JSCE (Journal of System and Computer Engineering)
Vol.7, No.1, January 2026, pp. 23-32
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i1.2270

Prediction of Protein Content of Carp Floss Based on Physical Characteristics and Processing Process Using Random Forest Regression Method

Irene Devi Damayanti [1,a,*]; Muhammad Sofwan Adha [2,b]; Lisna Junita Pairunan [3,c]
[1,2,3] Universitas Kristen Indonesia Toraja, Jln. Nusantara No. 12, Makale 91811, Indonesia
[a] irenedamayanti@ukitoraja.ac.id; [b] msofwan@ukitoraja.ac.id; [c] lisnajunita0020@gmail.com
** Corresponding author*

Article history: Received: October 03, 2025; Revised: November 13, 2021; Accepted: January 30, 2026

Abstract

Fish is one of the animal-based food sources that plays an important role in providing human nutrition. Fish meat is rich in macronutrients and micronutrients such as protein, fat, vitamins, and minerals. Protein is the dominant component after water, with a content of around 20%, making fish a potential source of animal protein. One type of fish with high protein content is carp (Cyprinus Carpio). However, carp that is not immediately processed after cultivation can deteriorate, so it is necessary to diversify processed products such as shredded carp. Shredded fish is a processed product made from shredded meat that is seasoned and fried until dry, with a distinctive taste and longer shelf life. In this study, the protein content of carp floss was predicted based on physical characteristics and processing parameters using the Random Forest Regression method. The input variables included moisture content, ash content, fat content, crude fiber, frying temperature, and frying time, while the output was protein content. The model was evaluated using MAE, MSE, RMSE, and R^2. The results showed MAE = 0.205, MSE = 0.051, RMSE = 0.225, and R^2 = 0.788, indicating a fairly high level of prediction accuracy. Thus, the Random Forest Regression method proved to be effective in predicting the protein content of carp abon and has the potential to be applied in quality control and optimization of fish processing.

Keywords—*Carp Floss, Protein Content, Random Forest Regression, Prediction, Food Processing.*

1. Introduction

Fish plays a significant role as an animal-based food source in providing nutrition for human life. The word “nutrition” comes from the Arabic word gizawi, which means nourishment. Fish meat has high nutritional value because it contains macronutrients and micronutrients that are important for humans, such as protein, fat, limited amounts of carbohydrates, vitamins, and minerals. Protein is the most dominant component in fish after water, with a significant amount, so fish can be considered a potential source of animal protein (Andhikawati et al., 2021). Fish protein has the highest nutritional content compared to other animal protein sources, reaching 20% (Damongilala, 2021). One type of fish with high nutritional content is carp (*Cyprinus Carpio*) (Fuadi & Surnaherman, 2017).

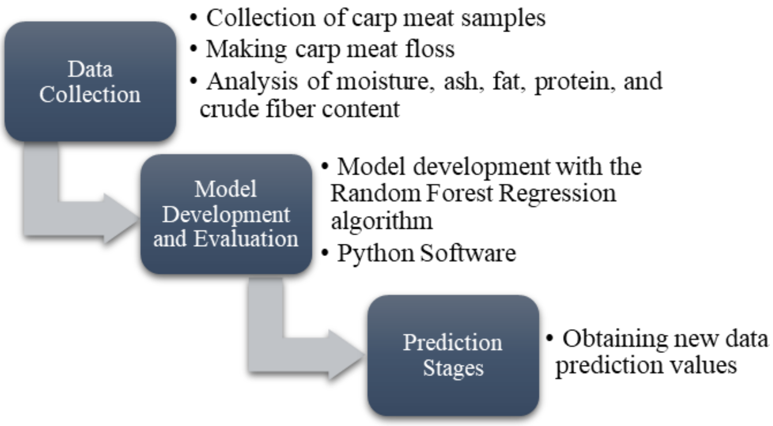
The results of carp farming that are not sold or consumed can cause carp products to be wasted due to damage and changes in taste over time. Carp, as a high source of protein, is very susceptible to spoilage if not processed carefully. Therefore, the solution to this problem is to diversify processed carp products. The right approach in this case is to optimize carp as a raw material for shredded carp (Pasinggi et al., 2023).

