JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6409

THE ROLE OF L1 REGULARIZATION IN ENHANCING LOGISTIC REGRESSION FOR EGG PRODUCTION PREDICTION

Nur Alamsyah[1*]; Budiman[2]; Elia Setiana[3]; Valencia Claudia Jennifer[4]

Informatic System[1,4]; Informatika[2,3]
Universitas Informatika Dan Bisnis Indonesia, Indonesia [1,2,3,4]
https://unibi.ac.id/[1,2,3,4]
nuralamsyah@unibi.ac.id[1*], budiman@unibi.ac.id[2], elia.setiana@unibi.ac.id[3], valencia@unibi.ac.id[4]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Poultry egg productivity is strongly influenced by various environmental factors, such as air and water quality. However, accurately predicting productivity remains a challenge due to the complex interplay of multiple environmental variables and the risk of overfitting in predictive models. This study improves egg productivity prediction using Logistic Regression with L1 regularization, which enhances model generalization by performing automatic feature selection. The research methodology includes data collection of key environmental indicators—Air Quality Index (AQI), Water Quality Index (WQI), and Humidex—followed by data preprocessing, exploratory data analysis (EDA), and model training using L1-regularized Logistic Regression. Model evaluation was performed using classification metrics and learning curve analysis to assess stability and effectiveness. Experimental results indicate that Logistic Regression without regularization achieved an accuracy of 90.7%, with misclassification occurring in the lower production categories. By applying L1 regularization, accuracy increased significantly to 97%, demonstrating its ability to reduce overfitting while improving classification performance. This study also compares its findings with previous research, such as De Col et al. (wheat epidemic prediction, 80–85% accuracy) and Jia Q1 et al. (heart disease prediction, 92.39% accuracy), confirming that our approach outperforms prior Logistic Regression models in similar predictive tasks. These findings suggest that L1 regularization is an effective solution for improving egg productivity prediction, particularly in scenarios with complex environmental influences. Future work will explore advanced ensemble learning techniques and real-time IoT-based monitoring to further enhance prediction accuracy and practical applicability.*

Keywords: *egg production prediction, environmental factors, logistic regression, l1 regularization.*

Intisari— *Produktivitas telur ayam sangat dipengaruhi oleh berbagai faktor lingkungan, seperti kualitas udara dan air. Namun, memprediksi produktivitas dengan akurat tetap menjadi tantangan karena kompleksitas interaksi variabel lingkungan serta risiko overfitting dalam model prediksi. Penelitian ini meningkatkan prediksi produktivitas telur dengan menggunakan Regresi Logistik dengan Regularisasi L1, yang meningkatkan generalisasi model melalui seleksi fitur otomatis. Metodologi penelitian mencakup pengumpulan data dari indikator lingkungan utama—Indeks Kualitas Udara (AQI), Indeks Kualitas Air (WQI), dan Humidex—dilanjutkan dengan preprocessing data, analisis eksploratif data (EDA), dan pelatihan model menggunakan Regresi Logistik dengan Regularisasi L1. Evaluasi model dilakukan menggunakan metrik klasifikasi dan analisis kurva pembelajaran untuk mengukur stabilitas serta efektivitas model. Hasil eksperimen menunjukkan bahwa Regresi Logistik tanpa regularisasi menghasilkan akurasi sebesar 90,7%, dengan kesalahan klasifikasi yang dominan pada kategori produksi rendah. Dengan menerapkan Regularisasi L1, akurasi meningkat secara signifikan menjadi 97%, membuktikan kemampuannya dalam mengurangi overfitting sekaligus meningkatkan performa klasifikasi. Studi ini juga membandingkan temuannya dengan penelitian sebelumnya, seperti penelitian De Col et al. (prediksi epidemi gandum, akurasi 80–85%) dan Jia Q1 et al. (prediksi penyakit jantung, akurasi 92,39%), yang mengonfirmasi bahwa pendekatan kami mengungguli*

model Regresi Logistik sebelumnya dalam tugas prediksi yang serupa. Temuan ini menunjukkan bahwa Regularisasi L1 merupakan solusi efektif untuk meningkatkan prediksi produktivitas telur, terutama dalam skenario yang dipengaruhi oleh faktor lingkungan yang kompleks. Penelitian selanjutnya akan mengeksplorasi teknik pembelajaran ansambel yang lebih canggih serta pemantauan berbasis IoT secara real-time guna lebih meningkatkan akurasi prediksi serta aplikabilitasnya di dunia nyata.

***Kata Kunci**: Prediksi Produktivitas Telur, Faktor Lingkungan, Regresi Logistik, Regularisasi L1*

INTRODUCTION

In the modern poultry industry, maintaining high egg production is crucial for the profitability and sustainability of operations [1]. Egg production is influenced by a wide range of factors, particularly environmental conditions such as air quality, water quality, temperature, humidity, and seasonal changes [2]. These environmental stressors can cause fluctuations in egg productivity, impacting not only the quantity of eggs produced but also the overall health and well-being of poultry [3]. Poor environmental conditions often lead to increased disease susceptibility and stress, which in turn reduces egg yield, causing significant economic losses for farmers [4]. To address these challenges, advancements in data analytics and machine learning have provided new opportunities to optimize farm management practices [5]. Predictive modeling has become an essential tool for anticipating production outcomes based on external variables such as weather patterns, pollution levels, and water quality indicators [6]. By leveraging these environmental factors, predictive models can help farmers make informed decisions, ensuring optimal conditions for poultry and ultimately improving production outcomes [7].

In recent years, machine learning has emerged as a powerful tool for improving prediction accuracy in various agricultural applications, including poultry farming [8]. Traditional regression-based models have often struggled to handle the complex, nonlinear relationships between environmental variables and egg production, leading to suboptimal predictive performance [9]. One key limitation is the presence of high-dimensional data, where irrelevant or weakly correlated features can introduce noise, increasing the risk of overfitting in predictive models. This challenge has prompted researchers to explore feature selection techniques and regularization methods to enhance model robustness and interpretability [10]. Among various regularization techniques, L1 Regularization (Lasso Regression) has gained prominence for its ability to perform automatic feature selection by eliminating irrelevant predictors while retaining the most influential ones [11]. By applying L1 Regularization to Logistic Regression, this study aims to develop a more efficient predictive model for egg production that balances model accuracy and generalizability. Unlike traditional logistic regression models, which can suffer from high variance when dealing with correlated features, the L1-regularized model optimally selects key environmental indicators—such as Air Quality Index (AQI), Water Quality Index (WQI), and Humidex—to improve classification performance [12].

