Original Article

Check for updates

Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 2, No. 2, 2024

Forecasting Upwelling Phenomena in Lake Laut Tawar: A Semi-Supervised Learning Approach

Muhammad Zia Ulhaq [1], Muhammad Farid [1], Zahra Ifma Aziza [1], Teuku Muhammad Faiz Nuzullah [1] Fakhrus Syakir [2] and Novi Reandy Sasmita [1,*]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; m_zia@mhs.usk.ac.id (M.Z.U.); m.farid21@mhs.usk.ac.id (M.F.); zahra_if@mhs.usk.ac.id (Z.I.A); tm_faiz@mhs.usk.ac.id (T.M.F.N); novireandys@usk.ac.id (N.R.S.)

[2] Department of Electrical Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; fakrus@mhs.usk.ac.id (F.S.)

* Correspondence: email@email.com

Article History

Received 29 July 2024
Revised 2 November 2024
Accepted 10 November 2024
Available Online 19 November 2024

Keywords:
Upwelling
Historical climate data
K-Means clustering
Support Vector Machine

Abstract

The current climate change is causing the upwelling phenomenon to occur frequently in lakes and reservoirs. As a result of this phenomenon, thousands of fish die, causing floating net cage fish farmers to suffer losses. From existing studies, temperature sensors are used to determine the current condition of a body of water experiencing upwelling or not. Therefore, this study applies clustering to historical climate data from 2017-2023 using a semi-supervised learning approach that produces two labels: "potential for upwelling" and "no potential for upwelling." In the clustering process, the data is divided into two clusters using K-Means Clustering, and Support Vector Machine (SVM) is chosen to classify them. The performance of the proposed algorithm is expressed with accuracy, precision, recall, and F1-score values of 0.99, 0.995, 0.970, and 0.985, respectively. The analysis results show that this model has excellent performance in identifying upwelling potential. By using this method, information about upwelling potential can be obtained more quickly and accurately, allowing fish farmers to take appropriate preventive measures. This study also shows that the combination of K-Means Clustering and Support Vector Machine (SVM) can be effectively used to analyze historical climate data and generate useful predictions.

1. Introduction

Climate change occurring worldwide has had significant impacts on various aspects of life, including the fisheries sector [1, 2]. The floating net cage fish farmers in Lake Laut Tawar, Aceh, Indonesia, also experience this phenomenon. Mass fish deaths have become one of the impacts experienced due to climate change. In 2017, 1.5 tons of fish in Lake Laut Tawar died simultaneously, resulting in losses of up to IDR 900,000,000 [3]. Similar incidents have also occurred in Lake Maninjau [4], Lake Toba [5], and recently in Saguling Reservoir [6]. For instance, 180 tons of fish died in Lake Toba in 2018 [5], and 15 tons of fish perished in Saguling Reservoir in 2024 due to extreme weather [6]. This mass mortality is caused by upwelling or overturning in a body of water. This phenomenon occurs when heavy rain over a body of water cools the surface, as cooler water is denser than warm water. Consequently, the cooler surface layer sinks to the bottom, causing water mass mixing [7]. The problem is that the water from the bottom is often low in oxygen and carries toxic substances for fish, such as ammonia (NH_3) and hydrogen sulfide (H_2S). This condition leads to mass fish deaths [8].

Mass mortality vividly illustrates how climate change can significantly destabilize fish farming. The threat to local food security arises due to the reduced availability of fish as the primary protein source. Additionally, this impact can trigger unemployment in the fisheries sector, damage the aquatic environment with the accumulation of fish carcasses, and create uncertainty for the future of fish farmers. Besides economic losses, upwelling also disrupts aquatic ecosystems, as the rise of toxic substances reduces biodiversity and increases ecosystem instability [9]. Therefore, knowing the classification of upwelling potential becomes an appropriate solution as an initial step for mitigation and adaptation to protect the sustainability of the fisheries sector and the welfare of fish farmers and local communities.

A previous study by Hidayat [8] developed an upwelling detection and early warning system using Raspberry Pi as the control system and two sensors to measure water temperature. This system can provide information through an Android application and website when upwelling occurs. Data transmission uses the HTTP protocol with the POST function. Test results show sensor accuracy levels between 98.00% and 99.00%. This system is applied in the lake to minimize losses due to upwelling. However, this warning system detects when the upwelling event is already happening (real-time), so farmers cannot prepare to address this phenomenon in advance.

Research by Maharani [9] states that upwelling in Maninjau Lake is influenced by climate variability and global climate change, such as rainfall, temperature fluctuations, and ENSO patterns, which trigger the mixing of nutrients and toxic substances from sediments to the surface. These changes increase the vulnerability of aquaculture systems, as fish confined in floating net cages are more exposed to hypoxic conditions and toxic environments during upwelling. While mitigation measures are in place, they tend to be reactive. Climate-based resilient strategies are needed to prevent the adverse impacts of more frequent and severe upwelling.

