding monitoring, mortality rates, average fish weight, and harvest results. The key factors affecting fish farming that inform the prediction system include the number of seeds, the average weight of the seeds, the volume of the pond, the average weight per fish at harvest, the percentage of survival rate, the total feed used and the overall results of the harvest. Data processing entails several steps, starting with data preparation to clean and filter pertinent information and ending with data visualization to explore relationships between variables. Subsequently, a dataset is created for modeling, leading to the establishment and training of various regression models such as multiple linear regression, Support Vector Regression (SVR), polynomial regression, and random forest regression. These models are then validated and tested for accuracy using models not part of the training dataset, employing metrics such as Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE).

The evaluation focused on measuring the precision of various models, specifically using RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error), while evaluating the influence of test data percentages (ranging from 5% to 95%) and hyperparameter settings on model performance. The results showed that the polynomial regression model had the lowest average RMSE (4.39%) and MAPE (0.41%), showing that it was more accurate than the other methods. For example, for 5% test data, its minimum RMSE was 0.15 and its MAPE was 0.02%. In contrast, Support Vector Regression displayed higher error metrics in general, with the average RMSE at 9.97 and the MAPE at 1.18%. The analysis used GridSearchCV for hyperparameter optimization, leading to significant improvements in RMSE and MAPE across models. Meanwhile, Random Forest Regression and Linear Regression exhibited higher error values, particularly under extreme test data conditions. In general, these results underscore the impact of model selection and hyperparameter tuning in achieving reliable predictive performance in fish harvest forecasting.

The rest of this article is organized as follows. Section 2 discusses several related theories to provide a more comprehensive and in-depth understanding of the topic. Section 3 covers the design and development process of the crop yield prediction system, detailing the steps taken to create it, and highlighting important factors that influence both the crop yield and the development of the application. Section 4 discusses the results of the experiment and analysis of the evaluation of the regression method. Lastly, this article concludes with some future works in Section 5.

2 Theoretical Basis

In this section, several theories related to the topic being discussed will be discussed to provide a more comprehensive and in-depth understanding of the topic.

2.1 Aquaculture

Aquaculture is the result of the production of fish obtained through cultivation in fields, rice fields, or other places on the land [16, 17]. There are several definitions related to aquaculture, including: According to the Central Statistics Agency of Malang Regency (2019), aquaculture is classified into types of cultivation, namely marine cultivation, ponds, and pools. Karamba, floating nets, and rice fields.

According to the Law of the Republic of Indonesia 31/2004, fish cultivation is an activity to maintain, raise and/or breed fish and harvest the results in a controlled environment, including activities that use ships to load, transport, store, cool, handle, cook and / or preserve them [18]. Aquaculture is the cultivation of aquatic organisms, including fish, mollusks, crustaceans, and aquatic flora which includes several forms of activities in the maintenance process to increase production, such as regular distribution, culinary/feed prizes, protection based on predators, and others [19].

2.2 Fish Cultivation Process

The fish farming process begins with careful planning [20, 21, 22, 23]. At this stage, the farmer estimates the amount of feed and seeds that will be the target harvest during the farming process. To achieve this, the number of seeds to be spread, the weight of the seeds in grams, and the calculation of the feed requirements required for each pond are recorded. The size of the pond used is determined based on the number of seeds to spread. For example, for 1500 catfish seeds, a pond with a diameter of 3 meters and a height of 1 meter is used, so the volume of the pond needed is around 7 cubic meters [24]. In addition, the cultivation process continues with the purchase of seeds. The fish seeds that are raised come from quality broodstock, which are obtained from seed farmers who have a good reputation in breeding quality broodstock.

After the process of obtaining feed and seeds has finished spreading in the pond, as shown in Figure 1, the next stage is the growing process which lasts approximately 90 days for catfish depending on the size of the seeds and the specified harvest target. The feeding process is carried out by continuing to observe the weather conditions and the health of the fish, it can be 2 to 3 times a day, this also affects the growth rate of the fish with a target size of 8-10 fish/kg so that the average weight of the fish per fish when harvested is around 100-125 grams. In this cultivation process, farmers sort the fish every 25 days or so to separate fish that have a higher growth rate than others because the cannibalistic nature of the catfish species can be detrimental to farmers. After the harvest period arrives, farmers have several sales options through wholesalers, resellers, or directly to end users with different levels of profit [25].

2.3 Survival Rate

The survival rate is the ratio of the number of fish that survive from the beginning to the end of the cultivation process. The survival rate has a significant impact on the production results obtained, where the higher the Survival Rate, the greater the harvest that the farmer can achieve. The low survival rate value is believed to be caused by the lack of implementation of Quality Control and proper acclimatization when distributing fish seeds. This results in the fish not being able to adapt well and eventually dying [26]. To calculate the survival rate, the following Equation 1 is used [27].