However, developing accurate and reliable models to predict egg production remains a challenge [13]. Traditional predictive models, such as Logistic Regression, have been widely used in various industries due to their simplicity and interpretability [14]. Nonetheless, when dealing with real-world agricultural datasets, these models often suffer from overfitting, where the model becomes too tailored to the training data and loses its ability to generalize well to new, unseen data [15]. This issue is particularly prevalent in datasets that contain a large number of features or complex, non-linear relationships between variables [16]. The problem arises from the difficulty in creating a model that can balance between capturing relevant patterns in the data while avoiding overfitting [17]. Logistic Regression, while useful, can struggle with large, noisy datasets unless properly regularized [18]. Overfitting not only reduces the predictive power of the model but also leads to inaccurate predictions, which can mislead farm management decisions [19]. L1 regularization offers a solution to this issue by shrinking the coefficients of less important features to zero, effectively simplifying the model and reducing the risk of overfitting [20]. This technique is particularly beneficial for datasets with numerous features, as it performs automatic feature selection, keeping only the most relevant predictors. By incorporating L1 regularization into Logistic Regression, the model becomes more robust, with improved generalization to new data [21].

Despite the proven advantages of regularization techniques, there has been limited research specifically examining the effectiveness of L1 regularization in improving egg production prediction using environmental data [22]. Given the

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

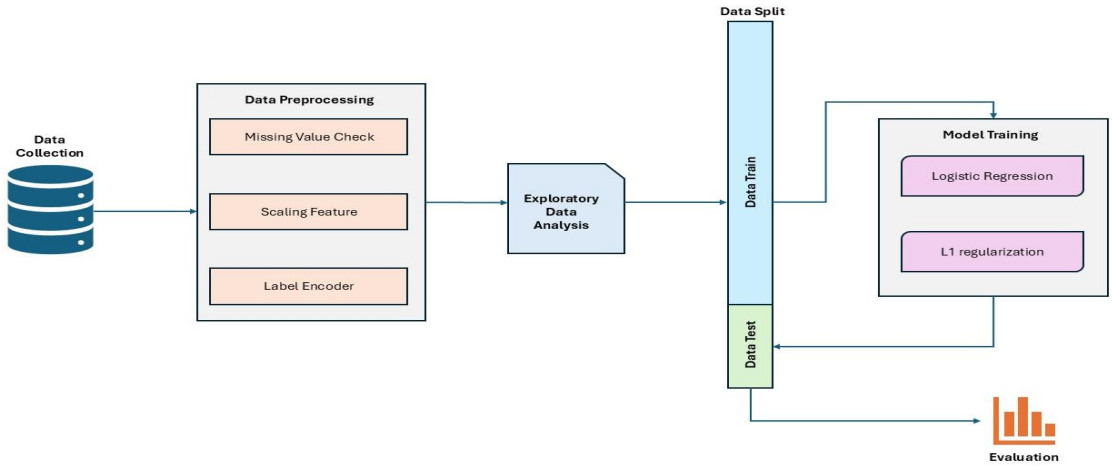
VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6409

complexity of agricultural environments and the variability of environmental factors, it is crucial to explore how L1 regularization can enhance predictive accuracy in this domain [23]. The main contributions of this study are twofold. First, it introduces the use of L1 regularization in Logistic Regression models for predicting egg production, effectively addressing the problem of overfitting commonly found in agricultural datasets. Second, this research demonstrates how L1 regularization significantly improves the model's ability to generalize to new, unseen data by performing automatic feature selection, leading to more accurate predictions in real-world agricultural environments.

MATERIALS AND METHODS

The methodology employed in this study consists of several essential steps, as illustrated in Figure 1. Figure 1 clearly outlines the step-by-step methodology, beginning with the collection of environmental data relevant to egg production, followed by essential preprocessing steps such as handling missing values, scaling, and label encoding. The diagram then moves to the exploratory data analysis (EDA), where correlations and patterns among variables are explored. This is followed by the model training phase, where Logistic Regression with L1 regularization is applied. Finally, the model is evaluated using performance metrics and learning curve analysis to assess generalization and prevent overfitting. This figure illustrates the overall methodological framework used in this study. The process begins with data collection, followed by preprocessing, exploratory data analysis (EDA), and dataset splitting. The Logistic Regression model is trained with L1 Regularization to enhance generalization and reduce overfitting. The final evaluation phase includes accuracy metrics and learning curve analysis to assess model stability and effectiveness [24]. The process begins with the data collection of environmental factors such as the Air Quality Index (AQI), Water Quality Index (WQI), alongside other variables like temperature and humidity. Once the data is gathered, preprocessing is applied to ensure the dataset is clean and ready for model training.

Preprocessing involves several key tasks: identifying and handling missing values, scaling numerical features for consistency, and converting the 'Production' column into numerical categories using a Label Encoder. After preprocessing, the dataset is divided into training and testing sets to facilitate the evaluation of the model's predictive performance. After that, Exploratory Data Analysis (EDA) was performed to examine the relationships between variables, visualize correlations, and identify outliers that may affect model accuracy. The next stage is to train the Logistic Regression model, utilizing L1 regularization to minimize overfitting and improve the model's ability to generalize effectively on unseen data.



Source: (Research Results, 2025)

Figure 1. Proposed Method

Finally, the model's performance is thoroughly assessed using a range of evaluation metrics, including accuracy, precision, recall, and f1-score. Additionally, a learning curve analysis is performed to evaluate how the model's performance improves as the training data size increases.

A. Dataset

The dataset used in this study, sourced from Kaggle, consists of 1501 records and includes various environmental and production variables that influence egg production in poultry farms. The humidex data, which is derived from temperature and humidity readings, was collected from both local meteorological stations and online weather services such as OpenWeatherMap and Weather.com, which provide historical and real-time weather data through APIs. For air quality data, measurements were obtained from two main sources: environmental monitoring agencies, which track air quality parameters like particulate matter and nitrogen dioxide, and on-farm air quality sensors, which specifically monitor pollutants in the poultry environment.