Abon is a dish made from shredded meat, mixed with various spices and then fried. The basic principle of fish abon involves a preservation method that combines boiling or steaming, frying, and the addition of certain spices (Aditya et al., 2016). The result is a product with a soft texture and distinctive taste and aroma. One of the purposes of processed fish products, such as fish floss, is that they are more practical to consume.

The application of computer software in food nutrition identification systems has become an issue in recent years. This is due to the significant impact of food consumption on health. One of the most popular and widely discussed machine learning algorithms is Random Forest (Forest & Learning, n.d.; Lu & Hardin, 2021; Ramosaj, 2021; Wang et al., 2023). Random Forest Regression is a machine learning method that has been proven effective and widely used as a standard approach for tabular data analysis, and has good predictive performance (Adiyati, 2021; Urrochman et al., 2025).

2. Method

This study predicts the protein content of carp floss based on physical characteristics and processing methods. Predictions are made using the Random Forest Regression algorithm by first dividing the dataset into two parts, namely training data and testing data. Prediction error calculations focus on the MSE matrix size and are supported by other matrix sizes, such as RMSE, MAE, and R-Squared. Data processing was performed using Python software. Broadly speaking, there were three stages in this study, as shown in Figure 1.



Data Collection
• Collection of carp meat samples
• Making carp meat floss
• Analysis of moisture, ash, fat, protein, and crude fiber content
Model Development and Evaluation
• Model development with the Random Forest Regression algorithm
• Python Software
Prediction Stages
• Obtaining new data prediction values

Figure 1. Conceptual Model.

The flowchart for this study can be seen in Figure 2 below:

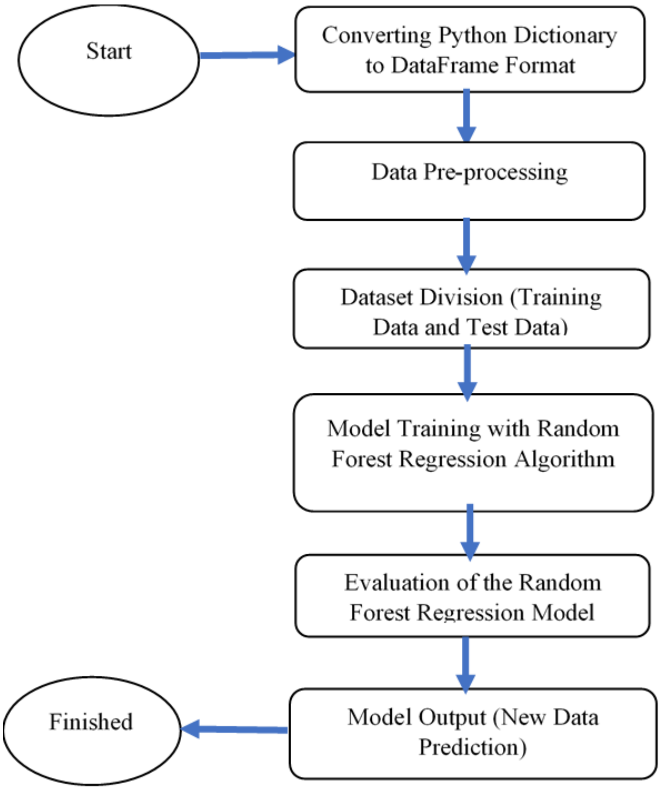

Start
Converting Python Dictionary to DataFrame Format
Data Pre-processing
Dataset Division (Training Data and Test Data)
Model Training with Random Forest Regression Algorithm
Evaluation of the Random Forest Regression Model
Model Output (New Data Prediction)
Finished

Figure 2. Research Flow Chart.

2.1 Data Collection

The initial stage of the research began with the collection of carp meat samples to be used as the main raw material for the research. The samples were then processed into shredded carp meat using standard processing methods. Next, a proximate analysis was conducted on the shredded carp meat to determine its proximate components, including moisture content, ash content, fat content, protein content, and crude fiber content. Proximate analysis is an indicator for determining the quality of a food product (Handhini Dwi Putri et al., 2022). Each test parameter is calculated using the following equation (Herson et al., 2023):

$$Water\ Content\ (\%) = \frac{\mathrm{B-C}}{\mathrm{B-A}} \times 100\% \tag{1}$$

where,
A = weight of empty cup (grams)
B = weight of cup + initial sample (grams)
C = weight of cup + final sample (grams)