$$SR = \frac{N_t}{N_0} \times 100\% \tag{1}$$

Description:
SR: Survival rate (Nt: Number of fish at the end of cultivation

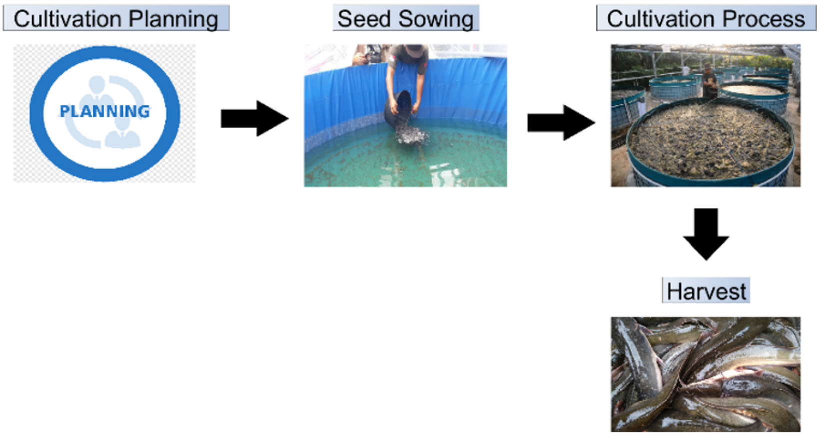

Figure 1: Fish cultivation process

NO: Number of fish at the beginning of cultivation

2.4 Prediction System

The system is a unity of elements or components that are interconnected to facilitate the transfer of material, energy, or information. This system is also called a single component that is interconnected and has a driving element. In general, the elements that make up a system and need to be known are goals, inputs, processes, and outputs. There are also other factors such as limitations, control and feedback mechanisms, system environment, and others. Prediction is the process of predicting future variables that are more intuitive than historical data, but more intuitively, quantitative data is often used in price forecasting as a complement to forecasting. So, it can be concluded that the prediction system is a unity of elements or components that has the aim of being able to process certain variables that are used to predict new variables based on variables that have been processed or exist [28].

2.5 Multiple Linear Regression (MLR)

MLR or Multiple Linear Regression is a technique to analyze the relationship between two variables, namely the dependent variable (y) to the independent variables ($x1, x2, x3, ..., xn$) linearly so that a model is formed. The purpose of the Multiple Linear Regression (MLR) method is to produce a model on the dependent variable (y) based on the values of the independent variables ($x1, x2, x3, ..., xn$) [29]. The MLR model is stated as a linear equation that includes a dependent variable, an intercept, coefficients, independent variables, and error terms. To estimate coefficients, use the least squares approach, which minimizes the sum of squared differences between the values of the observed and anticipated dependent variables [30]. The MLR approach is extensively used in a variety of scientific disciplines, including economics, engineering, and social sciences, and is beneficial to forecast results and assess the impact of several factors on a specific response variable.

2.6 Support Vector Regression

Support Vector Regression (SVR) is an SVM method applied to regression cases. SVR aims to find a function $f(x)$ as a hyperplane (separating line) in the form of a regression function which is under all input data by making the error (e) as small as possible. The SVR method is a method that can produce good performance because it can overcome the problem of overfitting, which is a condition where the machine learning model has a very low error on training data but is very high on testing data, and if the value of ($e = 0$) is obtained, it means that a perfect regression is obtained [31].

SVR aims to reduce the error while still maintaining a tolerance margin (denoted by ϵ) around the predicted function. This margin gives the model some flexibility, allowing it to generalize more well to previously unknown data. The SVR method determines the best hyperplane to maximize the margin while guaranteeing that the bulk of data points fall within the ϵ-tube around the regression line. Furthermore, SVR includes a regularization parameter, C, which governs the trade-off between maximizing the margin and reducing the error. A higher value of C produces a model that emphasizes reducing error above the increasing margin, which can lead to overfitting if not properly calibrated. In contrast, a smaller value of C allows for a larger margin, supporting a more generalizable model at the sacrifice of some training precision.

2.7 Polynomial Regression

Polynomial regression is a special type of regression that works on the curvilinear relationship between dependent and independent values. Polynomial regression is a Linear Regression model formed by adding the influence of each predictor variable (X) increased to the nth power. Polynomial regression is a modified result of the multiple linear regression model. To distinguish it from multiple linear regression, in general, polynomial regression can be crowned in exponent form. Finding the partial derivative of the SSE against the beta coefficient and setting it equal to zero is also similar to what is done in multiple linear regression [32].

Polynomial regression approximates the connection between independent and dependent variables by fitting a polynomial curve rather than a straight line. This method captures more intricate non-linear patterns in the data. Unlike multiple linear regression, which treats each predictor individually, polynomial regression accounts for predictor interactions by elevating them to different powers. The flexibility of the model increases with the degree of polynomial, allowing it to better match the training data. However, this flexibility can lead to overfitting, in which the model performs well on training data but badly on fresh, previously unknown data. To avoid this, it is crucial to determine the appropriate degree of the polynomial depending on the data and explore regularization approaches to manage the model.

2.8 Random Forest Regression

Random forest regression is an algorithm derived from a combination of several decision trees. Each decision tree in a random forest depends on the value of a random vector that is sampled independently and has the same distribution. In the random forest algorithm, the number of estimators parameter indicates the number of decision trees in the forest. In general, the more estimators, the better the results. However, at a certain point the prediction performance will decrease as a result of the high computational requirements. Therefore, it is necessary to find out at what number of estimators the algorithm produces the best prediction value [33].

Random forest regression improves model performance by averaging the predictions of numerous decision trees, reducing variance, and increasing accuracy. This ensemble approach performs particularly well with complicated datasets with high-dimensionality and variable interactions. It also makes reliable predictions in the presence of noise and outliers because of its capacity to decorrelate trees using random feature selection. Furthermore, random forests provide insights about feature relevance, helping researchers to determine which factors have the most impact on the model's predictions. However, although adding more trees improves model stability, it also increases computational time and memory utilization. Thus, it is crucial to balance the model's accuracy with computational efficiency by setting parameters such as the number of estimators and the maximum tree depth.