To address the increasing challenges posed by upwelling events in Lake Laut Tawar, this study utilizes a machine learning approach, specifically a Support Vector Machine (SVM), for upwelling potential classification. The study uses historical data to classify the potential occurrence of upwelling. The classification results are helpful for farmers in making more accurate decisions to anticipate upwelling and determine the best time for fish farming. This predictive model provides early warning so farmers can take proactive measures to reduce risks. By enhancing existing reactive systems, this approach offers significant advances in aquaculture resilience strategies amid the impacts of climate change. The hope is that the results of this study can contribute to improving the fisheries sector's resilience to climate change impacts in Lake Laut Tawar, Aceh.

2. Materials and Methods

2.1. Data Descriptions

The study uses secondary data from the National Aeronautics and Space Administration (NASA) regarding daily climate data in Lake Laut Tawar, accessed through the NASA POWER Data Access Viewer. Data acquisition was conducted by inputting the specific geographic coordinates of Lake Laut Tawar (latitude and longitude) to download relevant daily climate data for the area. This data is part of the Prediction of Worldwide Energy Resources (POWER) project, which is based on NASA's climate data assimilation models, including the Modern Era Retrospective-Analysis for Research and Applications (MERRA-2) and the GMAO Forward Processing – Instrument Teams (FP-IT) GEOS 5.12.4.

MERRA-2 provides real-time climate data from 1981 up to several months prior, while GEOS 5.12.4 supplies the latest data up to several days before the current date, allowing near real-time climate monitoring. This climate data has a spatial resolution of ½° latitude and ⅝° longitude and is presented in the WGS84 global grid format, making it suitable for applications that require broad spatial coverage with adequate precision. The time period used is 7 years, from January 1, 2017, to December 31, 2023, with a total of 2,556 observations. Using a semi-supervised learning approach, the clustering results aim to group the data into two labels: "potential for upwelling" and "no potential for upwelling." This classification considers six daily climate factors in Lake Laut Tawar: All Sky Insolation Clearness Index (Y1), Temperature (Y2), Earth Skin (Y3), Wind Speed (Y4), Precipitation (Y5), and Surface Pressure (Y6).

The semi-supervised learning technique has gained significant attention in various fields, including climate studies and environmental analysis. Semi-supervised learning improves the accuracy of clustering and classification by combining a small amount of labeled data with a larger set of unlabeled data [10]. This approach is beneficial when dealing with extensive datasets such as climate information, where manual labeling of all data points could be more practical [11].

The six climate factors were chosen for their direct impact on water stratification, density, and circulation, all essential to upwelling. Insolation affects surface heating, which can destabilize water layers [12]. Temperature and Earth's Skin Temperature help maintain this stability, while sudden temperature drops may trigger upwelling as cooler surface water sinks. Wind Speed further influences mixing as strong winds move surface layers, increasing upwelling potential. Precipitation and Surface Pressure also affect water stability by altering density and adding cooler water, which can drive the upwelling process [13]. These factors collectively enhance predictive accuracy, helping fish farmers take timely preventive action. By utilizing this methodology and considering these six daily climate factors, information about upwelling potential can be obtained more quickly and accurately, allowing fish farmers to take appropriate preventive measures.

2.2. Data Preprocessing

2.2.1. Linear Interpolation

Linear interpolation is a statistical technique used to estimate missing values in a dataset by leveraging the linear relationship between known data points surrounding the missing value. This approach involves drawing a straight line between two adjacent data points that are available before and after the missing value, then assigning a new value at the midpoint based on this line.

As mentioned in an earlier study, linear interpolation is effective in straightforwardly and intuitively filling missing values, especially for addressing small data gaps [14]. Other studies also indicate that this method can produce accurate predictions of missing values, sometimes outperforming more complex models, demonstrating its superiority in various situations [15].

In the context of this study, linear interpolation is applied to the variable Y4, which has one missing data point. This method fills the missing value by estimating the appropriate data point based on the linear trend of the surrounding data.

2.2.2. KNN Imputation

KNN imputation is an effective method for handling missing values in datasets by utilizing information from nearest neighbors in feature space. This approach identifies K instances most similar to the data with missing values to fill in the missing information [16]. The method is renowned for its effectiveness in dealing with missing data issues [17].

Fadhlil et al. explain that KNN imputation predicts missing values by considering a linear combination of singular vectors. This technique is widely used across various fields, such as metabolomics and well-being index estimation based on life logs [18]. The use of KNN imputation extends to environmental studies for filling in missing values in environmental data as well [19]. In this study, the KNN imputation method was chosen because variable Y1 contains consecutively missing values, which makes KNN a more suitable approach than linear interpolation. For KNN imputation, we used K=2, K = 2 selected through cross-validation to ensure an optimal balance between accuracy and computational efficiency. This configuration allowed us to effectively estimate the missing values for Y1, enhancing the dataset's completeness and reliability for subsequent analysis.

2.3. Semi-Supervised

Semi-supervised learning (SSL) is a machine learning paradigm that combines a small amount of labeled data with a large amount of unlabeled data during the training process. This approach is particularly useful in scenarios where acquiring labeled data is expensive or time-consuming while unlabeled data is abundant [20]. SSL lies between supervised and unsupervised learning, allowing models to leverage the strengths of both methodologies to improve learning performance [21]. Semi-supervised learning has gained recognition for enhancing predictive models by effectively utilizing labeled and unlabeled data. In upwelling prediction, integrating clustering and classification techniques can significantly improve model accuracy. Clustering helps organize data into meaningful groups, providing a structured foundation that enhances the subsequent classification process.

2.4. K-Means clustering

K-means clustering is a method that operates on the concept of descriptive modeling. The K-Means algorithm is used to determine the location of an object within a specific cluster based on the nearest average distance [22]. It is a type of clustering algorithm that uses partitioning methods, where objects are grouped into clusters based on centroids, representing each cluster's central points. The centroids are iteratively updated to minimize the variance within clusters, ensuring that objects in the same cluster are as similar as possible.

This method is often compared to algorithms like k-Medoids, which emphasize using actual objects as the central points for grouping rather than calculated centroids [23]. K-Means is particularly useful for large datasets due to its simplicity and computational efficiency, making it a common choice in clustering applications. The evaluation of clustering quality can be further validated using metrics such as the silhouette

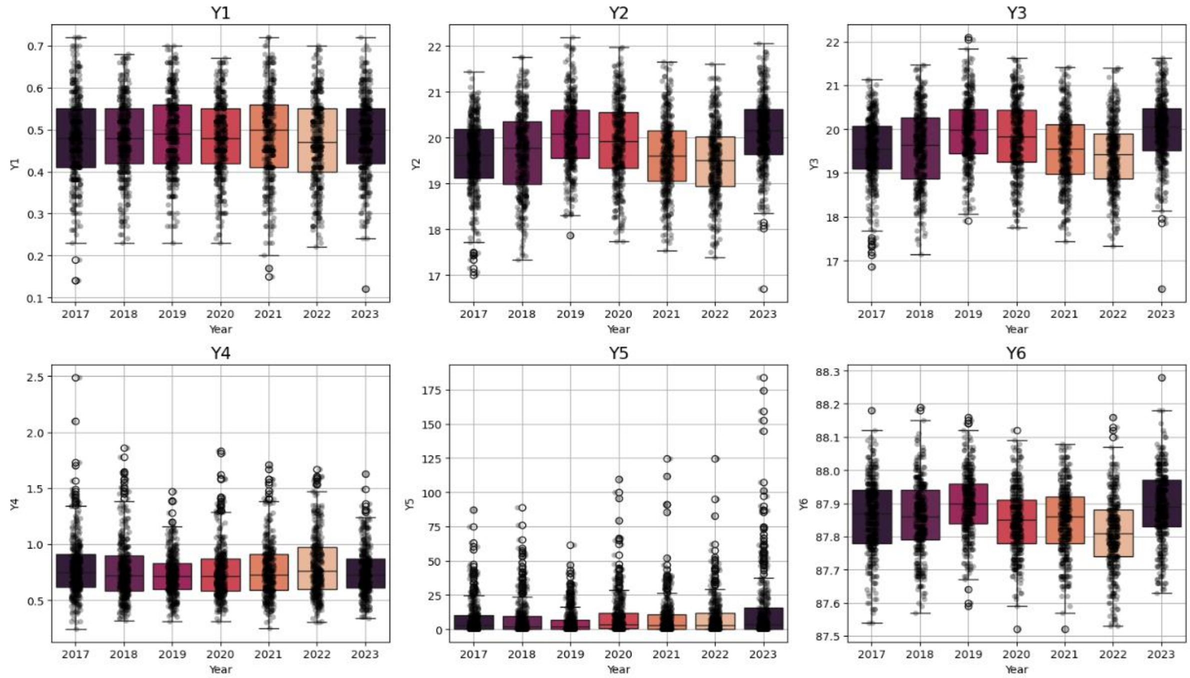

Figure 1. Distributed climate indicator data.

score or confusion matrix results based on a labeled test set, providing insights into the separation and cohesion of clusters.

2.5. Splitting data

After the labels in the dataset were determined through the semi-supervised clustering process, the data was divided into two main groups: training and testing data. This division was carried out to ensure that the machine learning model could be effectively trained and tested to evaluate its performance. A total of 80% of the 2,556 observations (2,044 data points) were allocated as training data, while the remaining 20% (512 data points) were used as testing data.

The training data was used to train the model so it could recognize patterns in the data, such as the relationship between daily climate variables and the potential for upwelling. During this process, the model learned the characteristics of the labeled data to optimize its internal parameters. Meanwhile, the testing data was used to evaluate the performance of the trained model. Using testing data that the model had not seen during training, the evaluation results objectively indicated how the model would perform on new, unseen data [24].

This data split used a stratified random sampling approach to ensure that the label distribution in both the training and testing data remained balanced. Maintaining a fair representation of each class, namely “potential for upwelling” and “no potential for upwelling, " is important. This approach helps the model avoid bias that might arise from label imbalance in the data subsets [25].

