JIKO (Jurnal Informatika dan Komputer) Accredited KEMDIKTISAINTEK, No.0173/C3/DT.05.00/2025
Vol. 8, No. 3, Desember 2025, pp. 285-293 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v8i3.10964 e-ISSN: 2656-1948

CLUSTERING ANALYSIS OF SIGNIFICANT WAVE HEIGHT DYNAMICS USING K-MEANS ALGORITHM IN THE SEMARANG–DEMAK COASTAL WATERS

Wahyu Sri Mulyani[1], Aji Supriyanto[2*]

[1,2] Master of Information Technology, Faculty of Information Technology and Industry, Universitas Stikubank, Semarang, 50241, Indonesia
Email: [1]wahyusri0034@mhs.unisbank.ac.id, [*2]ajisup@edu.unisbank.ac.id

(Received: 30 October 2025, Revised: 09 December 2025, Accepted: 17 December 2025)

Abstract

Global climate change has led to an increase in the frequency and intensity of extreme events at sea, including in the Semarang-Demak coastal area. This region is highly vulnerable to the dynamics of Significant Wave Height (SWH), sea level rise, and coastal land subsidence. As a result, in addition to disrupting maritime navigation, frequent occurrences of tidal flooding (rob) have caused significant disturbances to economic activities and settlements in the coastal area. This study aims to develop a clustering model for SWH in the Semarang-Demak waters using the K-Means algorithm. The data used includes oceanographic and meteorological parameters from the Tanjung Emas Semarang Maritime Meteorological Station (BMKG) for the period 2019-2024. The clustering results show that K-Means successfully formed three clusters of sea waves representing calm, moderate, and high waves. Model evaluation using the Silhouette Score with a value of 0.725 and the Davies-Bouldin Index (DBI) of 0.425 indicates good performance, with K=3 as the optimal cluster. Temporal analysis reveals a clear seasonal pattern, where high energy conditions dominate during the west season (December-February), while calm conditions are prevalent during the east season (June-August). These findings provide a foundation for early warning systems and disaster risk management in this region, with further clustering tests using other algorithms and the need for improved data quality.

Keywords: *K-Means, Clustering, Significant Wave Height (SWH), Semarang–Demak, Tidal Flooding (Rob)*

**Corresponding Author: Aji Supriyanto*

1. INTRODUCTION

Global climate change has increased the frequency and intensity of extreme events at sea, such as rising wave heights, tidal flooding (rob), and coastal erosion, which impact ecosystems and socio-economic activities in coastal communities[1][2]. According to the 2024 report from the Regional Disaster Management Agency (BPBD) of Central Java, economic losses from the effects of flooding and rob in Semarang-Demak reached 1.6 trillion IDR. Furthermore, the Tide Eye Indonesia research team reports that the economic losses from flooding and rob along the northern coast of Central Java amount to 2.5 trillion IDR per year. Climate change and regional oceanographic activities have contributed to the increase in Significant Wave Height (SWH) and Sea Level Rise (SLR) along the Semarang-Demak coastal area[1].

High waves pose a threat to maritime safety, one of which is caused by dominant wind patterns that contribute to the occurrence of high waves[3]. This phenomenon has become increasingly frequent in tropical regions such as Indonesia, which has complex oceanographic dynamics[4]. This condition demands the presence of an accurate and adaptive marine data monitoring and analysis system to support disaster mitigation and sustainable coastal zone management.

The frequency of high waves in the northern Java waters is relatively low compared to other islands, but the waves can reach up to 2.6 meters[3]. In addition to waves, which vary in height, wave period, and wave direction[1], one of the influences on waves is wind, particularly wind direction and speed[1][3][5][6]. Other influential variables include sea surface pressure, ocean currents (direction and speed) [1][7].

The Semarang-Demak coastal area is one of the most vulnerable regions to the impacts of climate change and sea level rise (SLR). The rise in sea level and land subsidence are the primary causes of tidal flooding (rob) [4][8][9]. Additionally, the increase in Significant Wave Height (SWH) and coastal erosion

causes damage to infrastructure as well as disruptions to port and fishing activities[1][4][8]. The Semarang-Demak coastal area is a strategic region with dense economic activity, including ports, fisheries, industry, and settlements[9].

The oceanographic conditions of the Semarang-Demak waters are highly complex, influenced by the interaction of monsoon winds, ocean currents, and the relatively flat topography of the seabed[4]. The issue of SLR is caused by climate change and increasing wave heights, which result in coastal land loss and tidal flooding in the region[1]. The wave patterns in this area show high spatial and temporal variability[9]. The global SLR phenomenon has increased the magnitude of HEX near the Javanese coast by 0.7m–0.8m during 2010–2017, which corresponds to seasonal sea level rise[10].

On average, the Java Sea experiences an SLR of 3.9 mm per year[1][10], including in the Semarang and Demak waters. The rise in the Semarang waters is 5.52 mm per year, which is validated with tidal data from Semarang[1]. The primary factors contributing to this are SLR caused by several main elements such as lunar and solar gravity, as well as SWH, especially due to severe weather. The impacts of climate change in Central Java include tidal flooding (rob) caused by tidal wave surges[11]. The phenomenon of sea level rise and land subsidence has occurred along the Central Java coast, particularly in Semarang, Pekalongan, and Demak[12][13][14]. This is the main cause of hydrometeorological disasters, particularly tidal flooding in these areas[13][15]. Consequently, further actions are needed from stakeholders to mitigate the worsening hydrometeorological disasters along the Semarang-Demak coast[1].