Water quality data was collected through on-site water testing, which involved regular checks for factors such as pH levels, nitrates, and bacterial contamination. Additionally, data was obtained from local water authorities, which provide detailed reports on the quality of water from municipal supplies or natural sources used in poultry farming. The egg production data was sourced from farm records, including daily or weekly logs, and was supplemented by poultry management systems, software that tracks various production metrics such as the number of eggs produced and their weight. This combination of diverse data sources provides a comprehensive understanding of the environmental factors impacting egg production and serves as the foundation for the predictive modeling used in this study. An explanation of these features is shown in Table 1.

Table 1. Dataset features

Feature	Description
AQI	Air Quality Index, measures the level of air pollution.
WQI	Water Quality Index, indicates the quality of water.
Humidex	Humidex, a combination of temperature and humidity.
Production	Egg production levels, categorized as 'High', 'Medium', or 'Low'.

Source: (Research Results, 2025)

B. Data Preprocessing

The data preprocessing phase was essential to ensure that the dataset was clean, standardized, and ready for model training [25]. First, the dataset was checked for any missing values, which could negatively impact the model's performance. For numerical features like AQI, WQI, and Humidex, missing values were handled through imputation, where they were replaced with the mean or median of the respective feature. This approach ensured that no data points were lost while maintaining the integrity of the dataset. Next, feature scaling was applied to the numerical variables to standardize their ranges. StandardScaler was used to transform the values of AQI, WQI, and Humidex to have a mean of 0 and a standard deviation of 1. This scaling process helped to prevent any one feature from dominating the learning process and improved the model's performance and convergence rate.

C. Exploratory Data Analysis (EDA)

The heatmap (shown in Figure 2) illustrates the correlation between various environmental variables such as Air Quality Index (AQI), Water Quality Index (WQI), Humidex, and the target variable, Production. Figure 2 reveals significant relationships among the environmental variables. A strong positive correlation (0.79) is observed between AQI and Humidex, indicating that higher air quality levels tend to be associated with warmer and more humid conditions. Conversely, there is a strong negative correlation (-0.88) between WQI and Humidex, implying that water quality tends to decrease as humidity and temperature rise. The correlation values in this heatmap provide crucial insights that inform the feature selection and model development process. The heatmap provides insight into the relationships among key environmental factors [26].

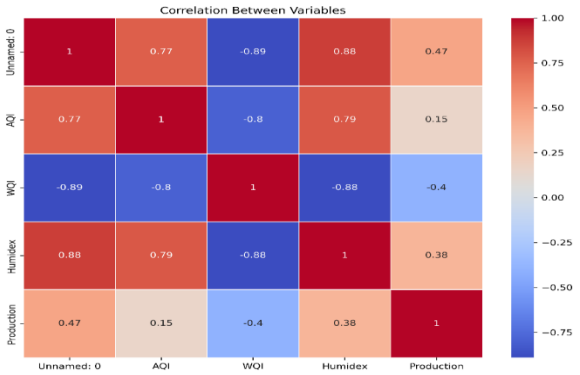
AQI and Humidex show a strong positive correlation (0.79), suggesting that improved air quality is associated with higher temperature and humidity levels. Meanwhile, WQI and Humidex exhibit a strong negative correlation (-0.88), indicating that lower water quality is linked to more humid conditions. Additionally, the weak correlation between AQI and production (0.15) suggests that air quality has a relatively minor impact on egg production compared to other variables. Correlation values span from -1 to 1, with values approaching 1 representing a strong positive correlation, values nearing -1 indicating a strong negative correlation, and values close to 0 suggesting minimal or no correlation between the variables. [27].

From the heatmap, we can observe that AQI and Humidex show a relatively strong positive correlation (0.79), suggesting that as air quality improves, the Humidex index, which reflects temperature and humidity, tends to increase. On the other hand, there is a strong negative correlation between WQI and Humidex (-0.88), indicating that higher humidity and temperature levels are associated with lower water quality. A negative correlation is also evident between AQI and WQI (-

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6409

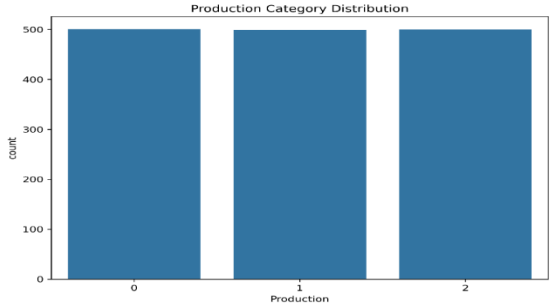
0.80), implying that as air quality worsens, water quality tends to degrade as well.



Source: (Research Results, 2025)
Heatmap Correlation

The relationship between Production and the environmental variables reveals key insights. WQI has a negative correlation with Production (-0.40), suggesting that lower water quality is associated with reduced egg production. Meanwhile, Humidex shows a moderate positive correlation with Production (0.38), indicating that higher humidity and temperature levels may be linked to an increase in egg production. The AQI has a weak positive correlation with Production (0.15), signifying a relatively minor influence of air quality on production.

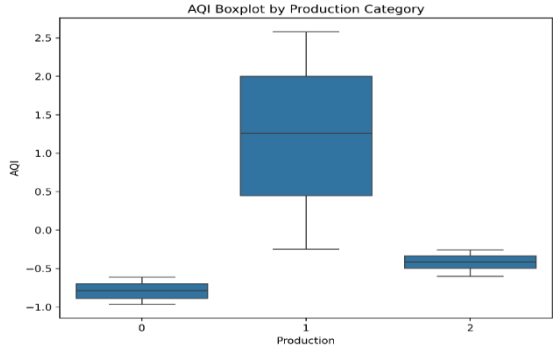
In order to better understand the relationship between environmental factors and egg production, we performed exploratory data analysis (EDA) focusing on key variables such as AQI (Air Quality Index), WQI (Water Quality Index), and Humidex. The following figures (Figures 1 to 4) illustrate the distribution and variability of these environmental factors across different egg production categories ("High," "Medium," and "Low"). The visualizations provide a clear picture of how these factors may influence production outcomes, offering valuable insights for further model development and optimization. Below is a detailed interpretation of each figure.