2.9 Laravel

Laravel is one of the best PHP frameworks currently developed by Taylor Otwell which is present as an open source web development platform. Laravel has a syntax designed to facilitate and accelerate the process of website development with an expressive and elegant style. Although Laravel is not the only widely used PHP framework, it can be an option that can be considered because it is open source and has many features that can be used [34].

Laravel syntax is intended to ease and expedite website building while maintaining an expressive and attractive style. Although Laravel is not the only commonly used PHP framework, it may be regarded since it is open source and contains many capabilities that can be employed. Laravel has built-in facilities for authentication, routing, sessions, and caching, allowing developers to construct strong and secure apps with little effort. Furthermore, Laravel's vast ecosystem offers packages such as Laravel Forge and Laravel Vapor, which provide deployment and server administration tools. The framework also supports the MVC (Model-View-Controller) architecture. This fosters separation of concerns, making the codebase easier to manage and scale. Laravel's community support and thorough documentation add to its popularity, making it a top choice for many developers.

2.10 MySQL

MySql is a network database access program, so it can be used for multi-user applications. Another advantage of MySql is that MySql uses the standard query language owned by SQL. The popularity of MySql is because MySql uses Sql as the basic language to access its database, making it easy to use. MySql is also open source and free on various platforms, except for Windows, which is shareware. MySql is distributed with an open source GPL (General Public License) license starting from version 3.23, in June 2000 [35].

MySQL can run on various operating systems such as Linux, macOS, and Windows, as it is cross-platform compatible. MySQL's fast performance and ability to handle increased capacity make it well suited for various applications on multiple computers [36]. The MySQL database engine adheres to the ACID standard, enhancing data integrity and reliability in transactional applications. MySQL also boasts robust security features such as user management, encryption, and access control to protect important data. The widespread involvement of the community and the extensive documentation of MySQL enhance usability through various troubleshooting and improvement tools.

3 Research Methods

This section discusses the design of the crop yield prediction system and the application development process. This section also describes the steps taken to create the crop yield prediction system, as well as some important factors that affect crop yield and application development.

3.1 Data collection

In this study, the targeted subjects were fish farmers who were PT harvest partners. Adma Digital Solusi. The method used to collect data came from ponds cultivated by harvest partners, which included seeding data, feeding monitoring, monitoring of cultivation mortality, average cultivation weight, and cultivation harvest results.

3.2 Fisheries Cultivation Factors

Based on studies in the literature on fish farming and data that have been collected, several important factors in fish farming have been identified that can be used to build a fish harvest prediction system. These factors include:

1. The number of seeds to be cultivated.
2. The average weight of seeds per fish in grams.
3. The volume of the pond used for cultivation in cubic meters.
4. The average weight of fish per fish at harvest (obtained from 1 kg divided by the number of fish to be sold).
5. Percentage of fish survival rate.
6. Total feed given during cultivation in kilograms.
7. Harvest results in kilograms.

3.3 Data processing

The collected data need to go through a data processing stage in order to implement the regression model. The data processing steps can be seen in Figure 2. From the block diagram of the flow of the data processing process above, there are several stages of data processing in this study with the following processes:

1. Data Preparation: in this process, data cleaning procedures are carried out, such as removing irrelevant data and handling empty data.
2. Data Visualization: in this process, previously prepared data is visualized to see the relationship between data. The following is an example of data visualization of the percentage of survival rate with the number of days as shown in Figure 3.

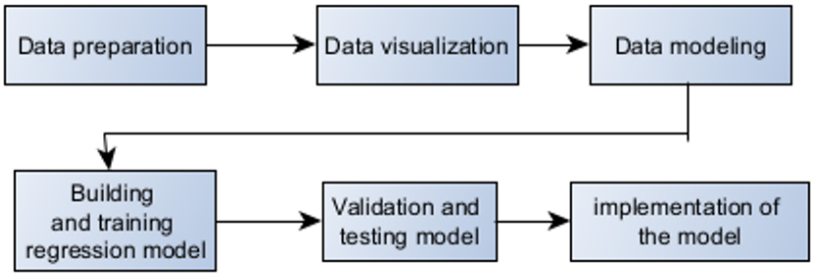
Data preparation
Data visualization
Data modeling
Building and training regression model
Validation and testing model
implementation of the model

Figure 2: Data processing block diagram

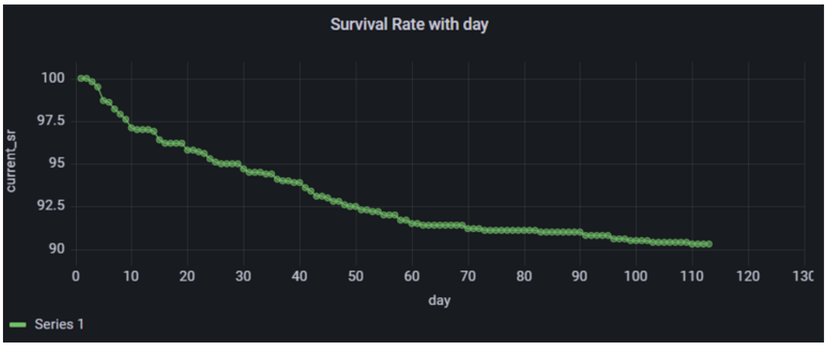

Figure 3: Data visualization of survival rate by number of days

3. Data Modeling: after knowing the relationship between data, the data modeling process is carried out which will later be used as a dataset or input for the regression model used. From the previous daily data, the data is then categorized based on the harvest and type of fish. The following is an example of a data set as in Table 1.