These events require an analytical approach capable of grouping wave characteristics based on dominant physical parameters. A comprehensive understanding of these wave patterns is crucial to support early warning systems, coastal spatial planning, and disaster mitigation[1][16]. Therefore, an effective data analysis approach such as clustering is needed to group sea wave data based on their characteristics, enabling the identification of key patterns that contribute to the formation of high waves[17]. A data-driven approach capable of accurately analyzing the patterns and dynamics of sea waves can support an early warning system in this area[1][17].

The use of artificial intelligence (AI) methods is highly relevant for providing solutions in oceanography, one of which is for sea wave detection. The K-Means algorithm is effective for clustering sea wave data based on shared characteristics[17][18]. In this study, K-Means was used to cluster sea level variability, which is related to wave behavior[17], including wave disasters[18]. K-Means has been shown to accurately identify differences in wave Stokes profiles, demonstrating its ability to differentiate complex wave characteristics[19].

Performance evaluation of K-Means for annual coastal bed evolution shows that the algorithm is robust in clustering coastal evolution data based on similar characteristics[20]. K-Means also identifies natural patterns in time series data without requiring initial labels, making it suitable for exploring previously uncharted wave characteristics[18]. K-Means effectively distinguishes complex wave profiles[19]. Therefore, K-Means is used for clustering coastal evolution data, which can be adapted to identify risk zones based on wave characteristics[20].

The K-Means algorithm has shown better performance compared to DBSCAN, based on silhouette index parameters for oceanographic data[21]. K-Means excels in large datasets and provides easily interpretable outputs for operational coastal analysis[22]. K-Means provides stable and representative cluster structures for wave dynamics, while DBSCAN is more sensitive to parameters and tends to label rare conditions as noise[23]. This is particularly relevant for wave datasets in tropical regions with strong seasonal variability, such as Semarang–Demak.

Based on the existing issues and supporting references, the objective of this study is to develop a clustering model for sea waves in the Semarang–Demak waters using the K-Means algorithm to support risk zoning systems and maritime disaster mitigation. The contribution of this research is the development of a clustering model for tropical oceanographic data, particularly in the northern Semarang-Demak waters, which can be used as a reference for maritime navigation and coastal disaster mitigation.

2. RESEARCH METHOD

In general, the stages of this research include literature study and data collection, data preprocessing, K-Means clustering modeling, and evaluation and recommendation of results. The complete stages can be seen in Figure 1.

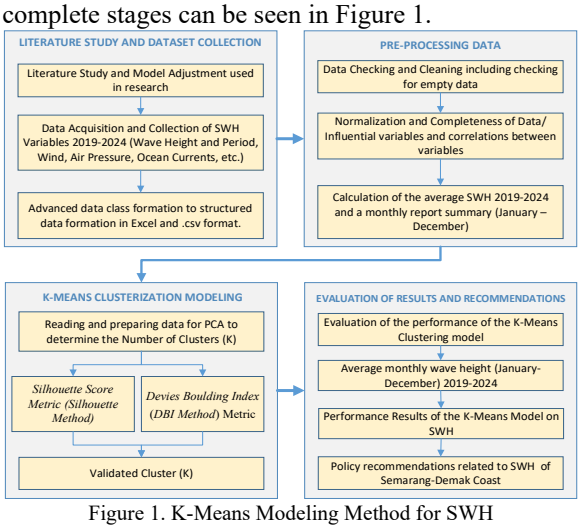

Figure 1. K-Means Modeling Method for SWH

2.1 Data Collection

This research is a development of the previous study conducted by Ganis & Supriyanto (2025)[1]. While the previous study focused on prediction, this study performs clustering using the K-Means algorithm. The data used is the same as in the previous study, specifically weather data from the Semarang City and Demak Regency regions that influence the Significant Wave Height (SWH). The data was obtained from the Maritime Meteorological Station of the Meteorology, Climatology, and Geophysics Agency (BMKG) Central Java. The data was collected from two observation points representing the water conditions in the study area.

The weather data used is daily data from 2019 to 2024 (6 years), consisting of 2,192 records. The data variables include: significant wave height (hs), wave period (t01), wave direction (dir), ocean current speed (cm/s), and sea surface pressure (hPa). All these parameters were chosen because they have a strong relationship with the formation and dynamics of sea waves. In addition to the primary variables, supporting variables were also required for conducting the wave clustering analysis, as shown in Table 1.

Table 1. SWH Dataset Structure

Variable Name	Description
Date	Data observation time is in daily datetime format.
Average Speed	Average daily wind speed (m/s), an indicator of the general condition of the surface atmosphere.
Max Speed	The maximum daily wind speed (m/s) reflects the highest wind intensity.
Direction	Wind direction (0–360°) indicates the dominant wind movement pattern.
Current Speed	Surface ocean current velocity (cm/s) plays a role in energy distribution in the waters.
Pressure	Air pressure (hPa) is used as an indicator of local atmospheric conditions.
hs	Significant wave height (m) is the target variable for prediction and classification.
t01	The average wave period (seconds) describes the interval between waves.
dir	Wave arrival direction (0–360°), representing the direction of wave energy propagation.
Average	The average aggregate value of daily parameters provides an overview of general conditions.
Max	The highest value of the combined parameter on a single observation day.
Min	The lowest value of the combined parameter over a daily period.
Year	The year extracted from the Date variable is used in annual trend analysis.
Month	Month derived from Date, used for seasonal identification.
SWH	Wave height class label (Calm – Extreme) based on hs value threshold.