$$Ash\ Content\ \ (\%) \quad = \frac{Ash\ Weight\ (grams\ )}{Sample\ \ Weight\ (grams\ )} \times 100\% \tag{2}$$

$$Fat\ Content\ \ (\%) \ = \frac{Fat\ Weight\ (grams\ )}{Sample\ \ Weight\ (grams\ )} \times 100\% \tag{3}$$

$$Protein\ Content\ (\%) \ = \frac{(\mathrm{V_A - V_B})\ \mathrm{HCl}\ \times \mathrm{NHCl}\ \ \times 14.007\ \times 6.25\ \times 100\%}{\mathrm{W} \times 100} \times 100\% \tag{4}$$

where,
V_A = ml HCl for sample titration

V_B = ml HCl for blank titration
N = Normality of HCl standard used
14.007 = Atomic weight of nitrogen
6.25 = Protein conversion factor for fish floss
W = Sample weight (g)

$$Crude\ Fiber\ Content(\%) = \frac{residual\ \ weig\,ht\ after\ \ digestion\ \ (grams\ ) - ash\ weig\,ht\ (grams\ )}{sample\ \ weig\,ht\ (grams\ )} \times 100\% \quad (5)$$

2.2 Data Preprocessing

Before proceeding to the model training stage, the dataset is first processed through a data pre-processing stage. The data pre-processing stage is carried out to ensure that the data to be used in modeling meets the required criteria and is of sufficient quality. Normalization is not performed because the model used is Random Forest Regression, which is a decision tree-based model that is not sensitive to data scale. Therefore, variations in scale between features do not significantly affect model performance. In addition, the data used is complete and clean. No duplicates or empty values were found, so no data entry imputation or deletion was required. Therefore, all data pre-processing stages have been completed successfully, and the data is now ready for the model training stage. Next, the data is organized in tabular form to facilitate further data manipulation and analysis.

2.3 Dataset Division (Training Data and Test Data)

After the pre-processing stage, the dataset is divided into two parts, namely training data and testing data. The data used for model training is separated from the data used to test the model's performance. The purpose of dividing the data is to measure how well the model can generalize to new data.

2.4 Random Forest Regression Model Training

The model training stage then continues after going through the pre-processing stage and the dataset division stage. This study uses the Random Forest Regression algorithm as one of the decision tree-based ensemble learning methods that is often used for regression. During this process, the model learns how input features interact with the target output. This stage is the basis of the modeling process, where the algorithm learns from previous data before being applied to new data.

2.5 Evaluation of Random Forest Regression Models

After training the Random Forest Regression model using training data, the next step is to evaluate the model's performance using test data. The purpose of this step is to determine how accurately the model can predict new data based on previously learned patterns. Model evaluation is carried out by calculating several regression matrix values, as follows (Ihzaniah et al., 2023; Tatachar, 2021):

2.5.1.1 Mean Absolute Error *(MAE)*

MAE is the average absolute difference between the actual value and the predicted value. The smaller the MAE value, the better the model performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

where,
y_i = actual value
$\hat{y}_i$ = predicted value
n = sample size

2.5.1.2 Mean Squared Error *(MSE)*

The MSE value is obtained from the average square difference between the actual value and the predicted value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \quad (7)$$

The smaller the MSE value, the closer the model prediction is to the actual value.

2.5.1.3 Root Mean Squared Error *(RMSE)*

RMSE is the square root of MSE. The RMSE value indicates the average prediction error, but is more sensitive to large errors.

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

2.5.1.4 R-squared *(*R^2*)*

The R^2 value indicates how much of the target data variation can be explained by the model, with values ranging from 0 to 1. If the R^2 value is closer to 1, it indicates that the model explains the data variation very well.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

where $\bar{y}$ is the actual average value.

2.6 New Data Predictions

After going through the evaluation stage, the next step is to predict new data that is not yet known from the previous sample data. To ensure that the prediction results remain contextually relevant, this data is constructed by combining feature values that are still within the training data interval.

3. Results And Discussion

In this study, the researchers developed a prediction model using the Jupyter Notebook platform with the Python programming language. The researchers will explain the results of the study in this section.

3.1 Data Preprocessing

Based on the results of the proximate test analysis of carp floss, all proximate components are in accordance with SNI standards. This can be seen in Table 1,

Table 1. Proximate Test Results for Carp Floss.

Type Sample	Water Content (%)	Ash Content (%)	Protein Content (%)	Fat Content (%)	Coarse Fiber (%)	Frying Temperature (°C)	Frying Time (menit)
Abon 1	12.7974	3.6504	20.22	13.95	7.60	91	55
Abon 2	13.0761	3.4842	20.49	14.24	8.67	93	50
Abon 3	13.0391	4.0095	21.11	13.74	8.31	109	32
Abon 4	13.0181	3.6977	20.66	14.18	7.41	111	29
Abon 5	12.8308	3.7346	20.78	14.10	8.14	97	48
Abon 6	13.1697	3.5978	20.69	14.16	8.36	99	40
Abon 7	13.0893	3.8276	20.29	14.24	7.70	91	32
Abon 8	12.9194	3.7482	20.57	14.30	7.51	106	57
Abon 9	12.9322	3.9033	21.36	13.86	8.08	111	39
Abon 10	13.0233	3.8823	21.21	13.84	8.59	112	48
Abon 11	12.7987	3.8877	20.93	13.75	8.27	102	45
Abon 12	12.9531	3.7565	20.15	14.29	9.23	99	42
Abon 13	12.9554	3.5718	20.16	14.13	8.44	106	45
Abon 14	12.9308	3.6813	20.35	14.44	7.17	108	43
Abon 15	13.0997	3.5919	20.23	14.05	9.02	95	42
Abon 16	12.9437	3.9201	21.29	13.77	8.50	93	34
Abon 17	12.8819	3.9532	21.08	13.80	8.81	106	43
Abon 18	12.8699	3.7276	20.31	14.26	7.45	91	34
Abon 19	13.1031	3.5156	20.37	14.01	8.62	93	44
Abon 20	13.0003	3.6475	20.23	14.22	7.23	96	47
Abon 21	12.8082	3.6260	20.30	14.25	7.77	119	48
Abon 22	13.0862	3.5273	20.14	14.32	8.76	94	51
Abon 23	12.8980	3.5080	20.49	14.19	8.55	91	32
Abon 24	13.0109	3.5676	20.39	14.15	9.08	96	56
Abon 25	12.8507	3.9621	21.31	13.90	8.47	91	38
Abon 26	12.9607	3.5847	20.63	14.12	8.72	120	41
Abon 27	12.9566	3.8726	21.43	13.82	8.48	93	21
Abon 28	12.9662	3.7123	20.57	14.36	7.51	98	22
Abon 29	12.9919	3.9343	20.75	13.77	8.46	113	32
Abon 30	12.8200	3.7288	20.42	14.03	7.53	115	24
Abon 31	12.9539	3.8749	20.99	13.93	8.22	91	40
Abon 32	12.9514	3.8899	21.58	13.90	8.46	95	56
Abon 33	12.9075	3.7093	19.97	14.22	7.37	90	42
Abon 34	12.9097	3.8455	21.30	13.71	8.09	120	60
Abon 35	12.8317	3.5265	20.66	13.99	8.61	115	45
Abon 36	12.8558	3.9172	21.05	13.95	8.50	107	36
Abon 37	12.7883	3.6795	20.34	14.08	7.17	93	41
Abon 38	12.9989	3.8107	20.67	14.29	7.65	99	22
Abon 39	13.1539	3.6023	20.67	14.15	8.86	101	40
Abon 40	12.9208	3.6007	20.40	14.27	8.62	112	42
Abon 41	12.7973	3.8688	21.59	13.69	8.44	112	22
Abon 42	12.8721	3.8912	21.42	13.91	8.35	116	54

Abon 43	13.0777	3.5703	20.55	13.97	8.56	114	40
Abon 44	13.1460	3.7977	20.74	14.36	7.75	113	49
Abon 45	12.8644	3.9332	20.58	13.88	8.33	102	54
Abon 46	13.0616	3.9162	21.34	13.94	8.44	101	20
Abon 47	13.0499	3.6156	20.73	14.08	8.61	107	45
Abon 48	13.1049	3.8169	20.45	14.27	7.90	103	60
Abon 49	12.8611	3.9650	21.19	14.04	8.46	103	30
Abon 50	12.9611	3.5846	20.53	13.96	8.64	110	27

The data from the proximate analysis in Table 1 was used as a dataset for the modeling process. The Python libraries used in this study were pandas, numpy, scikit-learn (sklearn), and matplotlib. The dataset used consists of 50 types of data and 7 features, as shown in Table 1. The dataset from each data with 6 input features, namely: water content, ash content, fat content, coarse fiber, frying temperature, and frying time, and 1 target output, namely protein content. The first step was to convert the data from a Python dictionary to a DataFrame format.