Table 1: Sample datasets

Fish id	Seed amount	Seed weight (g)	Survival rate (%)	Average weight (g)	Pond vol. (m^3)	Total feed (kg)
1	1000	4	90,3	115	3	167,05
1	1000	4,5	90,5	115	3	168,05
1	1500	4,2	90,1	125	7	246,45
1	2000	3,5	90,1	115	12,5	342,45
2	1000	15,5	90,4	500	7	639,85
2	1500	16,5	90,8	500	12,5	886,85
2	2000	17	90,2	500	19,5	1150,95
3	1000	16,5	72,5	500	7	453,15
3	1500	17	72,6	500	12,5	650,65
3	2000	18	81,5	500	19,5	1145,15

4. Building and Training Regression Model: in this process, 4 regression models are created and trained, namely multiple linear regression, SVR (Support Vector Regression), polynomial regression, and random forest regression using previously modeled data.
5. Validation and Testing Model: in this process, four models that have been created are validated and tested using data that is not used in the training process. Model testing uses the MAPE (Mean Absolute Percentage Error) and RMSE (Root Mean Square Error) methods.

6. Implementation of The Model: after building, training, validating, and testing the regression model, the next step is to apply it to real situations. In this process, the regression model with the best accuracy is implemented in the system to predict the results of fisheries cultivation harvests in the harvest prediction menu.

3.4 Harvest Prediction Process

In this process, the model with the best accuracy is implemented in the system to predict the results of the fish farming harvests in the "Harvest Prediction" menu available to the harvest partners. Through this menu, the harvest partners can predict the results of the fish harvests in kilograms before starting the cultivation. To make predictions, harvest partners must enter initial cultivation preparation data, such as fish type, number of seeds per tail, average weight of seeds per tail in grams, pond volume in cubic meters, and average target weight of fish per tail to be sold in grams (calculated by dividing 1000g by the number of tails to be sold in 1kg). Data on the percentage of fish survival rate will be entered automatically by the system using references from previous data, which provide 5 percentages of survival rates. This will produce 5 prediction results presented to harvest partners. Determination of 5 percentage survival rate data using minimum and maximum data, then the interval is adjusted to obtain 5 input data on the percentage of survival rate. The system will then automatically input the total feed data to be provided during cultivation, using references from previous data considering the type of fish, the percentage of survival rate, and the number of seeds. From these data, the average total feed will be sought.

The data that has been entered will be used as input for the best machine learning model that has been previously tested. After processing the input data, the model will produce output in the form of 5 predictions of cultivation harvest results in kilograms based on 5 survival rate percentage scenarios. The system will automatically provide guidance and simulation to provide cultivation feed according to the prediction results presented, model illustrations as shown in Figure 4.

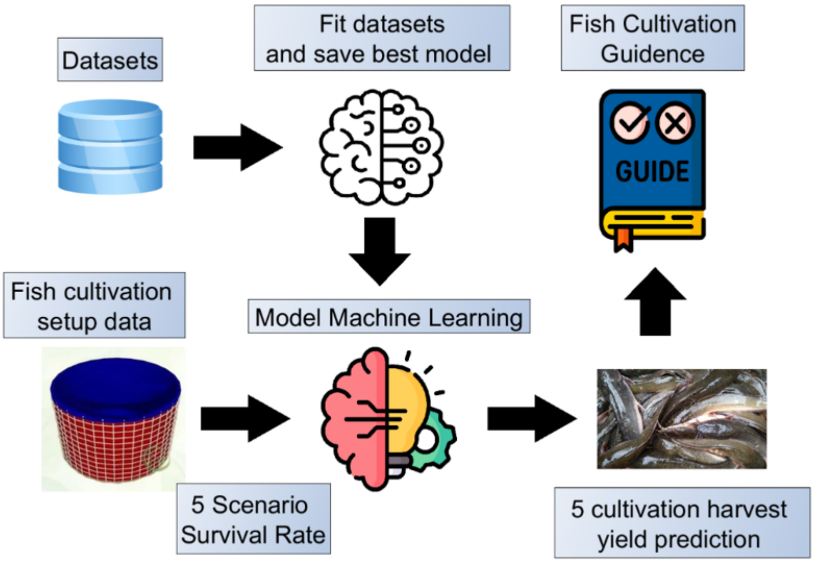
Datasets
Fit datasets and save best model
Fish Cultivation Guidence
GUIDE
Fish cultivation setup data
Model Machine Learning
5 Scenario Survival Rate
5 cultivation harvest yield prediction

Figure 4: Harvest prediction process

4 Results and Analysis

In this section, we discuss the results and analysis of the evaluation of the regression method used.