The data, which has been saved in .xlsx or .csv format, is reorganized by including important attributes that represent the features or labels for each observation record. This step is crucial to ensure that each variable to be used in the detection analysis process is well-structured and consistent.

2.2. Data Preprocessing

The raw data that has been collected undergoes a preprocessing stage to ensure its completeness, consistency, and quality. This stage includes data cleaning, handling missing values, normalization, and variable transformation according to the needs of machine learning algorithms. The preprocessing steps are as follows:

1. Data Cleaning**.** This step is performed after the data has been collected, and the data saved in .xlsx and .csv formats is reorganized by including important attributes that represent features or labels for each observation record. This step is essential to ensure that each variable to be used in the classification and clustering analysis process is well-structured and consistent.
2. Normalization and Data Completeness Check**.** This process checks for missing values and ensures that the values are within reasonable statistical limits. Data that is deemed valid is data that does not contain missing values and has a distribution of values within statistically reasonable boundaries. The results of the check indicate that out of a total of 2,192 daily observation records from 2019 to 2024, there was one record with missing values in some parameters, such as Current Speed, vektor_u, and vektor_v. Therefore, the total valid (clean) data used in the analysis process consists of 2,191 records. The structure of the data after cleansing and feature selection, which forms the input predictors and SWH detection labels, can be seen in Figure 2**.**

	Tanggal	Kec Rata2	Arah Terbanyak	Kec Terbesar	Arah	hs	\
0	2019-01-01	3.863636	360.0	10.0	240.0	1.084145	
1	2019-01-02	3.318182	290.0	8.0	290.0	0.759486	
2	2019-01-03	3.227273	360.0	8.0	40.0	0.599004	
3	2019-01-04	4.454545	350.0	12.0	10.0	0.335136	
4	2019-01-05	3.636364	290.0	8.0	320.0	0.103225	
2187	2024-12-27	0.166667	210.0	6.0	20.0	0.335938	
2188	2024-12-28	0.166667	210.0	6.0	20.0	0.416992	
2189	2024-12-29	0.166667	210.0	6.0	20.0	0.434570	
2190	2024-12-30	0.166667	210.0	6.0	20.0	0.789062	
2191	2024-12-31	0.166667	210.0	6.0	20.0	0.908203	

	t01	dir	RATA2	MAKS	MIN	Kec Arus (cm/s)	\
0	5.131293	315.275523	1009.758333	1011.1	1008.2	9.825789	
1	4.895896	314.921814	1009.754167	1011.2	1007.2	10.040464	
2	5.064625	315.959632	1010.891667	1012.5	1009.1	10.255138	
3	4.910140	316.153704	1011.250000	1012.8	1008.9	8.798804	
4	4.511729	312.710474	1011.370833	1013.4	1009.1	8.556446	
2187	3.801758	321.706035	1009.820833	1011.5	1007.1	13.483292	
2188	3.708008	316.450726	1009.820833	1011.5	1007.1	13.483292	
2189	3.850586	323.915007	1009.820833	1011.5	1007.1	13.483292	
2190	4.131836	322.546425	1009.820833	1011.5	1007.1	13.483292	
2191	4.908203	319.191034	1009.820833	1011.5	1007.1	13.483292	

	vektor_u (cm/s)	vektor_v (cm/s)	Arah ke (derajat)
0	-9.399663	-1.175148	262.873843
1	-9.618452	-0.937620	264.401554
2	-9.837240	-0.700092	265.929265
3	-8.287500	-2.312500	254.409054
4	-7.912501	-2.562500	252.055213
2187	11.047485	-1.757568	99.039553
2188	11.047485	-1.757568	99.039553
2189	11.047485	-1.757568	99.039553
2190	11.047485	-1.757568	99.039553
2191	11.047485	-1.757568	99.039553

Figure 2. Data Structure of SWH Variables After Data Cleansing

These steps aim to ensure that the data used for detection analysis meets quality standards and enhances the reliability of the classification and regression models built. This preprocessing approach aligns with practices in studies of prediction, classification, or clustering of waves, emphasizing

the cleaning of missing values, normalization, and feature selection to maintain the reliability of the machine learning models [1]. During this stage, correlations between influential variables are also created in the form of heatmaps and bar charts.

3. Calculation of Average SWH. This is done by calculating the average SWH for each month (January-December) from 2019 to 2024. This average is used as a basis for determining which months correspond to calm, low, moderate, high, and extreme sea waves (hs). Based on these monthly SWH averages, an accumulative SWH clustering for the Semarang-Demak waters can be established. Additionally, it is used to identify the SWH and SLR trends for each year.