2.6 Support Vector Machine (SVM)

The SVM is a powerful supervised machine learning algorithm widely used for classification tasks. SVM works by finding the optimal hyperplane separating different feature space classes [26]. SVM is distinct from perceptrons in that they can define a linear classifier by maximizing the margin in the feature space and can also function as nonlinear classifiers through kernel tricks [27]. The equations for linear, Radial Basis Function (RBF), and polynomial kernels are shown in Equations 1, 2, and 3, respectively [28]:

$$K(xi, x) = x_i^T x \tag{1}$$

$$K(xi, x) = exp\,(-\gamma |x_i^T x|^2) \tag{2}$$

$$K(xi, x) = (\gamma . x_i^T x + r)^p \tag{3}$$

where p is degree, r is coefficient and γ is gamma.

3. Results and Discussion

Figure 1 displays the distribution of climate indicator data from 2017 to 2023, highlighting key variables that may influence upwelling potential. The Y1 variable (All Sky Insolation Clearness Index) shows a consistent annual variation, which could impact solar heating of surface waters, indirectly affecting thermal stratification and

Table 1. Silhouette evaluation results for number of clusters optimization.

No.	N Cluster	Silhouette Score
1	2	0.774
2	3	0.734
3	4	0.681
4	5	0.653
5	6	0.636
6	7	0.614
7	8	0.556

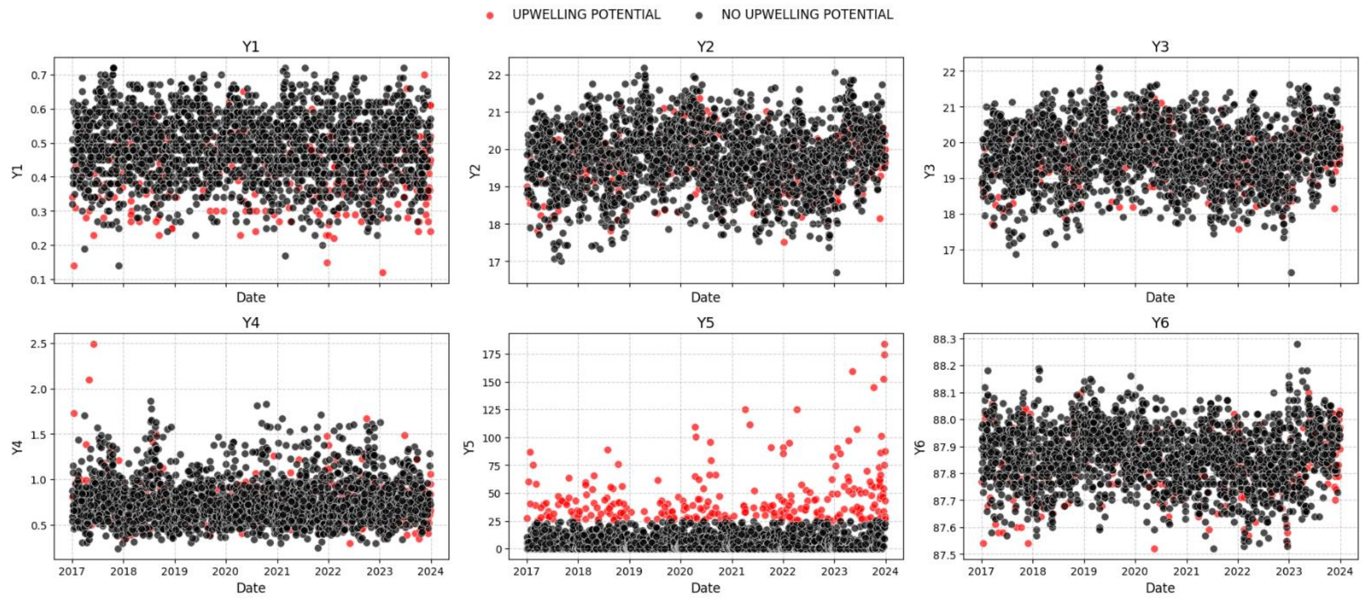

Figure 2. Distribution of clustered data.

upwelling potential. The Y2 variable (Temperature) and Y3 variable (Earth Skin Temperature) have similar trends, with relatively narrow ranges, indicating stable thermal conditions; however, any significant deviations may disrupt water layering, triggering upwelling.

The Y4 variable (Wind Speed) exhibits high variability, with extreme values up to 183.99 m/s, suggesting that wind can play a significant role in upwelling events, as stronger winds are known to promote surface water movement and mixing. Y5 (Precipitation) shows low variability but could influence surface water density and contribute to upwelling conditions when combined with other factors. Y6 (Surface Pressure) remains stable, yet minor fluctuations could impact atmospheric pressure dynamics and water column stability. These distributions suggest that wind speed, precipitation, and temperature variations are the most relevant indicators in assessing upwelling potential. Higher wind speeds and fluctuations in temperature or precipitation could lead to conditions conducive to upwelling, thus affecting aquatic ecosystem dynamics and fish survival.