4.1 Evaluation Scenario

In this study, an evaluation of the adopted method was performed. The evaluation of the adopted method aims to measure the accuracy of the regression model using the RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) methods. This test aims to understand the effect of the percentage of test data and hyperparameter settings on the performance of four different regression models. The test scenario involves the use of 19 percentages of test data that vary from 5% to 95%, with an increasing interval of 5%. In addition, the two hyperparameter settings used are the default setting and the best setting. The main purpose of this test is to analyze the ability of the regression model to recognize data patterns, as well as to evaluate the resulting RMSE and MAPE values. The five survival rate percentages (5%, 15%, 30%, 50%, and 95%) were chosen based on the previous data analysis from the harvest partners. This range includes situations ranging from the worst to the greatest, offering a realistic view of the expected harvest results. Low survival rates can be attributed to poor environmental circumstances, whereas high survival rates represent ideal settings. This range was intended to give predictability and allow harvest partners to plan better management scenarios.

4.2 RMSE Evaluation Results with Default Hyperparameter Settings

Figure 5 shows the test results with the default hyperparameter settings. The linear regression model has the lowest RMSE value in the percentage of test data 45% (34,196) and the highest in the percentage of test data 5% (49,032). The polynomial regression model has the lowest RMSE value in the percentage of test data 5% (1,675) and the highest in the percentage of test data 90% (46.37). Random Forest Regression achieves the lowest RMSE value in the percentage of test data 5% (8.182) and the highest in the percentage of test data 95% (77.004), while support vector regression achieves the lowest RMSE value in the percentage of test data 15% (218.493) and the highest in the percentage of test data 5% (283.76). 5% test data percentage (283.761).

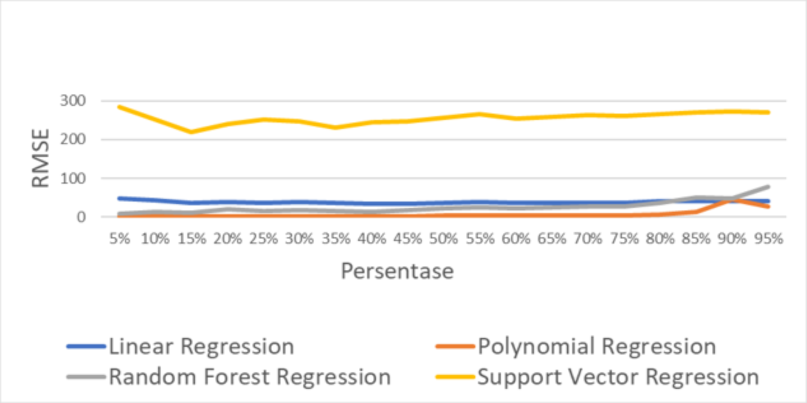

Figure 5: RMSE evaluation results with default hyperparameter settings

The polynomial regression model, random forest regression (25,940), linear regression (38,390), and support vector regression (255,470) have the lowest RMSE (7,170) on average. The lowest RMSE value overall is 1.675 for a percentage of 5% test data in the polynomial regression model, while the highest RMSE value is 283.761 for a 5% test data in the support vector regression model.

4.3 MAPE Evaluation Results with Default Hyperparameter Settings

Based on the test results as shown in Figure 6 with default hyperparameter settings, it can be seen that the Linear Regression model has the lowest MAPE value at the test data percentage 95% (8%) and the highest at the test data percentage 35% (10. 21%), while polynomial regression has the lowest MAPE value at the test data percentage 20% (0. 47%) and the highest at the test data percentage 90% (2. 93%). Random Forest Regression achieves the lowest MAPE value in the percentage of test data 5% (0. 91%) and the highest in the percentage of test data 95% (12. 87%), while Support Vector regression achieves the lowest MAPE value in the percentage of test data 10% (35. 45%) and the highest in the percentage of test data 95% (54. 49%). On average, the Polynomial Regression model has the lowest MAPE value (0. 9%), followed by random forest regression (3. 16%), linear regression (9. 18%) and support vector regression (45. 56%). The lowest overall

MAPE value is 0. 47% for the percentage of 20% test data in the polynomial regression model, while the highest MAPE value is 54. 49% for the percentage of 95% test data in the support vector regression model.

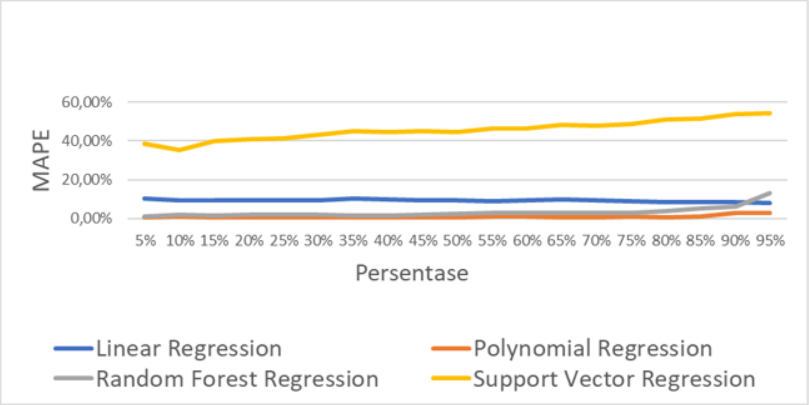

Figure 6: MAPE evaluation results with default hyperparameter settings

4.4 Best Hyperparameter Setting Results

In this study, the *GridSearchCV* algorithm is used to set the hyperparameters of the regression model and obtain the best parameters. One of the main advantages of GridSearchCV is its excellent compatibility with various machine learning algorithms. This algorithm can find the best parameters objectively using the cross-validation method, where the model performance score is obtained for each combination of parameters. In addition, GridSearchCV is also effective in preventing overfitting by applying the cross-validation technique. The first step to determine the best hyperparameters is to initialize the estimator parameters for each regression model as in the following program code snippet. The combination of the Random Forest Regression approach with hyperparameter optimization using GridSearchCV is the key novelty of the study. GridSearchCV searches for the optimal parameters of the regression model, resulting in more accurate and robust predictions. This paper proposes a hybrid technique that addresses the difficulty of overfitting while still producing a reliable prediction model.