2.3. Clustering Modeling with K-Means

The K-Means method was chosen as the clustering model for the SWH in the Semarang-Demak waters based on the studies by Ganis and Supriyanto (2025) [24], Li et al. (2025) [17], and Lin et al. (2025)[18]. The K-Means clustering steps are as follows:

1. Cluster Data Preparation**.** Select relevant variables (features), and ensure that the data has been normalized to ensure the suitability of input for the K-Means model.
2. Determining the Optimal Number of Clusters (K)**.** The Elbow Method is used to determine the optimal K value by examining the trade-off between the number of clusters and WCSS (Within-Cluster Sum of Squares). The formula:

$$WCSS\_k = \Sigma\,(x\_i - c\_k)\text{^}2 \quad (1)$$

Information:
x_i : data points in cluster k
c_k : centroid of cluster k

3. K-Means Implementation**.** The K-Means algorithm is applied to divide the data into clusters based on the proximity to centroids, with cluster quality evaluation using Silhouette Score (s) and Davies-Bouldin Index (DBI).

– The Silhouette Score metric is used to calculate the ratio between the proximity of a data point to its own cluster (cohesion) and the proximity to the nearest cluster (separation). A value close to +1 indicates well-formed clusters. The formula:

$$s = \frac{b-a}{\max(a,b)} \quad (2)$$

Information:
a : average intra-cluster distance
b : average inter-cluster distance

– The DBI metric is used to measure the comparison between the distance between clusters and the dispersion within clusters. A lower DBI value indicates better-formed clusters, as it suggests a larger distance between clusters and greater compactness within each cluster. The formula:

$$DBI = \frac{1}{n}\sum_{i=n}^{n} \underset{j \neq i}{Max} \left(\frac{s_i + s_j}{d_{ij}}\right) \quad (3)$$

Information:
n : number of clusters
S_i : measure of dispersion of cluster i
d(c_i, c_j): distance between centroids(i and j)
max: taken for each cluster i relative to j

2.4. Evaluation of Results and Recommendations

The evaluation of clustering results aims to assess how well K-Means can map sea wave conditions in the Semarang-Demak region based on the selected parameters. The evaluation is conducted using two main metrics, namely Silhouette Score (S) and Davies-Bouldin Index (DBI), which provide insights into the cohesion and separation of clusters based on the average monthly wave heights (January–December) from 2019 to 2024. The activities conducted include:

1. Analysis of clustering results**.** The clustering results are analyzed to identify key patterns in the dynamics of sea waves in the study area, such as clusters with low, high, or extreme SWH. This mapping will help understand how sea conditions change over time and form seasonal patterns.
2. Recommendations for management**.** Based on the clustering results, this study provides recommendations for managing and mitigating sea wave risks in the Semarang-Demak region. For example, identifying high-risk zones can aid in coastal disaster planning and response, as well as port and fisheries infrastructure management.
3. Model improvements and future work**.** Recommendations for future research may include using other algorithms such as DBSCAN to address outliers or more complex wave density variations, or time-series-based modeling to dynamically predict SWH and SLR trends.

These evaluation steps provide a deeper understanding of the characteristics of sea waves in the Semarang-Demak area and offer data-driven solutions for improving maritime navigation and better coastal management.

3. RESULT AND DISCUSSION

This study applies the K-Means Clustering algorithm to group sea wave data in the coastal area of Semarang–Demak using daily weather datasets from the 2019–2024 period. Based on the stages and methods described in the previous chapter, the analysis and results can be explained as follows.

3. 1 Pre-processing Results

Based on the data structure resulting from the normalization process of the Significant Wave Height (SWH) as shown in Figure 1, the correlation between the involved variables can be calculated. The correlations between the variables in the dataset and the target SWH (hs) are illustrated in the heatmap shown in Figure 3.

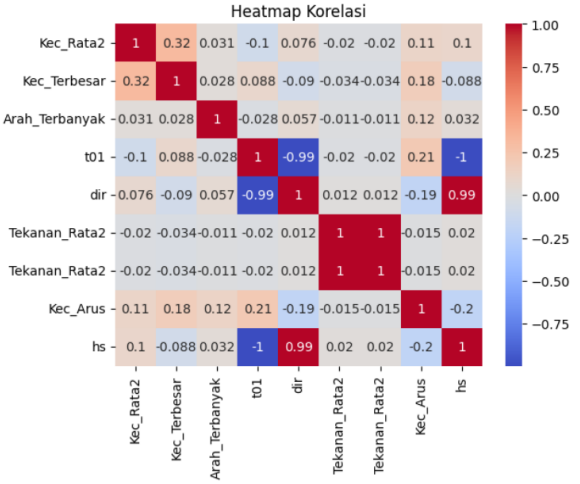

Figure 3. Correlation Heatmap of SWH Variables

Based on Figure 3, the variables that influence SWH are wave height (hs): 1.0, wave period (t01): 1.0, wind direction (dir): 0.994, and current velocity (cm/s): 0.197. The wind direction (dir) variable shows a very high correlation with SWH, indicating that wind direction strongly affects sea wave height. Similarly, the wave period (t01) has a very strong and negative correlation with SWH, meaning that the longer the wave period, the smaller the likelihood of large waves occurring.

Furthermore, since sea wave height (hs) is the main variable and the target of the clustering process, its average value needs to be calculated. The average value is computed based on monthly averages from 2019 to 2024. The results are shown in Figure 4.

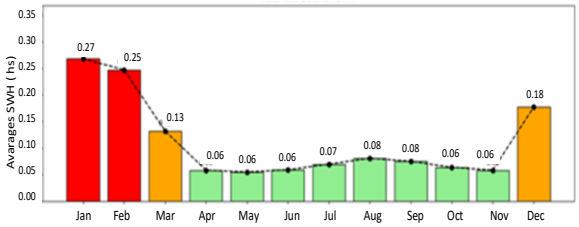

Figure 4. Average SWH from 2019 to 2024

In Figure 4, the highest average values occur in January (0.27 m) and February (0.25 m), reflecting the peak of the high wave season caused by westerly winds. Subsequently, the hs values decrease sharply and reach minimum conditions from April to November, ranging between 0.06–0.08 m, indicating a calm sea period. In December, wave heights rise again to around 0.18 m, marking the transition toward the next high wave season. This pattern is highly consistent on a seasonal basis and shows that the December–February period represents the highest wave risk season. This information is crucial as a basis for maritime navigation planning, coastal activities, and early warning systems for marine hazards.

3. 2 K-Means Clusterization

The clustering results in this study show that the K-Means algorithm successfully divided the sea wave data in the Semarang–Demak coastal area into three main groups that are consistent with seasonal oceanographic conditions. To obtain an initial overview of the data characteristics, the significant wave height (hs) values were classified into categorical classes (SWH_Class). This classification is based on the standard BMKG/WMO thresholds: Calm (0–0.5 m), Slight (0.5–1.25 m), Moderate (1.25–2.5 m), and Rough (>2.5 m). The frequency distribution resulting from the categorization of daily observation data for the 2019–2024 period is presented in Table 2.

Table 2. Distribution of SWH Classes

Class	Total	Percentage (%)
Smooth	1475	67.3
Slight	444	20.3
Moderate	236	10.8
Rough	36	1.6
Total	**2191**	**100**

The class distribution in Table 2 was obtained from daily observation data of significant wave height (hs) during the 2019–2024 period at the Tanjung Emas Maritime Meteorological Station, Semarang. The class determination process was carried out by converting hs values into SWH_Class categories based on the standard BMKG/WMO thresholds: Calm (0–0.5 m), Slight (0.5–1.25 m), Moderate (1.25–2.5 m), and Rough (>2.5 m). The distribution, which is dominated by calm–slight categories, is consistent with the characteristics of short-fetch tropical waters and remains relevant for the development of computationally efficient support vector models on daily wave data[1].

Based on the data shown in Figure 2, out of a total of 2,191 valid records, the majority fall into the Calm category with 1,475 records (67.3%), followed by Slight with 444 records (20.3%), Moderate with 236 records (10.8%), and Rough with 36 records (1.6%). These results indicate that wave conditions in the Semarang–Demak waters during the observation period were generally dominated by calm to low sea states, while high wave conditions occurred only occasionally and in very limited numbers.

It should be noted that the distribution in Table 2 represents a classification result based on the official BMKG/WMO thresholds, which differs from Table 3 that presents the clustering results using the K-Means algorithm. Classification uses fixed thresholds to define categories, whereas clustering groups data based on natural patterns of oceanographic variables.

To determine the optimal number of clusters (K) in K-Means, the Elbow Method was applied, and the visualization results are shown in Figure 5.

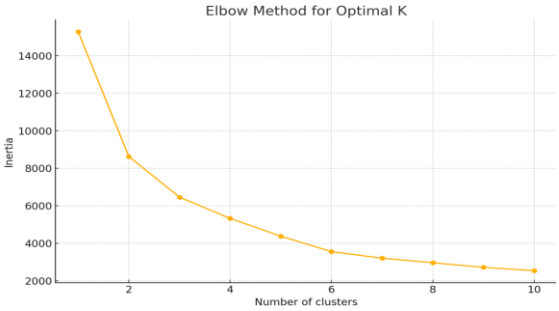

Figure 5. Visualization of the Elbow Method for SWH

Based on Figure 5, further analysis was conducted to determine the optimal number of SWH clusters. In Figure 5, the X-axis represents the number of clusters (K) ranging from 1 to 10, while the Y-axis represents the Inertia value. Inertia (Within-Cluster Sum of Squares / WCSS) indicates the total squared distance between each data point and its cluster centroid. As the value of K increases, the inertia decreases because the data is divided into more clusters, reducing the average distance between data points and their centroids. However, after a certain point, the decrease in inertia begins to slow down—this point is known as the 'elbow point'. In Figure 5, the most distinct elbow appears around K = 3 or 4. Beyond K > 4, the reduction in inertia becomes less significant, meaning that adding more clusters does not substantially improve accuracy but increases model complexity.

The choice of K = 3 as the optimal number of clusters is based on the substantial decrease in inertia from K = 2 to K = 3. After K = 3, the reduction becomes relatively slow, indicating that adding more clusters provides diminishing returns in representation efficiency.

This reasoning is further supported by the Silhouette Score method, which measures how similar a data point is to its own cluster compared to other clusters. The Silhouette Score obtained is 0.725, where K = 3 still yields a high value, indicating well-separated clusters (a value close to 1 signifies very good separation). This is also supported by the Davies-Bouldin Index (DBI) value of 0.425. A DBI value below 0.5 is generally considered good, suggesting that K = 3 provides a satisfactory clustering result for tropical oceanographic data. This finding is consistent with prior studies showing that appropriate feature selection enhances the separability of K-Means clusters[17].

Figure 6 presents the 2D PCA visualization of the SWH clustering results. The visualization shows the distribution of data grouped into three clusters, each represented by a different color. Cluster 0 (red), Cluster 1 (blue), and Cluster 2 (green) are distinctly separated. Clusters 0 and 2 (green and red) are more clearly separated along PCA1 (the first principal component), while Cluster 1 (blue) appears more dispersed along the same axis. This indicates that PCA successfully reduced the data dimensionality and effectively separated the clusters based on the most relevant principal components.

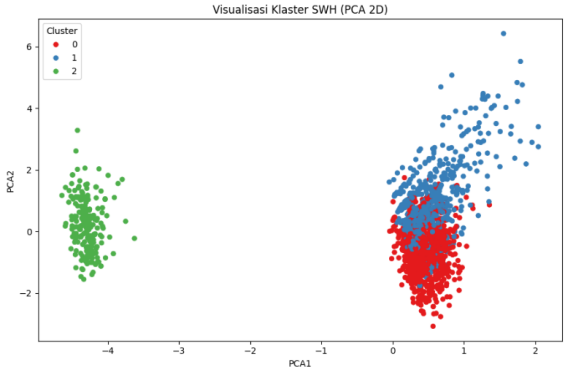

Figure 6. Visualization of SWH Clusters using PCA (3 Clusters)

The data shown in Figure 6 can be grouped into the following clusters:

a. Cluster 0 (green). This cluster exhibits high density along the PCA2 axis, indicating that the data within this cluster are more compact and concentrated within a specific area. It represents conditions with calmer seas (low SWH), as reflected by the concentrated distribution and lower variation along both principal components.
b. Cluster 1 (blue). This cluster appears more dispersed, suggesting greater variability in SWH. It represents more diverse sea conditions, ranging from moderate to high SWH. The wider spread along both PCA1 and PCA2 indicates larger variations within this cluster.
c. Cluster 2 (red). This cluster shows a broader distribution, indicating higher variability across both PCA1 and PCA2. It represents conditions of extreme sea waves (very high SWH), characterized by a wide dispersion along PCA2, which reflects high fluctuations in sea wave activity.

PCA effectively reduced the data into two principal dimensions, clearly distinguishing the clusters. This demonstrates that the first and second principal components contain significant information for separating SWH data based on clustering results. To further understand the characteristics of each cluster and to predict cluster membership for new data generated in the K-Means clustering process, statistical summaries and cluster classifications were developed.

After determining that the optimal number of clusters is K = 3, the number of data points and their respective percentages were calculated, along with descriptive statistics for each cluster. Cluster 0 contains 1,407 data points (74.48%), Cluster 1 contains 67 data points (3.55%), and Cluster 2 contains 415 data points (21.97%). Additionally, descriptive statistics were computed for each K-Means cluster, including the mean, standard deviation, minimum, median, and maximum values. The results are presented in Table 3.

Table 3. Statistical Values K = 3 for the Period 2019–2024

Cluster	Sum	Mean	Std	Min	25%	50%	75%	Max
0	1,407	0.144	0.131	0.000	0.050	0.100	0.205	1.2832
1	67	47.80	0.429	47.078	47.441	47.805	48.169	48.532
2	415	0.611	0.411	0.005	0.299	0.538	0.837	2.145

The difference in proportions represents the variation of sea conditions, ranging from calm seas to transitional and high-energy wave conditions. These results are consistent with the PCA and hs–t01 scatter visualizations, which show a clear separation between clusters, confirming that the Semarang–Demak waters during the 2019–2024 period were predominantly characterized by calm to transitional sea states, with high-wave events occurring only occasionally.

It should be noted that the cluster numbering (0, 1, 2) in Table 3 is an automatic label generated by the K-Means algorithm and does not indicate an ordered ranking of average wave height. Therefore, even though Cluster 1 has the highest hs value, its position is not placed at the top. The numbering simply serves as a group identifier. Furthermore, the clustering results in Table 3 differ from the wave classification in Table 2, which is based on the SWH thresholds (calm, low, moderate, high). Thus, clustering reveals the natural patterns of data without predefined categories, whereas classification uses official categorical standards.

Moreover, the clustering results can be interpreted within the context of risk mitigation. Cluster 0, representing calm sea conditions, can be categorized as "Safe"; Cluster 1, with transitional characteristics, as "Alert"; and Cluster 2, with relatively higher wave height and current speed, as "Hazardous." Hence, this table not only illustrates the statistical separation of data patterns but also provides practical relevance in supporting an early warning system for the Semarang–Demak coastal waters. The observed pattern of 'calm → transition → higher energy,' emerging from hs, wave period, and current variables, aligns with prior evidence that K-Means can effectively identify representative annual sea conditions for coastal prediction and risk management purposes.

In a previous study by Erutjhahjo and Supriyanto (2025)[1], SLR and SWH in the northern coast of Semarang-Demak were predicted without creating SWH clusters, thus leaving the hazardous SWH conditions for maritime navigation and coastal hydrometeorological disasters unidentified. This study determines SWH clusters, including extreme heights, to enable mitigation of maritime navigation risks and coastal hydrometeorological disasters.