The results in Table 1 show that the highest silhouette score is obtained when the number of clusters is 2, with a value of 0.77375. This score indicates the best separation of clusters compared to other cluster numbers. The silhouette score is a metric used to assess how similar data points within a cluster are compared to data points in other clusters. The silhouette score ranges from -1 to 1, where values approaching 1 indicate that data points are well-clustered within their own cluster, while values approaching -1 suggest that data points may be in the wrong cluster. A silhouette score of 0.773 for two clusters indicates that the data within each cluster are very similar to each other and distinct from data in other clusters. Therefore, the study chooses two clusters as the best solution for grouping the data. Next, the formed clusters are analyzed to determine areas that potentially exhibit upwelling and areas that do not exhibit upwelling.

Based on Figure 2, the graph displays time series data for several climate indicators, highlighting "upwelling potential" (red points) and "no upwelling potential" (black points). Variable Y1 (All Sky Insolation Clearness Index) ranges from 0.12 to 0.72, with upwelling scattered across this range, suggesting it can occur at various levels of Y1, averaging 0.41 during upwelling. This indicates that lower

Table 2. Comparison of support vector machine model evaluation based on different kernels.

Kernel	Classification Report			
	Precision	Recall	F1-Score	Accuracy
Linear	0.985	0.970	0.980	0.99
RBF	0.995	0.970	0.985	0.99
Polynomial	0.995	0.970	0.985	0.99

Table 3. Model performance.

Model	TN	FP	FN	TP
Linear	51	3	1	457
RBF	51	3	0	458
Polynomial	51	3	0	458

solar radiation may promote conditions favorable for upwelling by cooling the surface water, thereby destabilizing the water column. Y2 (Temperature) spans 16.7°C to 22.2°C, suggesting that upwelling can occur across a wide temperature range. The average temperature during upwelling is 19.68°C, indicating that upwelling tends to happen in moderately cool water, which aligns with the process where surface waters cool and sink, triggering the overturning of deeper, nutrient-rich waters. Y3 (Earth Skin Temperature) shows a similar range, reinforcing the close relationship between the surface temperature and the occurrence of upwelling. Y4 (Wind Speed) ranges from 0.25 m/s to 183.99 m/s, with higher wind speeds often associated with upwelling. This is consistent with strong winds enhancing the vertical mixing of water layers, promoting upwelling by moving cooler, deeper waters to the surface. Y5 (Precipitation) ranges from 0 to 183.99 mm, with higher values frequently linked to upwelling, especially post-2022. This suggests that heavy rainfall increases the surface water's density, causing instability in the water column that can trigger upwelling. Y6 (Surface Pressure) ranges from 87.52 to 88.28 kPa, with upwelling occurring at various pressure levels. While surface pressure does not exhibit as strong a pattern as other factors, fluctuations in pressure can still influence water column stability, potentially contributing to upwelling under specific conditions. Upwelling is associated with lower clearness index values, slightly lower temperatures, higher wind speeds, and increased precipitation.

Table 2 presents the performance evaluation of the SVM model across different kernels for classification tasks, revealing that the RBF and Polynomial kernels slightly outperform the Linear kernel, although all three kernels achieve a very high overall accuracy of 0.99. The choice of kernel in SVM is essential because each kernel maps data into different feature spaces, with each kernel being more suited for certain types of data. The high precision of 0.995 for the RBF and Polynomial kernels suggests that most positive predictions made by the SVM model are highly relevant to the target class. Kernel RBF and Polynomial are particularly effective for non-linear data, as they map data to higher dimensions, enabling the model to capture complex patterns that a linear kernel might miss. This indicates that the model effectively reduces the number of false positives, which is crucial in applications where the accuracy of positive predictions is paramount. Consistent recall scores of 0.970 for all kernels indicate that the model identifies the most true positive instances from the target class. While the Linear kernel is effective for linearly separable data, both the RBF and Polynomial kernels are better suited for non-linearly separable data, as they allow the model to capture more intricate relationships between features. While there is no significant difference in recall between the RBF, Polynomial, and Linear kernels, the higher precision of the RBF and Polynomial kernels suggests that they may be more suitable for classification tasks that require minimizing false positives. The F1-Score, which is the harmonic mean of precision and recall, also demonstrates the superiority of the RBF and Polynomial kernels with a score of 0.985 compared to the Linear kernel's score of 0.980. This higher F1-Score for RBF and Polynomial kernels indicates that they strike a better balance between precision and recall, ensuring that the model performs optimally in terms of both minimizing false positives and capturing the most true positives.

Table 3 shows the evaluation results of three classification models based on the confusion matrix. Each model, Linear, RBF, and Polynomial, is evaluated based on its ability to classify samples in the dataset into positive and negative categories. The Linear model achieves nearly perfect results with 457 true positives (TP) and only one false negative (FN), indicating its strong capability to correctly identify samples that belong to the positive category. However, it also experiences three false positives (FP), meaning some samples that should be negative are incorrectly classified as positive. On the other hand, both the RBF and Polynomial models show perfect results in terms of false negatives (FN), with a value of 0, indicating accuracy in identifying all positive samples. They also achieve 458 true positives (TP), demonstrating high accuracy in positive classification. They also have three false positives (FP), similar to the Linear

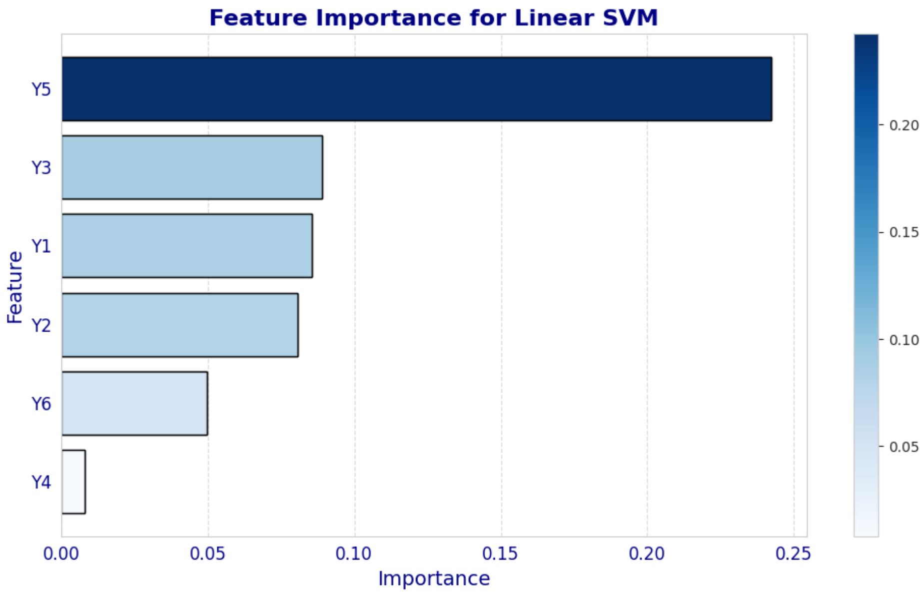

Figure 3. Feature Importance.

model, indicating slight inaccuracies in classifying negative samples. Overall, all three models exhibit very good performance in classifying positive samples, but further evaluation is needed to consider the trade-off between false positives and false negatives depending on specific application requirements.

Figure 3 shows the importance of each feature in the linear SVM model used to classify the potential for upwelling in Lake Laut Tawar. According to the chart, precipitation (Y5) emerges as the most significant predictor, showing the highest importance compared to other variables. This aligns with the understanding that heavy rainfall increases surface water density, leading to instability and vertical mixing. Next, earth skin temperature (Y3) and the clearness index (Y1) are ranked as the second and third most important features. These variables influence thermal stratification and the stability of the water column. Meanwhile, temperature (Y2) and surface pressure (Y6) show lower contributions but remain relevant, as fluctuations in atmospheric pressure can impact local dynamics and environmental conditions. Interestingly, wind speed (Y4) has the lowest contribution in the model, which contrasts with the general expectation of its critical role in driving upwelling processes. This could reflect the unique local characteristics of Lake Laut Tawar or limitations in the dataset used in this study. The importance of these features highlights the need to prioritize monitoring efforts on precipitation (Y5) and earth skin temperature (Y3) while acknowledging the complementary roles of other variables. These insights can strengthen mitigation and management strategies for upwelling events in Lake Laut Tawar.

The findings of this study demonstrated the high effectiveness of a semi-supervised learning approach, combining K-Means clustering and SVM, in predicting upwelling potential in Lake Laut Tawar. The SVM model with a Polynomial kernel achieved impressive performance metrics: precision of 0.995, recall of 0.970, F1-Score of 0.985, and accuracy of 0.99. Compared to previous studies, such as Hidayat's real-time detection system [8], which relied on temperature sensors, this study offers a predictive advantage by analyzing historical climate data to anticipate upwelling events. Additionally, this study builds upon Maharani's [9] work on climatic influences by quantifying the significance of precipitation, earth skin temperature, and clearness index as key indicators of upwelling potential. However, the study also identified limitations, such as the dependence on historical data quality and potential variations in model performance across different geographic contexts. These findings highlight the need for future studies to incorporate additional variables, such as dissolved oxygen levels and lake-specific characteristics, to improve model reliability.

From a practical perspective, this study provides valuable tools for fish farmers and policymakers. The predictive model enables fish farmers to take preventive measures, such as adjusting operations or relocating fish cages, to mitigate losses caused by upwelling. Policymakers can leverage these findings to develop sustainable aquaculture strategies by allocating resources to early warning systems and monitoring infrastructure. Despite these advantages, the study's occasional false positives underline the need for refining the model using higher-resolution datasets and real-time monitoring. Future studies should also explore the model's application in other lakes or reservoirs to assess its adaptability and generalizability, ensuring that it remains a robust tool for improving aquaculture resilience in diverse environmental and socio-economic conditions.

4. Conclusions

This study developed a semi-supervised predictive model for upwelling potential in Lake Laut Tawar, combining K-Means clustering and SVM to enhance aquaculture resilience. Using historical climate data (2017–2023), it identified precipitation (Y5), Earth Skin Temperature (Y3), and Clearness Index (Y1) as the most influential factors in predicting upwelling risks, with lower contributions from wind speed (Y4) and surface pressure (Y6). The SVM model with a Polynomial kernel demonstrated excellent performance (accuracy: 99%), confirming its utility as an early warning tool for fish farmers. By integrating climate data and machine learning, this study offers a novel, proactive approach to managing upwelling risks, contributing to sustainable fisheries and resilience against climate change, while also highlighting areas for future research, such as incorporating higher-resolution datasets and exploring additional environmental factors.

Author Contributions: Conceptualization, M.Z.U., M.F., and F.S.; methodology, M.Z.U., N.R.S., T.M.F.N., and Z.I.A.; software, M.F., F.S., and N.R.S.; validation, N.R.S., T.M.F.N., and Z.I.A.; formal analysis, M.Z.U., M.F., and F.S.; investigation, M.Z.U., N.R.S., and T.M.F.N.; resources, N.R.S. and T.M.F.N.; data curation, N.R.S., T.M.F.N., and Z.I.A.; writing—original draft preparation, M.Z.U., M.F., and F.S.; writing—review and editing, N.R.S., T.M.F.N., and Z.I.A.; visualization, F.S.; supervision, N.R.S.; project administration, N.R.S. All authors have read and agreed to the published version of the manuscript

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available upon request to the authors.

Acknowledgements: NASA POWER (Prediction of Worldwide Energy Resources) provided free access to data and made valuable contributions to this study.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Lam, V. W. Y., Allison, E. H., Bell, J. D., Blythe, J., Cheung, W. W. L., Frölicher, T. L., Gasalla, M. A., and Sumaila, U. R. (2020). Climate Change, Tropical Fisheries and Prospects for Sustainable Development, *Nature Reviews Earth & Environment*, Vol. 1, No. 9, 440–454. doi:10.1038/s43017-020-0071-9.
2. Hewindati, Y. T., and Novita Sari, D. (2023). Effect of Life History on Alopias pelagicus Overexploitation Vulnerability: A Literature Review, *Leuser Journal of Environmental Studies*, Vol. 1, No. 2, 39–46. doi:10.60084/ljes.v1i2.54.
3. Agronet.id. (2017). Ribuan Ikan Mati di Danau Laut Tawar Aceh, from https://www.agronet.co.id/detail/indeks/berita/818-Ribuan-Ikan-Mati-diDanau-Laut-Tawar-Aceh, accessed 10-8-2024.
4. Vinolia. (2023). Lagi, Puluhan Ton Ikan di Danau Maninjau Mati, from https://www.mongabay.co.id/2023/05/30/lagi-puluhan-ton-ikan-di-danau-maninjau-mati/, accessed 8-8-2024.
5. Kompas.com. (2018). Ikan Mati di Danau Toba Mencapai 180 Ton, Ini Penyebabnya..., from https://nasional.kompas.com/read/2018/08/29/11290411/ikan-mati-di-danau-toba-mencapai-180-ton-ini-penyebabnya, accessed 9-8-2024.
6. Kompas.com. (2024). Cuaca Ekstrem, Belasan Ton Ikan Mati di Keramba Waduk Saguling, from https://bandung.kompas.com/read/2024/11/05/055942178/cuaca-ekstrem-belasan-ton-ikan-mati-di-keramba-waduk-saguling, accessed 5-11-2024.
7. Chang, P., Xu, G., Kurian, J., Small, R. J., Danabasoglu, G., Yeager, S., Castruccio, F., Zhang, Q., Rosenbloom, N., and Chapman, P. (2023). Uncertain Future of Sustainable Fisheries Environment in Eastern Boundary Upwelling Zones under Climate Change, *Communications Earth & Environment*, Vol. 4, No. 1, 19. doi:10.1038/s43247-023-00681-0.
8. Hidayat, A. I. (2021). Sistem Pendeteksi Dan Peringatan Dini Upwelling, *Jurnal Ilmiah Ilmu Komputer*, Vol. 7, No. 1, 55–61. doi:10.35329/jiik.v7i1.187.
9. Susanti, S. (2020). Kadar Senyawa Toksik Dalam Darah Ikan Nila (Oreochromis niloticus) Saat Musim Up-Welling Di Danau Maninjau, Sumatra Barat, *Jurnal Sain Veteriner*, Vol. 38, No. 1, 25. doi:10.22146/jsv.29322.
10. Dang, V.-T., Vu, V.-V., Do, H.-Q., and Oanh Le, T. K. (2021). Graph Based Clustering with Constraints and Active Learning, *Journal of Computer Science and Cybernetics*, Vol. 37, No. 1, 71–89. doi:10.15625/1813-9663/37/1/15773.
11. C A Padmanabha Reddy, Y., Viswanath, P., and Eswara Reddy, B. (2018). Semi-Supervised Learning: A Brief Review, *International Journal of Engineering & Technology*, Vol. 7, No. 1.8, 81. doi:10.14419/ijet.v7i1.8.9977.
12. Sun, F., Yu, F., Si, G., Wang, J., Xu, A., Pan, J., and Tang, Y. (2022). Characteristics and Influencing Factors of Frontal Upwelling in the Yellow Sea in Summer, *Acta Oceanologica Sinica*, Vol. 41, No. 7, 84–96. doi:10.1007/s13131-021-1967-z.
13. Kim, D., Choi, J.-G., Park, J., Kwon, J.-I., Kim, M.-H., and Jo, Y.-H. (2023). Upwelling Processes Driven by Contributions from Wind and Current in the Southwest East Sea (Japan Sea), *Frontiers in Marine Science*, Vol. 10. doi:10.3389/fmars.2023.1165366.
14. Prediction of Dam Horizontal Displacement Based on CNN-LSTM and Attention Mechanism. (2021). *Academic Journal of Architecture and Geotechnical Engineering*, Vol. 3, No. 1. doi:10.25236/AJAGE.2021.030103.
15. Cho, B., Dayrit, T., Gao, Y., Wang, Z., Hong, T., Sim, A., and Wu, K. (2020). Effective Missing Value Imputation Methods for Building Monitoring Data, *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2866–2875. doi:10.1109/BigData50022.2020.9378230.
16. Lee, J. Y., and Styczynski, M. P. (2018). NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data, *Metabolomics*, Vol. 14, No. 12, 153. doi:10.1007/s11306-018-1451-8.
17. Jadhav, A., Pramod, D., and Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset, *Applied Artificial Intelligence*, Vol. 33, No. 10, 913–933.
18. Fadlil, A., Herman, and Praseptian M, D. (2022). K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction, *Jurnal RESTI (Rekayasa Sistem Dan*

Teknologi Informasi), Vol. 6, No. 4, 570–576. doi:10.29207/resti.v6i4.4173.

19. Saeipourdizaj, P., Sarbakhsh, P., and Gholampour, A. (2021). Application of Imputation Methods for Missing Values of PM 10 and O 3 Data: Interpolation, Moving Average and K-Nearest Neighbor Methods, *Environmental Health Engineering and Management*, Vol. 8, No. 3, 215–226. doi:10.34172/EHEM.2021.25.

20. Chung, H., and Lee, J. (2022). Iterative Semi-Supervised Learning Using Softmax Probability, *Computers, Materials & Continua*, Vol. 72, No. 3, 5607–5628. doi:10.32604/cmc.2022.028154.

21. Taherdoost, H. (2024). Beyond Supervised: The Rise of Self-Supervised Learning in Autonomous Systems, *Information*, Vol. 15, No. 8, 491. doi:10.3390/info15080491.

22. Noviandy, T. R., Hardi, I., Zahriah, Z., Sofyan, R., Sasmita, N. R., Hilal, I. S., and Idroes, G. M. (2024). Environmental and Economic Clustering of Indonesian Provinces: Insights from K-Means Analysis, *Leuser Journal of Environmental Studies*, Vol. 2, No. 1, 41–51. doi:10.60084/ljes.v2i1.181.

23. Lund, B., and Ma, J. (2021). A Review of Cluster Analysis Techniques and Their Uses in Library and Information Science Research: K-Means and K-Medoids Clustering, *Performance Measurement and Metrics*, Vol. 22, No. 3, 161–173. doi:10.1108/PMM-05-2021-0026.

24. Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset, *Scientific Reports*, Vol. 12, No. 1, 7166. doi:10.1038/s41598-022-11012-2.

25. Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., and Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery, *Malacca Pharmaceutics*, Vol. 2, No. 2, 68–78. doi:10.60084/mp.v2i2.217.

26. Noviandy, T. R., Idroes, G. M., Tallei, T. E., Handayani, D., and Idroes, R. (2024). QSAR Modeling for Predicting Beta-Secretase 1 Inhibitory Activity in Alzheimer's Disease with Support Vector Regression, *Malacca Pharmaceutics*, Vol. 2, No. 2, 79–85. doi:10.60084/mp.v2i2.226.

27. Rad, J. A., Parand, K., and Chakraverty, S. (Eds.). (2023). *Learning with Fractional Orthogonal Kernel Classifiers in Support Vector Machines*, Springer Nature Singapore, Singapore. doi:10.1007/978-981-19-6553-1.

28. Rochim, A. F., Widyaningrum, K., and Eridani, D. (2021). Performance Comparison of Support Vector Machine Kernel Functions in Classifying COVID-19 Sentiment, *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 224–228. doi:10.1109/ISRITI54043.2021.9702845.