Next, separate the input features and target outputs. Variable x is the input feature and variable y is the target output (protein content). This step is carried out for the prediction model training process, where the model will use the input features to be mapped to the target output.

3.2 Data Division (Training Data and Test Data)

After the pre-processing stage, the dataset was divided into two parts consisting of: training data (80%) and testing data (20%). This division is done randomly, but consistently by setting the random_state value to allow replication of results. This dataset only consists of 50 data points, so 80:20 ratio is considered balanced to maintain sufficient data availability for training and testing.

3.3 Random Forest Regression Model Training

Next is the model training stage with Random Forest Regression. The model was built with the basic parameter n_estimators=100, which indicates that the model consists of 100 decision trees. Random Forest combines the results of many trees to produce more reliable predictions, where these trees are then selected to improve the accuracy and stability of the predictions. Additionally, the random_state=42 parameter is also used to ensure that the training results can be replicated at each stage of execution.

The model training stages are carried out using the fit() function on the model object. This is done using the training data x_train and training target y_train that have been provided previously. During this process, the model learns how input features such as moisture content, ash, fat, crude fiber, temperature, and frying time interact with the output, namely protein content. After the training process is complete, the statement “model has been trained” appears as the program output. This indicates that the training process is complete and the model is ready to be used to predict protein content based on new input data. This stage is the basis of the modeling process, where the algorithm learns from previous data before being applied to new data.

3.4 Evaluation of the Random Forest Regression Model

After the model training stage, the next stage is the evaluation stage of model performance with test data. The model performance values were obtained from the regression matrix calculations as shown in Table 2.

Table 2. Model Performance Score.

Matrix	Value
Mean Absolute Error (MAE)	0.205
Mean Squared Error (MSE)	0.051
Root Mean Squared Error (RMSE)	0.225
R-squared (R^2)	0.788

Based on Table 2, the evaluation of the Random Forest Regression model's performance on carp fish floss protein content data produced an R-squared (R^2) value of 0.788, indicating that 78.8% of the variation in protein content can be explained by the input variables, namely water content, ash, fat, coarse fiber, frying temperature, and frying time. This value indicates that the model has good predictive ability in estimating protein content based on chemical parameters and process conditions. In addition, a Mean Absolute Error (MAE) value of 0.205 was obtained, which indicates that the average absolute error between the predicted value and the actual value is relatively small. The Mean Squared Error (MSE) value of 0.051 and the Root Mean Squared Error (RMSE) of 0.225 further reinforce these results, as both metrics indicate a low level of prediction deviation. Overall, the evaluation results show that the Random Forest Regression algorithm is capable of estimating the protein content of carp fish floss with a fairly high level of accuracy and minimal prediction error. Therefore, this method is considered effective and reliable in the process of predicting protein content based on chemical characteristics and processing parameters.

3.5 New Data Predictions

Figure 3 shows a graph comparing the actual and predicted values of carp fish floss protein content generated by the Random Forest Regression model. Based on the graph, it can be seen that the prediction line (dashed red) follows the trend of the actual value line (blue), indicating that the model is able to represent the relationship between input and output variables quite well.

Visually, the difference between the actual and predicted values is relatively small in most of the test samples. There are only a few deviations at several sample points, indicating minor variations due to nonlinear factors in the data or limitations in the number of training samples. Nevertheless, the predicted values remain within a range close to the actual values, suggesting that the Random Forest model has stable and accurate predictive capabilities for the protein content of carp floss.

These results are in line with the quantitative evaluation values obtained previously, namely R^2 of 0.788, MAE of 0.205, MSE of 0.051, and RMSE of 0.225. These values indicate that the model has a low prediction error rate and good generalization ability. Thus, the Random Forest Regression approach can be used effectively to estimate protein content based on the chemical characteristics and processing parameters of carp fish floss.

Figure 4 shows a scatter plot between the actual and predicted values of carp fish floss protein content generated by the Random Forest Regression model. Each point on the graph represents a comparison between the measured protein content value (actual) and the model's estimated value (predicted). The dotted red line shows the ideal line with a slope of 45°, which describes the condition where the predicted value is equal to the actual value.

Based on the data distribution in the graph, it can be seen that most of the data points are around the diagonal line. This shows that the predicted values produced by the model are very close to the actual values, which means that the prediction error rate is relatively small. The distribution of points that are not too far from the diagonal line also indicates that the model is

DOI : https://doi.org/10.61628/jsce.v7i1.2270

able to represent the relationship pattern between input variables (such as water content, fat, crude fiber, ash, temperature, and frying time) and protein content consistently.

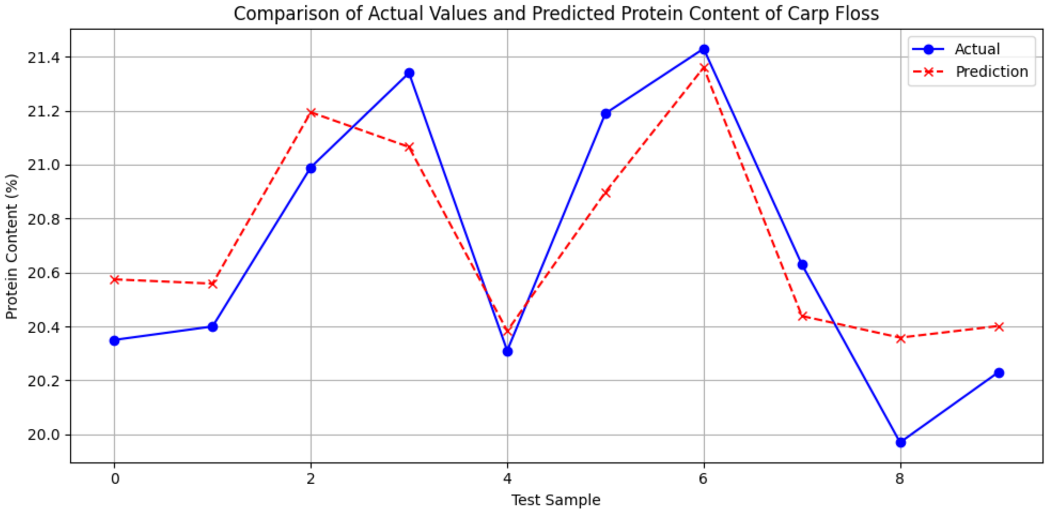


Figure 3. Comparison of Actual Values and Predicted Protein Content of Carp Floss.

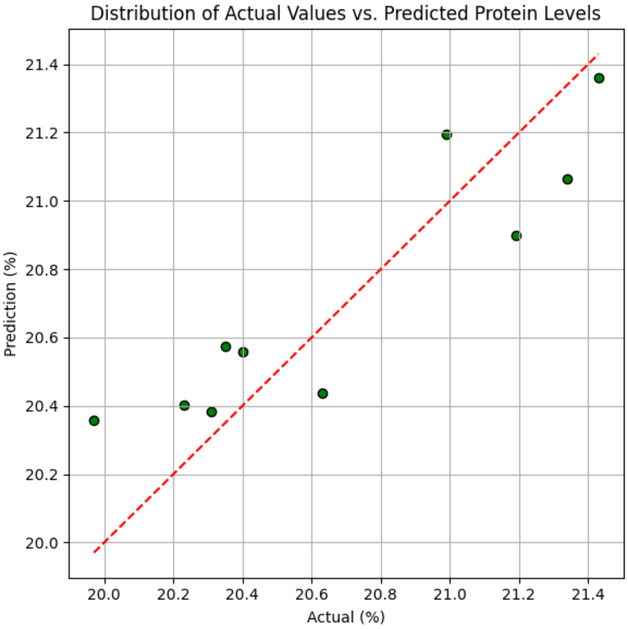


Figure 4. Distribution of Actual Values vs. Predicted Protein Levels.

4. Conclusions

Based on the results of the research and model evaluation, the Random Forest Regression method showed good ability in predicting the protein content of carp abon using input variables derived from physical characteristics and processing parameters, namely moisture content, ash content, fat content, crude fiber, frying temperature, and frying time. The model evaluation results show an MAE value of 0.205, an MSE of 0.051, an RMSE of 0.225, and an R^2 of 0.788. These values indicate that the model has a high level of accuracy and reliability in making predictions. The similarity between the actual values and the predicted values shown in the comparison graph and the scatter graph also reinforces that the Random Forest algorithm is able to effectively capture the non-linear relationship between the input variables and the protein content. Thus, it can be concluded that the Random Forest Regression method is an appropriate and reliable approach for predicting the protein content of carp fish flakes, and has the potential to be used in optimizing product quality and processing parameters in future research and industrial food applications.

References

Aditya, H. P., Herpandi, H., & Lestari, S. (2016). Karakteristik Fisik, Kimia dan Sensoris Abon Ikan dari Berbagai Ikan Ekonomis Rendah. *Jurnal FishtecH*, *5*(1), 61–72. https://doi.org/10.36706/fishtech.v5i1.3519

Adiyati, P. A. (2021). *Implementasi Algoritma Random Forest*. *8*(1), 70–73.

Andhikawati, A., Junianto, J., Permana, R., & Oktavia, Y. (2021). Review: Komposisi Gizi Ikan Terhadap Kesehatan Tubuh Manusia. *Marinade*. https://doi.org/10.31629/marinade.v4i02.3871

Damongilala, L. J. (2021). Kandungan Gizi Pangan Ikan. *Patma Media Grafindo Bandung*, 1–60.

Forest, R., & Learning, M. (n.d.). *Random Forest in Machine Learning Random Forest in Machine Learning*. 2–4.

Fuadi, M., & Surnaherman, S. (2017). Cara Pengawetan Ikan Mas (Cyprinus carpio L) Dengan Menggunakan Fermentasi Limbah Kubis (Brassica oleracea). *Agrintech: Jurnal Teknologi Pangan Dan Hasil Pertanian*, *1*(1), 55–63. https://doi.org/10.30596/agrintech.v1i1.1669

Handhini Dwi Putri, Haninah, Elfidasari, D., & Sugoro, I. (2022). Kandungan Nutrisi Abon Ikan Sapu-Sapu (Pterygoplichthys Pardalis) Asal Sungai Ciliwung, Indonesia. *Jurnal Pengolahan Pangan*, *7*(1), 14–19. https://doi.org/10.31970/pangan.v7i1.62

Herson, N. A., Sumual, M. F., Rumengan, I. F. M., Pongoh, J., & Mandey, L. C. (2023). Proximate Analysis Of Collagen Cockatoa Fish Scales (Scarus sp.). *Jurnal Agroekoteknologi Terapan*, *4*(2), 428–433. https://doi.org/10.35791/jat.v4i2.52484

Ihzaniah, L. S., Setiawan, A., & Wijaya, R. W. N. (2023). Perbandingan Kinerja Metode Regresi K-Nearest Neighbor dan Metode Regresi Linear Berganda pada Data Boston Housing. *Jambura Journal of Probability and Statistics*, *4*(1), 17–29. https://doi.org/10.34312/jjps.v4i1.18948

Lu, B., & Hardin, J. (2021). A unified framework for random forest prediction error estimation. *Journal of Machine Learning Research*, *22*.

Pasinggi, E. S., Damayanti, I. D., & Kannapadang, S. (2023). *Pendampingan UMKM Barrent Foods , Desa Wisata Randanan untuk Peningkatan Kapasitas Produksi dan Pemasaran Abon Ikan Mas*. *1*(10), 2547–2552.

Ramosaj, B. (2021). *Interpretable Machines: Constructing Valid Prediction Intervals with Random Forests*. http://arxiv.org/abs/2103.05766

Tatachar, A. V. (2021). Comparative assessment of regression models based On model evaluation metrics. *International Research Journal of Engineering and Technology*, *8*(9), 853–860. https://d1wqtxts1xzle7.cloudfront.net/73250877/IRJET_V8I9127-libre.pdf

Urrochman, M. Y., Asy'ari, H., & Hizham, F. A. (2025). Performance Comparison of Random Forest Regression, Svr Models in Stock Price Prediction. *Jurnal Pilar Nusa Mandiri*, *21*(1), 16–23. https://doi.org/10.33480/pilar.v21i1.6072

Wang, Y., Wu, H., & Nettleton, D. (2023). Stability of Random Forests and Coverage of Random-Forest Prediction Intervals. *Advances in Neural Information Processing Systems*, *36*(1), 1–28.