3.3 Evaluation and Recommendations

Based on the results of data analysis and testing, this study produced the following assessments:

a. After collecting daily weather data from 2019–2024, the main influencing variables were identified as follows: significant wave height (hs): 100%, wave period (t01): 100%, and wind direction (dir): approximately 99%. Meanwhile, other variables such as ocean current velocity (cm/s): ~20%, and surface air pressure (hPa): <1%, showed minor influence.
b. During the data cleansing and normalization process, some missing values were detected. However, during the testing phase, several outliers were still observed, although their effect was not significant. This was particularly evident during the PCA test, which suggests that the presence of outliers might be due to human input errors during manual data entry by weather station operators.
c. Although the clustering result with K = 3 produced good performance based on the Silhouette Score and Davies–Bouldin Index (DBI), further testing with K = 4 is recommended. This would help determine whether the clustering quality improves, and whether the data distribution becomes more consistent with the linearity of wave classification based on WMO (World Meteorological Organization) standards.
d. The research would be more comprehensive if extended to include prediction and classification analyses, allowing for a clearer understanding of inter-variable relationships and model consistency.

Based on the testing and evaluation results, several recommendations are proposed in this study as follows:

a. Further clustering tests should be conducted using alternative algorithms, directly applying the main influencing variables, namely significant wave height (hs), wave period (t01), and wind direction (dir).
b. During the data cleansing and normalization process, invalid or inconsistent data should be carefully reviewed and, if necessary, verified with the original weather data operators before removal or correction. This verification process is essential to reduce potential outliers during PCA (Principal Component Analysis) and clustering tests.
c. Additional testing using different values of K, such as K = 4, is recommended to confirm whether better clustering performance can be achieved and to ensure higher accuracy in representing oceanographic patterns.
d. A more comprehensive study combining prediction and classification analyses on the same dataset and study area would provide deeper insights and more complete information for marine navigation risk mitigation and coastal management, particularly for the Semarang–Demak coastal region.

4. CONCLUSION

This study successfully developed a clustering model for sea waves in the Semarang–Demak coastal waters using the K-Means algorithm. The clustering

results show that Significant Wave Height (SWH) can be categorized into three primary clusters: calm, moderate, and high sea waves. Cluster 0 represents calm sea conditions with low SWH, Cluster 1 encompasses moderate to high waves, and Cluster 2 represents extreme wave conditions. The evaluation metrics—Silhouette Score (0.725) and DBI (0.425), indicate that clustering with K = 3 provides well-separated and representative clusters.

The findings of this research make a significant contribution to understanding the dynamics of sea waves in the Semarang–Demak region, providing valuable insights for disaster mitigation, coastal management planning, and maritime navigation safety. The application of PCA for dimensionality reduction prior to clustering also proved effective in enhancing the separability of the data.

Although the developed model demonstrates strong performance, further improvement is recommended through the application of alternative clustering algorithms and enhanced data quality verification. Future studies integrating other clustering methods, along with predictive and classification analyses, will offer a more comprehensive understanding and support more informed decision-making for coastal risk mitigation and marine disaster management in the Semarang–Demak coastal region.

5. REFERENCE

[1] G. Erutjahjo and A. Supriyanto, "Prediksi Tinggi Gelombang Laut di Perairan Semarang – Demak dengan Menggunakan Random Forest dan XGBoost," *Jurnal Informatika : Jurnal pengembangan IT*, vol. 10, no. 4, pp. 869–881, 2025, doi: 10.30591/jpit.v10i4.9315.

[2] Z. Ahmad and M. Mansurova, "Machine Learning Approach To Predict Significant Height," *Journal of Mathematics, Mechanics and Computer Science*, vol. 2, no. 110, pp. 87–96, 2021.

[3] F. Retika, D. N. Sugianto, and R. Widiaratih, "Analisis Terjadinya Gelombang Tinggi Akibat Pola Pergerakan Angin Terkait Keselamatan Pelayaran di Perairan Utara Jawa Tengah," *Indonesian Journal of Oceanography*, vol. 6, no. 4, pp. 334–343, 2024, doi: 10.14710/ijoce.v6i4.24678.

[4] P. Raharjo, F. B. Prasetio, G. N. Hawari, and N. A. Kristanto, "Dinamika Pantai Kota Semarang, Jawa Tengah," *Jurnal Geologi Kelautan*, vol. 22, no. 2, pp. 130–145, 2025, doi: 10.32693/jgk.22.2.2024.926.

[5] H. T. Mudho, I. A. Azies, J. Setiyadi, E. A. Kisnarti, and W. S. Pranowo, "Karakteristik Tinggi Gelombang Laut di Perairan Halmahera Utara dan Morotai pada Periode Waktu ENSO Tahun 2012-2021," *Jurnal Kelautan Tropis*, vol. 28, no. 1, pp. 11–24, 2025, doi: 10.14710/jkt.v28i1.25192.

[6] S. V. Haiyqal, A. Ismanto, E. Indrayanti, and R. Andrianto, "Karakteristik Tinggi Gelombang Laut pada saat Periode Normal, El Niño dan La Niña di Selat Makassar," *Jurnal Kelautan Tropis*, vol. 26, no. 1, pp. 190–202, 2023, doi: 10.14710/jkt.v26i1.17003.

[7] J. Mo, X. Wang, S. Huang, and R. Wang, "Advance in Significant Wave Height Prediction: A Comprehensive Survey," *Complex System Modeling and Simulation*, vol. 4, no. 4, pp. 402–439, 2025, doi: 10.23919/csms.2024.0019.