Listing 1: Parameter estimator initialization code snippet

```
model_params = {
    ’linear_regression’: {
        ’model’: LinearRegression(),
        ’params’ : {
            # No Parameter
        }
    },
    ’polynomial_regression’: {
        ’model’: pipeline_poly,
        ’params’ : {
            ’poly__degree’ : [2,3,4]
        }
    },
    ’random_forest’: {
        ’model’: RandomForestRegressor(),
        ’params’ : {
            ’n_estimators’: [1,10,100],
            ’min_samples_leaf’ : [1,2,3],
            ’random_state’: [0]
        }
    },
    'svr': {
        'model': pipeline_svr,
        'params': {
            'svr__kernel': ["linear", "poly", "rbf", "sigmoid", "precomputed"],
            'svr__degree': [2,3,4],
            'svr__gamma': ['scale', 'auto'],
            'svr__coef0': [0,1,2],
            'svr__C': [1,10,100],
        }
    }
}
```

Next, GridSearchCV will explore and evaluate the parameter combinations of the estimators, then apply the scenario to the model to find the best parameters. The results of the best parameter search can be seen in Table 2 below.

Table 2: Hyperparameter setting results using GridSearchCV

No	Model	Best Score	Best Parameters
1	Linear Regression	-32,151	Has no parameters
2	Polynomial Regression	-0,237	Degree: 3
3	Random Forest Regression	-13,144	Min_samples_leaf: 1, n_estimators: 100, random_state: 0
4	Support Vector Regression	-0,477	C: 100, coef0: 2, degree: 3, gamma: scale, kernel: poly

After obtaining the best parameters for each regression model using GridSearchCV, the next step is to conduct a trial to measure the RMSE value of each regression model using the best hyperparameter settings.

4.5 RMSE Evaluation Results with the Best Hyperparameter Settings

Based on the test results as shown in Figure 7 with the best hyperparameter settings, it can be seen that the Linear Regression model has the lowest RMSE value at the 45% test data percentage (34.196) and the highest at 5% test data percentage (49.032), while the Polynomial Regression model has the lowest RMSE value at 5% test data percentage (0.168) and the highest at 95% test data percentage (34.72). Random Forest Regression achieves the lowest RMSE value in the percentage of test data 5% (8.182) and the highest in the percentage of test data 95% (77.004), while support vector regression has the lowest RMSE value in the percentage of test data 5% (9.127) and the highest in the percentage of test data 95% (75.72). On average, the Polynomial Regression model has the lowest RMSE value (4.39), followed by Support Vector Regression (9.97), Random Forest Regression (25.94), and Linear Regression (38.39). The lowest RMSE value overall is 0.15 for 5% of the test data in the Support Vector Regression model, while the highest RMSE value is 77.004 for 95% of the test data in the Random Forest Regression model.

4.6 MAPE Evaluation Results with the Best Hyperparameter Settings

Based on the test results as shown in Figure 8 with the best hyperparameter settings, it can be seen that the Linear Regression model has the lowest MAPE value at the 95% test data percentage (8%) and the highest at the 35% test data percentage (10. 21%), while polynomial regression has the lowest MAPE value at 5% test data percentage (0.02%) and the highest at 95% test data percentage (4.46%). Random Forest Regression achieves the lowest MAPE value in the percentage of test data 5% (0. 91%) and the highest in the percentage of test data 95% (12. 87%), while Support Vector regression achieves the lowest MAPE value in the percentage of test data 5% (0. 04%) and the highest in the percentage of test data 95% (8. 31%). The polynomial regression model has the lowest MAPE value (0. 41%), then the support vector regression (1.18%), the random forest regression (3. 16%), and the linear regression (9.18%). The lowest overall MAPE value is 0.02% for 5% of the test data in the Polynomial Regression model, while the highest MAPE value is 10.55% for 95% of the test data in the Random Forest Regression model.

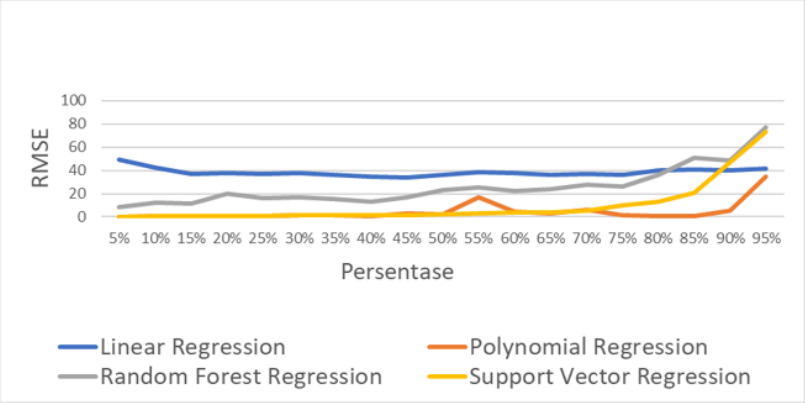

Figure 7: RMSE evaluation results with the best hyperparameter settings

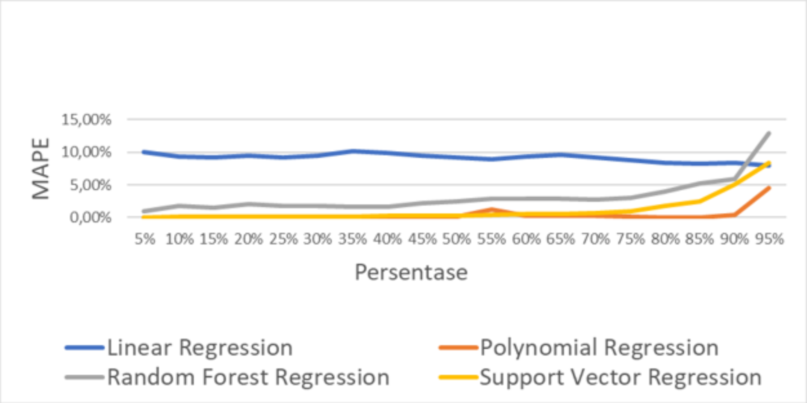

Figure 8: MAPE evaluation results with the best hyperparameter settings

5 Conclusion

The design and development of the crop yield prediction system involved a meticulous data collection process that collected comprehensive information from fish farmers associated with PT. Adma Digital Solusi. Significant attention was paid to crucial factors that influence crop yield, encompassing variables such as seed quantity, average weight, pond volume, and survival rates. Following data collection, a robust data processing workflow enabled the creation of a regression model. This workflow included several essential stages: data preparation, visualization, modeling, building, and validation of the regression models. The outcome of this process resulted in a well-structured system capable of accurately predicting harvest yields based on varying input parameters.

The implementation of the selected regression model for harvest prediction is a significant milestone in supporting fish farming ventures. The system allows entrepreneurs to input crucial variables, such as type of fish, seed amount, and pond specifications, to generate predictions for potential harvest outcomes. Through intelligent allocation of feed and resources based on reliable predictions, farmers can optimize their operations. The system is built to be user-friendly, offering insights that can help to make strategic decisions, ultimately promoting sustainability and efficiency in fish farming practices.

Looking towards the future, further development of the crop yield prediction system may introduce enhancements that improve predictive accuracy and user experience. Future works could focus on integrating machine learning techniques that adapt and update predictions based on real-time data input, such as weather conditions or disease outbreaks. Expanding the data set to include various farming scenarios could improve the robustness and applicability of the model in various geographical regions.

References

[1] D. Dahliah, A. Kurniawan, and A. H. P. K. Putra, "Analysis and strategy of economic development policy for smes in indonesia," *Journal of Asian Finance, Economics and Business*, vol. 7, 2020.

[2] J. Puszkarski and O. Śniadach, "Instruments to implement sustainable aquaculture in the european union," *Marine Policy*, vol. 144, 2022.

[3] C. B. of Statistics, "Marine and coastal resources statistics 2023," p. 3312002, 2023.

[4] I. S. B. Ntsama, B. A. Tambe, J. J. T. Takadong, G. M. Nama, and G. Kansci, "Characteristics of fish farming practices and agrochemicals usage therein in four regions of cameroon," *Egyptian Journal of Aquatic Research*, vol. 44, pp. 145–153, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1687428518300293

[5] Republic of Indonesia, *Law Number 45 of 2009 concerning Amendments to Law Number 31 of 2004 on Fisheries.* State Secretariat of the Republic of Indonesia, 2009, article 7, Paragraph (1) states that the government is obliged to develop aquaculture to increase production, improve income, and enhance the welfare of the fisheries community.

[6] Indonesian government, *Government Regulation Number 23 of 2021 concerning Forestry Administration.* State Secretariat of the Republic of Indonesia, 2021, article 167 states that the use of conservation areas and areas for aquaculture must adhere to the principles of environmental, social, and economic sustainability.

[7] X. Li, H. Wang, H. Abdelrahman, A. Kelly, L. Roy, and L. Wang, "Profiling and source tracking of the microbial populations and resistome present in fish products," *International Journal of Food Microbiology*, vol. 413, p. 110591, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168160524000357

[8] K. Zander, F. Daurès, Y. Feucht, L. Malvarosa, C. Pirrone, and B. le Gallic, "Consumer perspectives on coastal fisheries and product labelling in france and italy," *Fisheries Research*, vol. 246, p. 106168, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165783621002964

[9] N. A. Hasan, R. D. Heal, A. Bashar, A. L. Bablee, and M. M. Haque, "Impacts of covid-19 on the finfish aquaculture industry of bangladesh: A case study," *Marine Policy*, vol. 130, p. 104577, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308597X21001883

[10] J. A. Guajardo, R. Weber, and J. Miranda, "A model updating strategy for predicting time series with seasonal patterns," *Applied Soft Computing Journal*, vol. 10, 2010.

[11] A. Somers, J. Stevenson, A. Gudmundsson, and G. Onder, "Ss2.01: Prescriptions across borders: a multifaceted, multidisciplinary approach to adverse drug reactions," *European Geriatric Medicine*, vol. 5, 2014.

[12] A. A. Bery, "Development of soil cohesion and friction angle models using multiple linear regression (mlr) statistical techniques," *Indonesian Journal on Geoscience*, vol. 10, 2023.

[13] S. A. Lafta and M. Q. Ismael, "Trip generation modeling for a selected sector in baghdad city using the artificial neural network," *Journal of Intelligent Systems*, vol. 31, 2022.

[14] K. N. Çerçi and E. Hürdoğan, "Comparative study of multiple linear regression (mlr) and artificial neural network (ann) techniques to model a solid desiccant wheel," *International Communications in Heat and Mass Transfer*, vol. 116, 2020.

[15] H. Nourali and M. Osanloo, "Mining capital cost estimation using support vector regression (svr)," *Resources Policy*, vol. 62, 2019.

[16] C. E. Boyd, A. A. McNevin, and R. P. Davis, "The contribution of fisheries and aquaculture to the global protein supply," *Food Security*, vol. 14, 2022.

[17] S. B. Longo, B. Clark, R. York, and A. K. Jorgenson, "Aquaculture and the displacement of fisheries captures," *Conservation Biology*, vol. 33, 2019.

[18] Republic of Indonesia, "Law of the Republic of Indonesia Number 31 of 2004 on Fisheries," 2004, this law regulates the cultivation of fish and other aquaculture practices, including guidelines for sustainable development and management of fishery resources.

[19] F. S. Kibenge, B. Baldisserotto, and R. S.-M. Chong, *Aquaculture toxicology.* Academic Press, 2020.

[20] E. Mikkelsen, P. B. Sørdahl, and A. M. Solås, "Transparent and consistent? aquaculture impact assessments and trade-offs in coastal zone planning in norway," *Ocean and Coastal Management*, vol. 225, 2022.

[21] I. Kvalvik and R. Robertsen, "Inter-municipal coastal zone planning and designation of areas for aquaculture in norway: A tool for better and more coordinated planning?" *Ocean and Coastal Management*, vol. 142, 2017.

[22] A. Gimpel, V. Stelzenmüller, S. Töpsch, I. Galparsoro, M. Gubbins, D. Miller, A. Murillas, A. G. Murray, K. Pınarbaşı, G. Roca, and R. Watret, "A gis-based tool for an integrated assessment of spatial planning trade-offs with aquaculture," *Science of the Total Environment*, vol. 627, 2018.

[23] S. Larson, M. A. Rimmer, S. Hoy, and S. Thay, "Is the marine finfish cage farming value chain in cambodia inclusive?" *Aquaculture*, vol. 549, 2022.

[24] Y. Susanti, Z. Pramudia, A. Amin, L. Salamah, A. Yanuar, and A. Kurniawan, "Peningkatan produksi pangan melalui sistem integrasi teknologi aquaponics-recirculating aquaculture system (a-ras) pada budidaya ikan lele di desa kaliuntu kabupaten tuban," *Rekayasa*, vol. 14, pp. 121–127, 04 2021.

[25] Food and Agriculture Organization of the United Nations (FAO), *The State of World Fisheries and Aquaculture: Meeting the Sustainable Development Goals.* Rome: FAO, 2018. [Online]. Available: http://www.fao.org/3/i9540en/i9540en.pdf

[26] C. Akbar, D. S. C. Utomo, S. Hudaidah, and A. Setyawan, "Feed time and quantity management in increase growth rate and survival rate of snakehead fish farming, channa striata (bloch, 1793)," *Journal of Aquatropica Asia*, vol. 5, 2020.

[27] D. Torres-Salinas, Álvaro Cabezas-Clavijo, R. Ruiz-Pérez, and E. D. López-Cózar, "State of the library and information science blogosphere after social networks boom: A metric approach," *Library and Information Science Research*, vol. 33, 2011.

[28] J. E. Jelovsek, K. Chagin, M. Gyhagen, S. Hagen, D. Wilson, M. W. Kattan, A. Elders, M. D. Barber, B. Areskoug, C. MacArthur, and I. Milsom, "Predicting risk of pelvic floor disorders 12 and 20 years after delivery," *American Journal of Obstetrics and Gynecology*, vol. 218, 2018.

[29] P. Senawongse, A. R. Dalby, and Z. R. Yang, "Air quality prediction by machine learning methods," *the University of British Columbia*, 2015.

[30] J. Khoo, S. Haw, N. Su, and S. Mulaafer, "Kiwi fruit iot shelf life estimation during transportation with cloud computing," in *3rd IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET 2021*, 2021.

[31] K. Tjahjadi and M. S. Uria, "The influence of compensation, organization culture and motivation on employee performance," *Media Bisnis*, vol. 13, 2021.

[32] A. Rahbari, T. R. Josephson, Y. Sun, O. A. Moultos, D. Dubbeldam, J. I. Siepmann, and T. J. Vlugt, "Multiple linear regression and thermodynamic fluctuations are equivalent for computing thermodynamic derivatives from molecular simulation," *Fluid Phase Equilibria*, vol. 523, 2020.

[33] P. Ariwala, "9 real-world problems that can be solved by machine learning," 2023.

[34] Z. Subecz, "Web-development with laravel framework," *Gradus*, vol. 8, 2021.

[35] A. K. Wicaksono, F. Nurrakhman, and S. Samidi, "Comparation of distributed database model by clustering method in e-government system. study at kemenkeu ri," *Jurnal Teknik Informatika (Jutif)*, vol. 4, 2023.

[36] S. Sotnik, V. Manakov, and V. Lyashenko, "Overview: Php and mysql features for creating modern web projects," 2023.