Source: (Research Results, 2025)
Figure 3. Production Category Distribution Train / Test Data Splitting

Figure 3 bar plot illustrates the distribution of the three production categories: "High," "Medium," and "Low." The bar chart in Figure 3 confirms that the dataset is well-balanced, with nearly equal numbers of instances in each production category: High, Medium, and Low. This balance ensures fair model training, preventing bias toward any one class and allowing for more reliable generalization across categories. The dataset is well-balanced, with each category containing approximately 500 data points. This balanced distribution is crucial for ensuring that the model does not become biased towards one class, leading to fairer and more reliable predictions. The equal distribution also facilitates better model training, as each production level is adequately represented. Each category has nearly equal representation with approximately 500 data points per category, which indicates a balanced dataset across the production categories. This balance is crucial for model training, ensuring that no category is overrepresented or underrepresented, thus preventing bias in model predictions.

Figure 4 shows the distribution of Air Quality Index (AQI) values across the three production categories. The Medium Production category exhibits a higher median AQI with a wider interquartile range, indicating more variability in air quality for this group. In contrast, High and Low production categories tend to have more stable AQI conditions, suggesting that extremely good or poor air quality is less favorable for optimal production. The "Medium" production category shows a significantly higher median AQI, with a wide interquartile range, indicating more variability in air quality for this group. The "High" and "Low" production categories exhibit lower AQI values, suggesting that extreme production levels are associated with more stable air quality conditions, either low or relatively high.

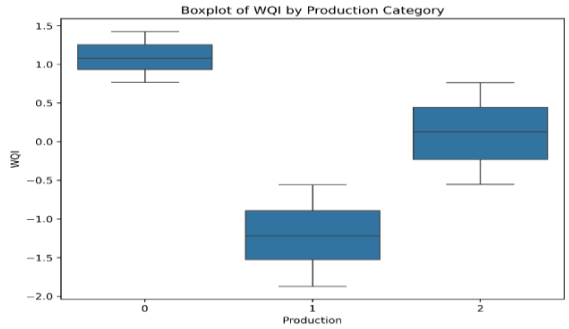


Source: (Research Results, 2025)
Figure 4. AQI Boxplot by Production Category

illustrates the Water Quality Index (WQI) for each production category. From the boxplot in

Figure 4, it is evident that the Medium category experiences the highest variability in AQI values, as reflected in the larger interquartile range. Meanwhile, the High and Low categories tend to cluster more closely around their medians, indicating more consistent air quality within those groups. Interestingly, the "High" production category corresponds to the highest WQI values, with a narrow range of variation, indicating optimal water quality conditions for maximum production. On the other hand, the "Medium" production category has the lowest WQI values, possibly suggesting that suboptimal water quality can reduce production to intermediate levels.

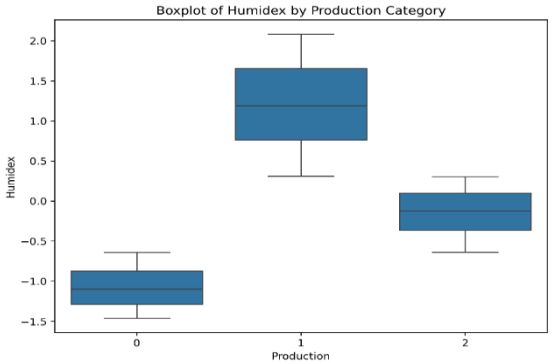
Figure 5 illustrates the Water Quality Index (WQI) for each production category. The boxplot clearly shows that the High production group benefits from optimal water quality, with a higher median WQI and less variability. In contrast, the Medium group shows the lowest WQI scores and widest spread, suggesting that suboptimal water conditions might correlate with reduced egg production. The High Production category is associated with the highest WQI values, suggesting that optimal water quality supports increased egg production. Conversely, the Medium Production category has the lowest WQI values, implying that poor water quality negatively affects egg yield. The results highlight the significant impact of water conditions on poultry productivity. Interestingly, the "High" production category corresponds to the highest WQI values, with a narrow range of variation, indicating optimal water quality conditions for maximum production. On the other hand, the "Medium" production category has the lowest WQI values, possibly suggesting that suboptimal water quality can reduce production to intermediate levels.



Source: (Research Results, 2025)

Figure 5. Boxplot of WQI by Production Category

Figure 6 shows for Humidex that the "Medium" production category tends to occur in the highest temperature and humidity conditions. The visualization highlights that extreme Humidex levels do not necessarily result in High production. Instead, High and Low production levels appear to benefit from more moderate temperature and humidity ranges, as indicated by their tighter box distributions and lower medians. This suggests that while moderate egg production thrives in warmer and more humid environments, extreme humidity levels do not necessarily lead to peak production. Both High and Low production categories exhibit lower Humidex values, reinforcing the idea that poultry production performs best under more moderate climatic conditions. Conversely, both the "High" and "Low" production categories are associated with lower Humidex values, which suggests that poultry production at extreme levels (either high or low) occurs under more moderate climatic conditions.



Source: (Research Results, 2025)

Figure 6. Boxplot of Humidex by Production Category

D. Data Split

After the data preprocessing steps, the dataset was split into training and testing sets to evaluate the model's performance effectively [28]. The features (AQI, WQI, Humidex, and other environmental factors) were used as input (X), while the "Production" column, which represents the egg production levels, served as the target variable (y). The dataset was divided into an 80-20 split, with 80% allocated for training the model and 20% set aside for testing. This approach allows the model's performance to be evaluated on unseen data, reducing the risk of overfitting and offering a more accurate assessment of its generalization ability.

E. Model Training

In this study, the Logistic Regression model was selected for its capability to handle multi-class classification problems effectively and its interpretability [29]. Logistic Regression works by modeling the probability of each class as a linear function of the input features and applying a logistic function (also known as a sigmoid function) to

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6409

produce probabilities [30]. For a binary classification, the logistic regression equation is:

$$P(y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \quad (1)$$

Where ($P(Y = 1|X)$) represents the probability that the output class (Y) belongs to category 1 given input(X), (β_0) is the intercept (bias), ($\beta_1, \beta_2, \ldots, \beta_n$) are the coefficients for each independent feature, ($x_1, x_2, \ldots, x_n$) are the feature values and (e) is the base of the natural logarithm.