[8] K. K. Khairullah, A. Rifai, and E. Indrayanti, "Studi Luasan Genangan Banjir Rob Akibat Kenaikan Muka Air Laut Dan Penurunan Muka Study of the Area of Flood Inundation Due to Sea Level Rise and Land Subsidence in Sayung District , Demak," *Indonesian Journal of Oceanography (IJOCE)*, vol. 06, no. 04, pp. 316–323, 2024, doi: 10.14710/ijoce.v6i4.24645.

[9] A. Supriyanto, D. A. Diartonor, B. Hartono, and A. Jananto, "Classification Of Sea Wave Heights On The North Coast Of Central Java Using Random Forest," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 4, pp. 2263–2280, 2025, [Online]. Available: https://jutif.if.unsoed.ac.id/index.php/jurnal/article/view/5108/915

[10] W. Han *et al.*, "Sea level extremes and compounding marine heatwaves in coastal Indonesia," *Nature Communications*, vol. 13, no. 1, pp. 1–12, 2022, doi: 10.1038/s41467-022-34003-3.

[11] C. Tay *et al.*, "Sea-level rise from land subsidence in major coastal cities," *Nature Sustainability*, vol. 5, no. 12, pp. 1049–1057, 2022, doi: 10.1038/s41893-022-00947-z.

[12] G. Mulyasari, Irham, L. R. Waluyati, and A. Suryantini, "Understanding and adaptation to climate change of fishermen in the northern coastal of Central Java, Indonesia," *IOP Conference Series: Earth and Environmental Science*, vol. 724, no. 1, 2021, doi: 10.1088/1755-1315/724/1/012094.

[13] S. Susilo *et al.*, "GNSS land subsidence observations along the northern coastline of Java, Indonesia," *Scientific Data*, vol. 10, no. 1, pp. 1–8, 2023, doi: 10.1038/s41597-023-02274-0.

[14] K. Triana and A. J. Wahyudi, "Sea level rise in Indonesia: The drivers and the combined impacts from land subsidence," *ASEAN Journal on Science and Technology for Development*, vol. 37, no. 3, pp. 115–121, 2020, doi: 10.29037/AJSTD.627.

[15] A. Supriyanto, E. Zuliarso, E. T. Suharmanto, H. Amalina, and F. Damaryanti, "Drought Prediction Using Lstm Model With Standardized Precipitation Index on the North Coast of Central Java," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 6, pp. 1873–1882, 2024, doi: 10.52436/1.jutif.2024.5.6.4159.

[16] C. Murtiaji, M. Irfani, I. Fauzi, A. S. D.

Marta, C. I. Sukmana, and D. A. Wulandari, "Methods for addressing tidal floods in coastal cities: An overview," *IOP Conference Series: Earth and Environmental Science*, vol. 1224, no. 1, 2023, doi: 10.1088/1755-1315/1224/1/012019.

[17] J. Li *et al.*, "Prediction of Seawater Intrusion Run-Up Distance Based on K-Means Clustering and ANN Model," *Journal of Marine Science and Engineering*, vol. 13, no. 2, pp. 1–18, 2025, doi: 10.3390/jmse13020377.

[18] Z. Lin, N. F. S. Zulkepli, M. S. Bin Mohd Kasihmuddin, and R. Gobithaasan, "A Topological-Indicators-Based k-Means Clustering Algorithm and Its Application in Time Series Data: A Case Study on Sea Level Variability in Peninsular Malaysia," *IEEE Access*, vol. 13, no. March, pp. 46514–46533, 2025, doi: 10.1109/ACCESS.2025.3548558.

[19] T. E. Moe, T. M. D. Pereira, F. Calvo, and J. Leenaarts, "Shape-based clustering of synthetic Stokes profiles using k -means and k -Shape," *Astronomy and Astrophysics*, vol. 675, no. A130, pp. 1–12, 2023, doi: 10.1051/0004-6361/202346724.

[20] A. Papadimitriou and V. Tsoukala, "Evaluating and enhancing the performance of the K-Means clustering algorithm for annual coastal bed evolution applications," *Oceanologia*, vol. 66, no. 2, pp. 267–285, 2024, doi: 10.1016/j.oceano.2023.12.005.

[21] A. Aprianti, A. Jufriansah, P. B. Donuata, A. Khusnani, and J. Ayuba, "Comparison of K-Means Algorithm and DBSCAN on Aftershock Activity in the Flores Sea: Seismic Activity 2019-2022," *Journal of Novel Engineering Science and Technology*, vol. 2, no. 03, pp. 77–82, 2023, doi: 10.56741/jnest.v2i03.393.

[22] D. Wang, D. Conley, M. Hann, K. Collins, S. Jin, and D. Greaves, "Power output estimation of a RM3 WEC with HF radar measured complex representative sea states," *International Marine Energy Journal*, vol. 5, no. 1, pp. 1–10, 2022, doi: 10.36688/imej.5.1-10.

[23] H. Li *et al.*, "A Novel Sea State Classification Scheme of the Global CFOSAT Wind and Wave Observations," *Journal of Geophysical Research: Oceans*, vol. 129, no. 11, 2024, doi: 10.1029/2023JC020686.

[24] G. Erutjahjo, "Sistem informasi pasang surut menggunakan alat palem pasut".