For multi-class classification, Logistic Regression extends the probability calculation using the softmax function, which converts logits into probabilities:

$$P(Y = k|X) = \frac{e^{(\beta_k X)}}{\sum_{j=1}^{K} e^{(\beta_j X)}} \quad (2)$$

Where (K) is the total number of output classes, (β_k) represents the coefficient vector for class (k) and The denominator ensures that all class probabilities sum to 1. To prevent overfitting, L1 Regularization (Lasso Regression) is applied, which introduces a penalty term in the loss function:

$$L(\beta) = -\sum_{i=1}^{n}[y_i \log P_i + (1 - y_i)\log(1 - P_i)] + \lambda\sum_{j=1}^{p}|\beta_j| L(\beta) = -\sum_{i=1}^{n}[y_i \log P_i + (1 - y_i)\log(1 - P_i)] + \lambda\sum_{j=1}^{p}|\beta_j| \quad (3)$$

Where ($L(\beta)$) is the log-likelihood loss function, (y_i) represents the actual label for the (i)-th observation, (P_i) is the predicted probability from the logistic function, (λ) is the regularization parameter controlling the strength of the penalty and (β_j) are the feature coefficients.

To improve the model's performance and reduce overfitting, L1 Regularization (also called Lasso) was applied. L1 regularization introduces a penalty to the model based on the absolute values of the coefficients, which encourages sparsity by shrinking some coefficients to exactly zero. The equation for Logistic Regression with L1 regularization is as follows:

$$L(\beta) = -\sum_{i=1}^{m}\left[y^{(i)} \log\left(h_\beta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_\beta\left(x^{(i)}\right)\right)\right] + \lambda\sum_{j=1}^{n}|\beta_j| \quad (4)$$

Where ($L(\beta)$) is the loss function (log-loss), ($h_\beta(x^{(i)})$) represents the logistic function, and (λ) is the regularization parameter that controls the strength of the penalty. The second term, ($\lambda\sum_{j=1}^{n}|\beta_j|$), applies the L1 penalty, which helps in performing feature selection by setting less important feature coefficients to zero. Through GridSearchCV, the hyperparameters of the model, including the regularization strength (λ) (represented by C in the model) and the penalty type, were tuned to find the best configuration. The final model was trained with these optimized parameters and evaluated on the test set, yielding accuracy and other performance metrics.

F. Evaluation

During the evaluation phase, the performance of the Logistic Regression model was measured using key classification metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. These metrics offered valuable insights into the model's effectiveness in predicting the various production categories (High, Medium, Low) based on environmental features. The application of L1 regularization significantly improved the model's performance by preventing overfitting, leading to a better balance between bias and variance. This improvement was particularly visible in metrics such as the F1-score, where the model exhibited higher precision and recall across the various production categories. Additionally, cross-validation was employed to ensure the robustness and generalizability of the model. Cross-validation divides the dataset into several folds, commonly 5 or 10, where the model is trained on all folds except one, which is used as the validation set. This process is repeated for each fold, and the average performance across all folds is calculated to provide a more dependable evaluation of the model's accuracy. The cross-validation accuracy is computed as:

$$\text{Cross-validation accuracy} = \frac{1}{k}\sum_{i=1}^{k}\text{Accuracy}_i \quad (5)$$

where (k) is the number of folds, and (Accuracy_i) is the accuracy for the (i^{th})fold. This technique ensures that the model is not overfitting to any particular subset of the data and helps evaluate the model's ability to generalize to unseen data. To gain a deeper understanding of the model's learning behavior, a Learning Curve was plotted. The learning curve provides a visual representation of how the model's performance changes as the training set size increases. It shows two key plots: The training score reflects the model's accuracy on the training dataset, whereas the cross-validation score demonstrates the model's ability to generalize to new, unseen data. The learning curve equation can be expressed as:

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i4.6409

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

$$\text{Learning Curve} = \left(\frac{1}{n}\sum_{i=1}^{n}\text{Train Error}_i, \frac{1}{m}\sum_{j=1}^{m}\text{Validation Error}_j\right) \quad (6)$$

where (n) is the number of training examples and (m) is the number of validation examples. This equation tracks both the training error and the validation error as the model is trained on progressively larger subsets of the dataset. In our analysis, the learning curve revealed that with a smaller training set, the model exhibited high variance, resulting in lower validation accuracy and signs of underfitting. However, as the training set size increased, the *xcross-validation score** improved, indicating that the model was learning better and generalizing more effectively to unseen data. The use of L1 regularization further improved this behavior by reducing overfitting, as seen in the reduced gap between the training and validation curves.

RESULTS AND DISCUSSION

The results of this study demonstrate the effectiveness of applying Logistic Regression and L1 regularization in predicting egg production levels based on environmental factors. The analysis compares the performance of both models, highlighting the impact of regularization on improving accuracy and other key metrics.

The performance of the Logistic Regression model and its improved version with L1 regularization were evaluated using several key metrics, including accuracy, precision, recall, and F1-score. In this dataset, Class 0 represents low egg production, Class 1 represents medium production, and Class 2 represents high production. The Logistic Regression model, without regularization, achieved an accuracy of 0.91. For Class 0 (low production), the model displayed perfect precision (1.00) but had a lower recall (0.75), resulting in an F1-score of 0.86. In Class 1 (medium production), the model performed exceptionally well, achieving both precision and recall near 1.00, yielding an F1-score of 0.99. However, in Class 2 (high production), the model had a recall of 0.98 and precision of 0.76, leading to an F1-score of 0.86, indicating room for improvement in predicting high production levels.

After applying L1 regularization, the model's accuracy improved to 0.97. For Class 0 (low production), the precision dropped slightly to 0.93, but the recall increased to 1.00, resulting in a much higher F1-score of 0.96. For Class 1 (medium production), the model maintained its strong performance with precision and recall values of 1.00 and 0.99, respectively, and an F1-score of 1.00. In Class 2 (high production), the model's precision improved significantly to 0.99, while recall slightly decreased to 0.91, resulting in an F1-score of 0.94.

Before implementing L1 regularization, a baseline Logistic Regression model was trained and evaluated. The initial model achieved an accuracy of 90.67%, highlighting its capability to predict egg production levels based on environmental factors. However, analysis of the confusion matrix revealed that misclassification errors were most prominent in the Medium production class, where a significant number of instances were incorrectly classified as Low or High. This suggests that overlapping environmental conditions across production categories made it challenging for the model to accurately distinguish between them.

After implementing L1 regularization, the model's accuracy increased to 97%, demonstrating a substantial improvement in classification performance. The regularization technique effectively reduced overfitting by setting less relevant feature coefficients to zero, allowing the model to focus on the most influential environmental variables. This enhancement resulted in a more balanced precision and recall across all production categories, reducing the number of misclassified instances. The comparison between the initial Logistic Regression model and the regularized model is summarized in the following table 2:

Table 2. The comparison between the initial Logistic Regression model and the regularized model

Metric	Before Regularization (Logistic Regression)	After Regularization (L1 Logistic Regression)
Accuracy	90.67%	97%
Precision (Low Production)	1.00	0.93
Recall (Low Production)	0.75	1.00
F1-Score (Low Production)	0.86	0.96
Precision (Medium Production)	0.98	1.00
Recall (Medium Production)	1.00	0.99
F1-Score (Medium Production)	0.99	1.00
Precision (High Production)	0.76	0.99
Recall (High Production)	0.98	0.91
F1-Score (High Production)	0.86	0.94

Source: (Research Results, 2025)

To examine the predictive performance in more detail, the Logistic Regression model without regularization initially achieved an accuracy of 90.67%, with noticeable misclassification especially in the 'High' and 'Medium' production classes. Particularly, only 75% of the 'High' class examples and 98% of the 'Medium' class examples were

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6409

correctly predicted, while the rest were misclassified.

After applying L1 regularization, the accuracy of the model significantly improved to 97%. The classification report shows that 100% of the 'High' production class examples were correctly predicted, along with 99% of the 'Low' production class examples and 91% of the 'Medium' production class examples. This considerable increase in accuracy highlights the ability of the improved model to properly generalize and identify egg production categories across a wide range of environmental conditions.

The numerical improvements shown in Table X provide strong evidence of the benefits of applying L1 regularization in Logistic Regression. However, to further validate this improvement, we analyze the model's learning behavior through the learning curve in Figure Y. The learning curve reinforces the numerical findings, demonstrating that the initial model (without regularization) suffered from high variance, as indicated by the large gap between the training and validation accuracy. This aligns with the low recall for Class 0 (low production) in Table X, confirming that the model had difficulty generalizing for this category.

With the L1 regularized model, the training and validation curves converge more smoothly, reflecting improved generalization ability, which directly corresponds to the increased recall for Class 0 and improved overall accuracy in Table X. This indicates that L1 regularization effectively prevented overfitting, allowing the model to perform consistently across all production categories. Additionally, the confusion matrix in Figure Z further supports these findings by highlighting the reduction in misclassified instances, particularly for medium and high production categories. This aligns with the precision-recall improvements seen in Table X, demonstrating that L1 regularization enhanced the model's confidence in predicting high-production cases without sacrificing accuracy in other categories.

The table clearly illustrates the improvements achieved after implementing L1 regularization. Notably, the recall for Low Production increased from 0.75 to 1.00, ensuring fewer misclassifications in this category. Similarly, the precision for High Production improved significantly from 0.76 to 0.99, indicating that the model is more confident in its classifications. These results validate the effectiveness of L1 regularization in enhancing model accuracy while improving its ability to generalize to new, unseen data.

This demonstrates that L1 regularization helped the model achieve better balance across all classes, particularly in improving the precision and recall for low and high production categories. The results of both models are summarized in the table 3 below:

Table 3. Evaluation Matrix

Metric	Logistic Regression	L1 Regularization
Accuracy	0.91	0.97
Precision (Class 0)	1.00	0.93
Recall (Class 0)	0.75	1.00
F1-Score (Class 0)	0.86	0.96
Precision (Class 1)	0.98	1.00
Recall (Class 1)	1.00	0.99
F1-Score (Class 1)	0.99	1.00
Precision (Class 2)	0.76	0.99
Recall (Class 2)	0.98	0.91
F1-Score (Class 2)	0.86	0.94

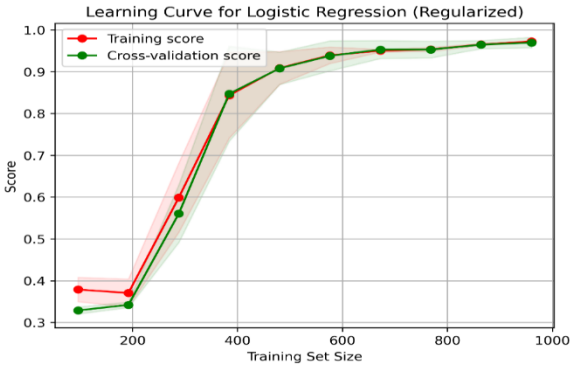
Source: (Research Results, 2025)

In summary, L1 regularization proved effective in reducing overfitting and improving the model's generalization, leading to more accurate predictions across all production categories. The improvement in accuracy from 0.91 to 0.97, along with the enhanced F1-scores for each class, highlights the benefits of regularization in optimizing the model's performance.

The learning curve shown in Figure 7 provides an in-depth evaluation of the Logistic Regression model with L1 regularization, showcasing how the model's performance improves as more training data is introduced. As shown in Figure 7, the model initially struggled with underfitting due to limited training samples, indicated by a large gap between training and validation performance. However, with more training data, the curves converge and stabilize near 1.0, confirming the model's improved generalization and robustness. This convergence demonstrates the effectiveness of regularization in preventing overfitting and achieving consistent performance. Initially, the model exhibits high variance when trained on a small dataset, leading to low validation accuracy. As more training data is introduced, both training and cross-validation scores increase significantly. At around 500 samples, they converge at approximately 0.90, indicating effective learning. The final stage, with 800–1000 samples, demonstrates strong generalization capabilities, proving that L1 regularization effectively prevents overfitting.

The curve illustrates the relationship between the training score (red line) and the cross-validation score (green line) across different training set sizes. Initially, with fewer training samples (100-200), both the training and cross-

validation scores are relatively low, indicating that the model is struggling to fit the data effectively. As the amount of training data increases, particularly between 200 and 500 samples, there is a significant improvement in both scores. At around 500 samples, the training and cross-validation scores converge, reaching approximately 0.90, which suggests that the model is learning effectively from the data and is generalizing well.



Source: (Research Results, 2025)

Figure 7. Learning Curve

In the later stages, with 800 to 1,000 samples, both scores stabilize close to 1.0, indicating that the model performs well across different datasets and is not overfitting. The convergence of the scores shows that the model has learned to generalize appropriately, without overfitting the training data, which is a direct benefit of using L1 regularization. The small gap between the training and cross-validation scores, along with their high values, confirms that the model maintains strong performance and balance between fitting the training data and generalizing to new data. This learning curve analysis highlights the robustness and efficiency of the Logistic Regression model with L1 regularization in this context.

The discussion from this study demonstrates the effectiveness of applying Logistic Regression with L1 regularization to predict egg production based on environmental factors. The initial Logistic Regression model achieved an accuracy of 90.67%, but with the application of stronger regularization (L1), the model accuracy increased significantly to 97%. This suggests that regularization plays an important role in improving model performance by overcoming overfitting and improving model generalization. In addition, the evaluation metrics, including precision, recall, and F1-score, showed balanced performance across all production classes, with the strongest performance recorded in the Low and Medium production categories.

Compared to other predictive modeling studies using Logistic Regression, this research shows a marked improvement in accuracy. For instance, De Col et al., in their study on Predicting Wheat Head Blast Epidemics, achieved an accuracy ranging between 0.8 and 0.85. Additionally, the research conducted by Jia Q1 et al. on Heart Disease Prediction using Logistic Regression with feature selection reported an accuracy of 92.39%. These results, while notable, fall short of the 97% accuracy achieved in this study. The use of L1 regularization in this research was a key factor in this success, allowing the model to select the most relevant features and reducing the impact of noise in the dataset, thus enhancing the overall predictive accuracy.

The findings of this study demonstrate that Logistic Regression with L1 regularization is highly effective for predicting egg production, outperforming similar Logistic Regression models applied in other domains. By employing regularization techniques, the model in this study achieved superior accuracy, highlighting the importance of regularization in improving predictive performance and offering insights into how Logistic Regression can be optimized for specific applications.

CONCLUSION

This study directly addresses the core issue outlined in the background, which highlights the difficulty of accurately predicting egg production due to complex environmental influences and the tendency of models to overfit when handling multiple correlated variables. By implementing L1 regularization within a Logistic Regression framework, the research effectively mitigates overfitting and enhances model generalization. The proposed model focuses on key environmental features—Air Quality Index (AQI), Water Quality Index (WQI), and Humidex—allowing for more reliable and interpretable predictions. This solution responds to the research problem by offering a method that not only improves accuracy but also ensures stability and scalability for real-world poultry production applications. This study successfully addresses the challenge of predicting egg production based on environmental factors by implementing L1 regularization in Logistic Regression, significantly improving model performance. The results demonstrate that applying L1 regularization enhances classification accuracy from 90.67% to 97%, effectively reducing overfitting while selecting the most relevant environmental variables such as Air Quality Index (AQI), Water Quality Index (WQI), and Humidex.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6409

This method provides a more generalizable model, ensuring balanced precision, recall, and F1-scores across different production levels (High, Medium, and Low).

Compared to previous studies, such as De Col et al. (Predicting Wheat Head Blast Epidemics) and Jia Q1 et al. (Heart Disease Prediction using Logistic Regression with Feature Selection), which achieved accuracies of 80–85% and 92.39%, respectively, our approach outperforms traditional Logistic Regression models by leveraging L1 regularization for automatic feature selection. This confirms that reducing noise from irrelevant features significantly enhances prediction reliability, making it applicable to real-world poultry farming scenarios. The findings of this study directly address the issue highlighted in the background: traditional prediction models often struggle to accurately forecast egg production due to complex environmental dependencies. By integrating L1 regularization, the proposed approach successfully mitigates this limitation, providing a scalable and interpretable solution. Future studies can extend this research by incorporating additional environmental and operational factors, such as feed quality, lighting conditions, and temperature control systems, to refine prediction accuracy further. Additionally, ensemble learning models (e.g., Random Forest, XGBoost) and deep learning architectures could be explored to compare their effectiveness against regularized Logistic Regression. Implementing real-time IoT-based monitoring systems in poultry farms could also enhance prediction capabilities, enabling dynamic adjustments to farming conditions and improving overall productivity and sustainability.

REFERENCE

[1] R. B. Bist *et al.*, "Sustainable poultry farming practices: A critical review of current strategies and future prospects," *Poult. Sci.*, p. 104295, 2024.

[2] W. E. Sawyer, A. O. Aigberua, M. U. Nwodo, and M. Akram, "Overview of Air Pollutants and Their One Health Effects," Springer, 2024.

[3] C. M. Ncho, J. I. Berdos, V. Gupta, A. Rahman, K. T. Mekonnen, and A. Bakhsh, "Abiotic stressors in poultry production: A comprehensive review," *J. Anim. Physiol. Anim. Nutr.*, vol. 109, no. 1, pp. 30–50, 2025.

[4] O. O. Apalowo, D. A. Ekunseitan, and Y. O. Fasina, "Impact of Heat Stress on Broiler Chicken Production," *Poultry*, vol. 3, no. 2, pp. 107–128, 2024.

[5] I. Attri, L. K. Awasthi, and T. P. Sharma, "Machine learning in agriculture: a review of crop management applications," *Multimed. Tools Appl.*, vol. 83, no. 5, pp. 12875–12915, 2024.

[6] E. Küçüktopçu, B. Cemek, and H. Simsek, "Modeling Environmental Conditions in Poultry Production: Computational Fluid Dynamics Approach," *Animals*, vol. 14, no. 3, p. 501, 2024.

[7] W. Y. Leong, Y. Z. Leong, and W. San Leong, "Eco-efficient poultry farms: Leveraging IoT for sustainable production," in *2024 9th International Conference on Applying New Technology in Green Buildings (ATiGB)*, IEEE, 2024, pp. 432–437.

[8] X. Yang *et al.*, "Computer vision-based cybernetics systems for promoting modern poultry farming: a critical review," *Comput. Electron. Agric.*, vol. 225, p. 109339, 2024.

[9] M. Aruna *et al.*, "Enhancing safety in surface mine blasting operations with IoT based ground vibration monitoring and prediction system integrated with machine learning," *Sci. Rep.*, vol. 15, no. 1, p. 3999, 2025.

[10] A. H. Alsaeedi *et al.*, "Fractal feature selection model for enhancing high-dimensional biological problems," *BMC Bioinformatics*, vol. 25, no. 1, p. 12, 2024.

[11] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," *Knowl. Inf. Syst.*, vol. 66, no. 3, pp. 1575–1637, 2024.

[12] R. Dey, N. Das, S. Mondal, B. Sadhukhan, and A. Dey, "Geo-AQI: A Real-Time Air Quality Monitoring and Forecasting System using ARIMA Model," in *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, 2024, pp. 1102–1108.

[13] A. J. Siciliano, C. Zhao, T. Liu, and Z. Wang, "EGG: Accuracy Estimation of Individual Multimeric Protein Models Using Deep Energy-Based Models and Graph Neural Networks," *Int. J. Mol. Sci.*, vol. 25, no. 11, p. 6250, 2024.

[14] N. Alamsyah, B. Budiman, T. P. Yoga, and R. Y. R. Alamsyah, "XGBOOST HYPERPARAMETER OPTIMIZATION USING RANDOMIZEDSEARCHCV FOR ACCURATE FOREST FIRE DROUGHT CONDITION PREDICTION," *J. Pilar Nusa Mandiri*, vol. 20, no. 2, pp. 103–110, 2024.

[15] M. J. Karim, M. O. F. Goni, M. Nahiduzzaman, M. Ahsan, J. Haider, and M. Kowalski, "Enhancing agriculture through real-time grape leaf

disease classification via an edge device with a lightweight CNN architecture and Grad-CAM," *Sci. Rep.*, vol. 14, no. 1, p. 16022, 2024.

[16] A. Soltani and C. L. Lee, "The non-linear dynamics of South Australian regional housing markets: A machine learning approach," *Appl. Geogr.*, vol. 166, p. 103248, 2024.

[17] N. Alamsyah, T. P. Yoga, B. Budiman, and others, "IMPROVING TRAFFIC DENSITY PREDICTION USING LSTM WITH PARAMETRIC ReLU (PReLU) ACTIVATION," *JITK J. Ilmu Pengetah. Dan Teknol. Komput.*, vol. 9, no. 2, pp. 154–160, 2024.

[18] J. Duan, J. Xiong, Y. Li, and W. Ding, "Deep learning based multimodal biomedical data fusion: An overview and comparative review," *Inf. Fusion*, p. 102536, 2024.

[19] L. N. Habibi, T. Matsui, and T. S. Tanaka, "Critical evaluation of the effects of a cross-validation strategy and machine learning optimization on the prediction accuracy and transferability of a soybean yield prediction model using UAV-based remote sensing," *J. Agric. Food Res.*, vol. 16, p. 101096, 2024.

[20] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "Regularization in machine learning models for MVT Pb-Zn prospectivity mapping: applying lasso and elastic-net algorithms," *Earth Sci. Inform.*, vol. 17, no. 5, pp. 4859–4873, 2024.

[21] X. Cheng, "A Comprehensive Study of Feature Selection Techniques in Machine Learning Models," 2024.

[22] L. A. van Veen, H. van den Brand, A. C. van den Oever, B. Kemp, and A. Youssef, "An adaptive expert-in-the-loop algorithm for flock-specific anomaly detection in laying hen production," *Comput. Electron. Agric.*, vol. 229, p. 109755, 2025.

[23] Y. Ge, J. Ma, and G. Sun, "A structural pruning method for lithium-ion batteries remaining useful life prediction model with multi-head attention mechanism," *J. Energy Storage*, vol. 86, p. 111396, 2024.

[24] N. Alamsyah, A. P. Kurniati, and others, "Event Detection Optimization Through Stacking Ensemble and BERT Fine-tuning For Dynamic Pricing of Airline Tickets," *IEEE Access*, 2024.

[25] N. Alamsyah, A. P. Kurniati, and others, "A Novel Airfare Dataset To Predict Travel Agent Profits Based On Dynamic Pricing," in *2023 11th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2023, pp. 575–581.

[26] S. Zhao *et al.*, "Comparative analysis of chloroplast genome of Meconopsis (Papaveraceae) provides insights into their genomic evolution and adaptation to high elevation," *Int. J. Mol. Sci.*, vol. 25, no. 4, p. 2193, 2024.

[27] N. Alamsyah, Budiman, V. R. Danestiara, I. Akbar, and E. Setiana, "Optimizing Computational Efficiency in Feature Selection for Machine Learning Models: A Study Crime Detection Based on Criminal Data," in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, 2023, pp. 1–6. doi: 10.1109/ICIC60109.2023.10382057.

[28] A. G. Putrada, I. D. Oktaviani, M. N. Fauzan, and N. Alamsyah, "CNN Pruning for Edge Computing-Based Corn Disease Detection with a Novel NG-Mean Accuracy Loss Optimization," *Telematika*, vol. 17, no. 2, pp. 68–83, 2024.

[29] J. Shang, Z. Xiao, T. Tao, J. Wang, and Z. Wu, "A heuristic method for discovering multi-class classification rules from multi-source data in cloud–edge system," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 36, no. 2, p. 101962, 2024.

[30] N. Alamsyah, B. Budiman, T. P. Yoga, and R. Y. R. Alamsyah, "COMPARISON LINEAR REGRESSION AND RANDOM FOREST MODELS FOR PREDICTION OF UNDERGROUND DROUGHT LEVELS IN FOREST FIRES," *J. Techno Nusa Mandiri*, vol. 21, no. 2, pp. 81–86, 2024.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri