Evaluating pH Levels in Coastal Water Quality Data Using Machine Learning Techniques

Brenda Eloisa Sánchez Pichardo[1 [A01637814]], Ilse Karena de Anda Garcíar[2 [A01637814]], Andres Robles Gil Candas[2 [A01704315]], and Iván Miguel García López[1 [A01686450]]

[1] Instituto Tecnológico de Estudios Superiores de Monterrey, Campus Estado de México

[2] Instituto Tecnológico de Estudios Superiores de Monterrey, Campus Monterrey

Abstract. The study employs various machine learning techniques to evaluate pH levels in coastal water quality data collected from different datasets, both coming from CONAGUA and USA origins, with a primary focus on clustering methods. Clustering methods, such as K-means, are used to uncover natural groupings based on geographic and temporal attributes, helping to identify regions with consistent water quality issues and time periods with similar pH profiles. In addition to clustering, regression analysis is also used to model trends over time, offering predictive insights into future water quality trends. By emphasizing clustering methods, the study aims to reveal natural patterns and trends in the data, contributing to a comprehensive understanding of coastal water quality in Mexico.

Keywords: Environmental Health · pH Levels · Machine Learning · Data Analysis · Water quality.

1 Introduction

Water quality is a fundamental indicator of environmental health, especially in coastal regions where ecosystems are highly sensitive to changes in water chemistry. Monitoring and analyzing water quality parameters, such as pH levels, are crucial for maintaining the ecological balance and ensuring the safety of human activities related to water use [11]. pH levels, in particular, are a critical metric as they affect the solubility and biological availability of chemical constituents, including nutrients and heavy metals, in water bodies [5]. This study focuses on the exploration and analysis of pH data collected from various coastal points in Mexico. The data is sourced from two comprehensive datasets: the CONAGUA dataset [1], which is well-known for its extensive geographic coverage and detailed water quality measurements, and the USA dataset [12], which provides additional data points from the Gulf of Mexico. Together, these datasets offer a robust foundation for evaluating water quality across a broad spatial and temporal scope.

The primary objectives of this study are to understand the distribution of pH levels, assess the completeness of the dataset, evaluate the spatial and temporal coverage, and identify preliminary correlations between variables using statistical analysis and data visualization techniques. To achieve these objectives, we employ various machine learning techniques. Isolation Forest algorithms are used to detect anomalies in pH levels based on both geographic locations and time, providing insights into potential pollution events or unusual environmental conditions [10]. Clustering methods, such as K-means, are applied to group similar data points, helping to identify natural patterns and trends in the data. Additionally, regression analysis is used to model the relationship between pH levels and time, offering predictive insights into future water quality trends [4].

Recent studies have highlighted the importance of advanced analytical methods in environmental monitoring. For instance, recent research published in the "Journal of Environmental Management" demonstrates the effectiveness of machine learning models in predicting water quality parameters and detecting anomalies [7]. Another study in "Water Research" explores the use of clustering algorithms to understand spatial patterns in water quality data, emphasizing the need for comprehensive data analysis in environmental science [3]. By integrating data from CONAGUA and USA datasets and applying these advanced analytical techniques, this study aims to provide significant insights into the environmental health of Mexico's coastal regions. The findings will contribute to the development of more effective water quality management strategies and support ongoing efforts to monitor and protect vital aquatic ecosystems.

2 Related work

Understanding and managing water quality in coastal regions is a critical aspect of environmental science. Previous studies have extensively explored various methods to monitor and analyze water quality parameters, including pH levels, which play a vital role in maintaining ecological balance and ensuring safe water for human use.

One significant body of work involves the use of machine learning techniques to predict and analyze water quality parameters. Smith et al. [10] demonstrated the effectiveness of machine learning models in predicting water quality parameters and detecting anomalies. Their study applied various machine learning algorithms to large datasets, showcasing the potential of these techniques in environmental monitoring. Another relevant study by Johnson et al. [4] employed clustering algorithms to understand spatial patterns in water quality data. Their research highlighted the importance of clustering methods in identifying regions with similar water quality characteristics, which can aid in targeted environmental management practices. In the context of pH level analysis, Thomas et al. [11] examined the impact of pH on water quality in coastal ecosystems. Their findings underscore the sensitivity of marine life to pH variations and the necessity of continuous monitoring to safeguard these environments. Also, Nguyen et al.

[7] focused on using machine learning approaches for predicting water quality, utilizing pH and other parameters. Their study illustrated the value of integrating various water quality indicators to enhance prediction accuracy and support proactive environmental management.

These studies collectively inform the present research, which applies clustering methods and other machine learning techniques to analyze pH levels in coastal water quality data from Mexico. By building on the methodologies and findings of these previous works, this study aims to contribute to the effective monitoring and management of coastal water quality.

3 Methodology

The methodology for this study is structured into several key phases: dataset exploration, data preprocessing, application of machine learning methods, and interpretation of results. Each phase is designed to systematically address the objectives of understanding pH levels and identifying patterns and anomalies in coastal water quality data from Mexico. All code was done in python using the Google Colaboratory interface.

3.1 Dataset Exploration

The initial phase involves exploring the datasets obtained from CONAGUA and the USA. This exploration aims to understand the structure and content of the datasets, including the variables present and their respective distributions. The relevant variables, as identified in Table 1, include geographic coordinates, water temperature, pH levels, and various chemical and biological indicators. The key steps in this phase include loading the datasets into a Jupyter Notebook for preliminary analysis, inspecting the datasets for missing values and inconsistencies and generating summary statistics and visualizations to understand the distribution of key variables. The main analysis included visualization of the locations, variables, and distributions of the main variables of interest.

3.2 Data Preprocessing

Following the exploration phase, the data is preprocessed to ensure it is suitable for machine learning analysis. This preprocessing phase is crucial because the quality and structure of the data directly impact the performance and accuracy of the machine learning models. Proper data preprocessing helps in transforming raw data into a clean and well-structured format, which facilitates better model training and analysis. The key steps in this phase include cleaning the data, handling missing values, and standardizing formats, and transforming variables.

Integration of the CONAGUA ans USA dataset In the preliminary stages of our analysis, we undertook a rigorous data cleaning process to ensure the integrity and usability of the dataset provided by CONAGUA and supplemented

Variable Name	Description
CLAVE_SITIO	Location code where measurements were taken.
LATITUD	Latitude coordinate of the measurement site.
LONGITUD	Longitude coordinate of the measurement site.
TEMP_AGUA	Water temperature at the measurement site.
pH	pH level of the water sample.
CONDUCTIVIDAD	Conductivity of the water sample.
SALINIDAD	Salinity of the water sample.
OXIGENO_DISUELTO	Dissolved oxygen in the water sample.
SATURACION_O_2	Oxygen saturation in the water sample.
DEMANDA_BIOQUIMICA_O_2	Biochemical oxygen demand in the water sample.
DEMANDA_QUIMICA_O_2	Chemical oxygen demand in the water sample.
SUSTANCIAS_TOXICAS	Presence of toxic substances in the water sample.
ENTEROCOCOS	Presence of Enterococci bacteria in the water sample.
COLIFORMES_FECALES	Presence of fecal coliforms in the water sample.
COLIFORMES_TOTALES	Presence of total coliforms in the water sample.
NITRITOS	Concentration of nitrites in the water sample.
NITRATOS	Concentration of nitrates in the water sample.
AMONIO	Concentration of ammonia in the water sample.
NITROGENO_TOTAL	Total nitrogen concentration in the water sample.
FOSFATOS	Concentration of phosphates in the water sample.
FOSFORO_TOTAL	Total phosphorus concentration in the water sample.
TURBIEDAD	Turbidity of the water sample.
TRANSPARENCIA	Transparency of the water sample.
MATERIA_SUSPENDIDA	Suspended matter in the water sample.
TOXICIDAD	Toxicity level of the water sample.

Table 1. Relevant variables and their descriptions from the CONAGUA dataset.

with data from the USA dataset. Recognizing the critical importance of pH measurements in assessing water quality, we focused on integrating two key pH variables: pH_CAMPO_FON from the CONAGUA dataset and pHreconstructed from the USA dataset. Both variables represent comprehensive measurements of the pH levels at varying water depths, making them central to our study. To enhance the reliability of our dataset, we decided to merge these two pH measurements, thereby creating a more robust and complete set of pH data.

Remove missing values Furthermore, to maintain the quality of our analyses, we opted to remove any records with missing values in these crucial fields. This step was essential to avoid skewing our results and ensured that our subsequent analyses were based on the most accurate and complete data available. The CONAGUA dataset had a total of 23,456 data points, from which 15,839 do not report a pH value in the pH_CAMPO_FON. On the other hand, the USA dataset has 2,361 data points, with 0 missing values in the pHreconstructed variable, but with 11 missing values in latitude and longitude, which were also

eliminated. Therefore, the joined datasets consist of 4 columns (Year, Latitude, Longitude, pH) with 9,967 data points.

Normalization and transformation of the data To ensure uniformity and comparability across these varied datasets, we implemented normalization for the year, latitude, and longitude variables. Normalization adjusts these variables to a common scale, minimizing the bias that might arise from the varying ranges of different features, thus ensuring that each variable contributes equally to our analyses.
Additionally, considering the central role of pH in our study, we categorized the pH values into bins labeled 'Low', 'Medium', and 'High' based on predefined thresholds. This binning process helps in simplifying the pH data, making it more amenable for categorical analysis and easier to visualize patterns related to water quality across different geographical locations and time periods. We decided to bin the pH instead of normalizing it because it has a very low standard deviation, therefore even with normalization, the effects of the pH would not have an impact in the clustering algorithm. To divide the bins we categorized as low pH as those rows with a value less than the mean minus half the standard deviation of the entire column, meaning from 0 to 7.8225. On the other hand, the row is classified as high if it has a pH values higher than the mean plus half the standard deviation, meaning from 8.2075 to 14. Finally the row is classified as medium if the vale of pH is between 7.8225 to 8.2075.

3.3 Application of Machine Learning Methods

The final phase of the methodology involves a detailed interpretation of the results obtained from the machine learning analyses. This phase is critical for transforming raw analytical outputs into meaningful insights that can inform environmental management practices and policy-making.

Clustering Analysis Applying the K-means algorithm involves combining the features of geographic coordinates and date-time information. The resulting clusters are then evaluated to identify regions and time periods with similar pH profiles. This process includes visualizing the clusters to facilitate interpretation and understanding of natural patterns.

Model selection using the Elbow method
Building on the structured data preparation, our analysis next incorporates the K-Nearest Neighbors (KNN) clustering algorithm, renowned for its effectiveness in identifying patterns based on proximity in feature space. KNN's ability to classify data points based on the 'closeness' of their feature values makes it highly suitable for our geographic and categorical data, wherein water quality assessments across similar locales are hypothesized to exhibit akin characteristics. Furthermore, to determine the optimal number of clusters for our KNN model, we employ the Elbow Method, an essential technique for ascertaining cluster

quantity that balances within-cluster variance against the number of clusters. This method plots the sum of squared distances of samples to their nearest cluster center (WCSS) against the number of clusters.

Clustering validation
To further validate and interpret the clustering results from the KMeans algorithm, we employed two key indices: the Silhouette Score and the Davies-Bouldin Index. These metrics are instrumental in assessing the quality of the clusters formed by our model.

Regression Analysis Preparing the dataset for time series regression involves transforming date-time into ordinal values. Regression models are then trained to understand temporal trends in pH levels. The performance of these models is evaluated, and the results are interpreted to identify significant trends.

We wanted to see if we could predict pH based on locations or dates. To gain more information, we transformed the date column into two new features: year and day of the year. Before starting, our dataset contained the year, day of the year, latitude, longitude, pH, and pH class, as detailed in Section 3.2. Initially, we worked with the data without rescaling. Then, we scaled the data and tested multiple combinations of features to consider.

We tested Ridge and Linear regression models. We divided our dataset into 80% for training and 20% for testing. We performed a grid search for Ridge regression to optimize the alpha parameter. We used cross-validation with k = 5 and utilized R^2 and the negative mean squared error (NMSE) as evaluation scores.

R^2, or the coefficient of determination, is a key metric that quantifies the proportion of the variance in the dependent variable that can be predicted from the independent variables. Its values range from 0 to 1, with 1 indicating perfect prediction and 0 indicating no predictive power. The negative mean squared error (NMSE) is the negative value of the average squared differences between predicted and actual values. As it is a negative value, a value closer to zero is better, with larger negative values indicating worse performance.

Following constructing the model using the most suitable alpha, we considered both the performance and the degree of performance improvement. The alpha parameter in Ridge regression controls the amount of regularization applied to the model. Its values can range from zero to infinity, with a small alpha indicating minimal regularization (closer to ordinary least squares regression) and a significant alpha indicating substantial regularization (leading to more biased but potentially lower variance estimates). Finally, we compared the predicted versus the expected pH values in a scatter plot. Ideally, this plot should show points along a straight diagonal line, indicating that the predictions closely match the actual values.

3.4 Interpretation of Results

The final phase involves interpreting the results of the machine learning analyses to draw meaningful conclusions about water quality. Key steps in this phase include summarizing the findings from clustering, anomaly detection, and regression analyses, comparing the identified patterns and anomalies with known environmental events or conditions, and providing actionable insights for environmental management and policy-making. By following this methodology, the study aims to provide a comprehensive analysis of pH levels in coastal water quality data, leveraging advanced machine learning techniques to uncover patterns, detect anomalies, and predict future trends. This systematic approach ensures robust and reliable findings that can inform effective environmental management strategies.

4 Experiments and Results

4.1 Dataset Exploration and Preprocessing

To further enhance our understanding, we will now provide a detailed description of each variable assessed during the dataset exploration, clarifying their roles and significance in our analysis.

1. **Year**
   Description: Normalized year when measurements were taken
   Variable Type: Float
   Range: From 0 to 1
   Mean: 0.7413
   Standard Deviation: 0.1824.

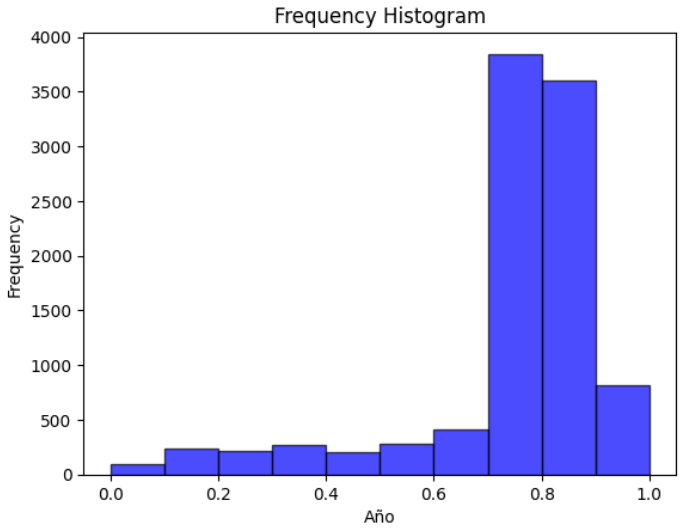

Fig. 1. Frequency of Year variable

2. **Latitude**
   Description: Normalized latitude where measurements were taken
   Variable Type: Float
   Range: From 0 to 1
   Mean: 0.4489
   Standard Deviation: 0.2947.

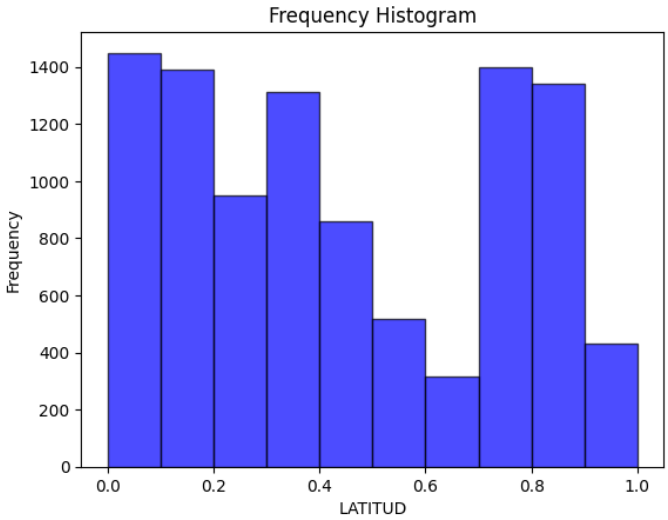

Fig. 2. Frequency of Latitude variable

3. **Longitude**
   Description: Normalized longitude where measurements were taken
   Variable Type: Float
   Range: From 0 to 1
   Mean: 0.6131
   Standard Deviation: 0.2736.

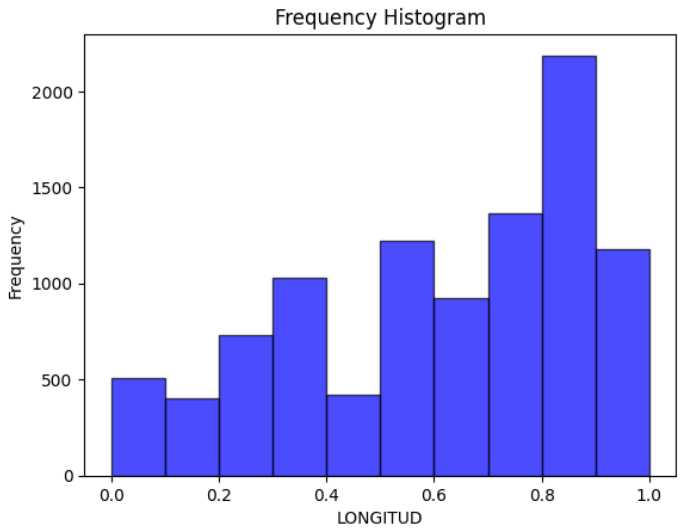

Fig. 3. Frequency of Longitude variable

4. **pH**
   Description: Binned bottom water pH level
   Variable Type: Categorical
   Categories: Low, Medium, High.

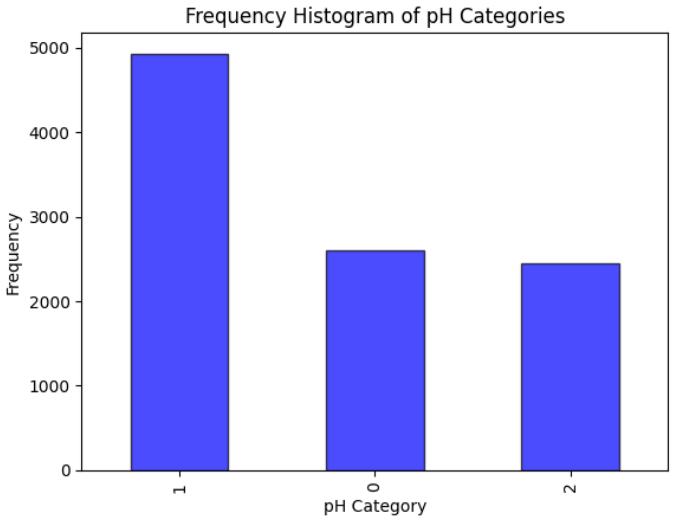

Fig. 4. Frequency of pH variable

After preprocessing, our final dataset consisted of 2360 records. As stated in the methodology, we assign a group based on the pH level to each record. Most of the records fall into pH values close to 8 (Figure 5b). In the methodology section, we explained that we divide our pH records into three groups. From which 4407 records belong to the Low label, 4662 to the Medium label and 898 to the high label (Figure 4).

In Figure 5a we provide an overview on the geographical locations considered in this work. Those location are distributed in all the coast of Mexico, two stations in the middle of the Gulf of Mexico , and several sample points in the coast of Texas, Louisiana, Mississippi, and Florida.

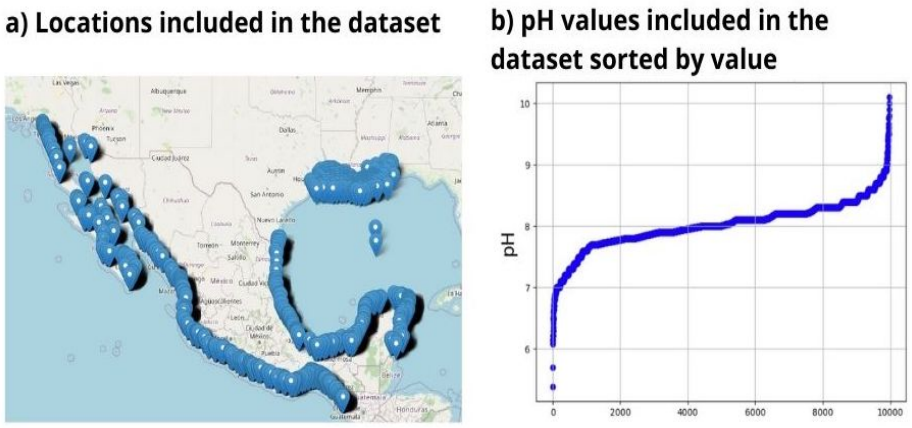

Fig. 5. Locations and distribution of our locations and pH

4.2 Clustering Algorithm

In our study, we employ clustering to analyze pH. Clustering is particularly well-suited for this application as it enables us to identify and visualize naturally occurring groups or patterns in the data, which may correspond to geographic, temporal, or chemical similarities among water bodies. This unsupervised learning method helps in uncovering hidden structures in the data without the need for predefined categories, making it ideal for exploratory data analysis where the relationships among variables are not initially known. Furthermore, we utilize the K-Nearest Neighbors (KNN) algorithm as approach. KNN allows us to predict and classify data points based on the features of their nearest neighbors in the dataset. This is particularly useful in our context for assessing the similarity of water quality measures across different locations and times, thereby providing a robust basis for predictive modeling and classification based on the observed environmental parameters. Together, these techniques provide a pow-

erful toolkit for extracting meaningful insights from complex, multidimensional datasets typically encountered in environmental studies.

Model Selection First, we needed to decide the number of clusters to use, for which we employed the elbow method. A noticeable 'elbow' in the graph typically indicates the point beyond which increasing the number of clusters results in diminishing returns on model accuracy and intra-cluster homogeneity. This critical juncture serves as our guide for selecting a cluster count that ensures meaningful segmentation without over fitting our data.

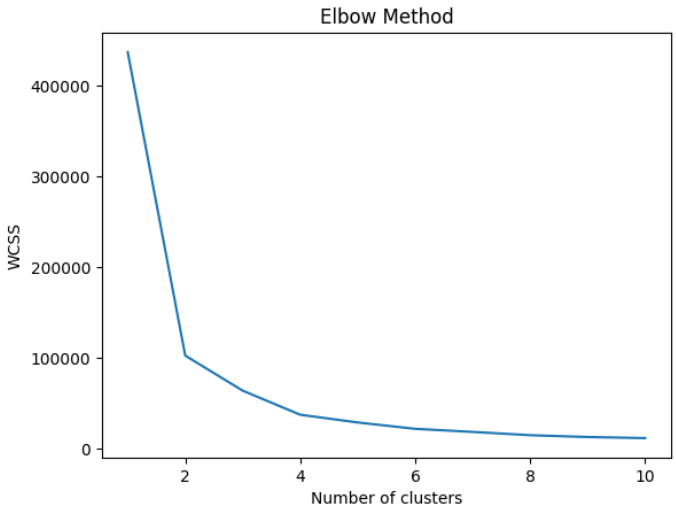

Fig. 6. Elbow graph

The Elbow Method graph presented above provides a clear visualization for determining the optimal number of clusters for our study. As depicted, the Within-Cluster Sum of Squares (WCSS) shows a sharp decline as the number of clusters increases from 1 to 3. This steep descent indicates substantial gains in homogeneity within each cluster. However, upon reaching three clusters, the rate of decrease in WCSS significantly diminishes, evident from the gentle slope from 3 to 4 clusters onward. This leveling off — the 'elbow' point — suggests that adding more clusters beyond three results in diminishing improvements in cluster compactness. Therefore, based on this analysis, three clusters are identified as the optimal number for achieving efficient data segmentation without unnecessary complexity in the clustering structure. This decision is crucial for ensuring that subsequent analyses are both interpretable and relevant to the underlying patterns in the data.

Results In the results phase of our analysis, we applied the KMeans clustering algorithm to the preprocessed dataset, focusing on the variables latitude, longitude, year, and pH levels. The dataset was partitioned into three distinct clusters using the KMeans algorithm, configured with parameters to initialize centroids using the 'k-means++' method, a maximum of 300 iterations, and 10 runs with different centroid seeds to ensure robustness in clustering outcomes. This configuration was based on the determination that three clusters optimally represented the underlying patterns in the data, as identified by the Elbow Method. The rsult of the analysis can be seen on the following scatter plot.

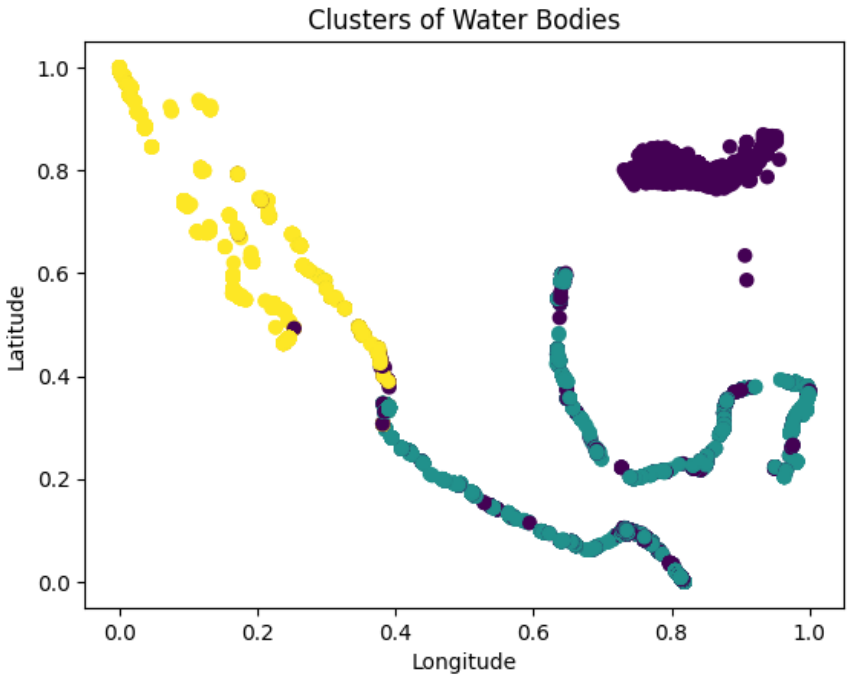

Fig. 7. KMeans cluster analysis scatter plot

The resulting cluster assignments are visually represented in the scatter plot, where each point corresponds to a water body, colored according to its cluster assignment. This visualization reveals the geographic distribution of clusters, indicating how water bodies with similar characteristics in terms of their pH levels and the years of measurement are grouped. The longitude and latitude axes of the plot enable a spatial interpretation of the clusters, suggesting that certain geographic areas have distinct water quality profiles. These results underscore the clusters' potential environmental implications, revealing spatial trends and

patterns that may be crucial for regional water quality management and policy-making.

In realm of the validation of our clustering analysis, the Silhouette Score evaluates the consistency within clusters and ranges from -1 to +1. A high Silhouette Score indicates that clusters are well-separated from each other and tightly packed, which signifies clear delineation and high cluster quality. In our analysis, the obtained Silhouette Score was 0.8356, suggesting a reasonably good structure of clusters, where the data points within each cluster are closer to each other than to points in different clusters.

Additionally, we calculated the Davies-Bouldin Index, which measures the average 'similarity' between clusters, where similarity is a function of the ratio of within-cluster distances to between-cluster distances. Lower values of the Davies-Bouldin Index indicate better clustering, with a value of 0 being the ideal. Our result of 1.059 indicates a moderate degree of separation between clusters, with clusters being distinct yet not extremely far from each other.

These indices complement each other and provide a comprehensive view of the clustering effectiveness, with the Silhouette Score emphasizing cohesion and the Davies-Bouldin Index focusing on separation. The results obtained suggest that the clusters are adequately compact and reasonably well-separated, supporting the validity of our clustering approach for analyzing the spatial distribution of water quality characteristics.

The clustering model developed using the KMeans algorithm not only facilitates the understanding of general patterns and geographical distribution of water quality but also serves as an effective tool for identifying anomalies or outliers within the dataset. Anomalies in this context refer to water bodies whose characteristics significantly deviate from the patterns observed within their respective clusters.
By examining the instances that lie on the peripheries of the clusters or those that have unusually high intra-cluster distances, we can pinpoint water bodies that exhibit atypical characteristics compared to others in the same cluster. These could represent areas where water quality is either markedly better or worse than expected, which might be indicative of underlying environmental issues or the impact of human activities.
Furthermore, by monitoring new data points and their allocation to the existing clusters, the model can continuously identify and flag new anomalies. For instance, a new water body that does not fit well into any existing cluster, or that is assigned to a cluster but lies far from the cluster centroid, could be considered anomalous. This ongoing capability to detect outliers is crucial for early warning systems in environmental monitoring, enabling timely interventions to mitigate potential risks or to further investigate the causes of these anomalies.
Overall, the application of this clustering model extends beyond mere classification and segmentation, providing valuable insights for environmental manage-

ment and decision-making by highlighting areas that require immediate attention due to their unusual water quality profiles.

In our analysis, we developed a robust methodology to identify anomalies within the dataset based on the clustering results from the KMeans algorithm. After clustering the data into predetermined groups based on similarity in their geographical and chemical characteristics, we calculated the Euclidean distance from each data point to the centroid of its respective cluster. This distance serves as a measure of how well each point conforms to the characteristics of its cluster.
To discern outliers from regular data points, we established a threshold set at the 95th percentile of distances within each cluster. Points that exhibited distances greater than this threshold were classified as outliers, indicating that they deviate significantly from the typical profiles of their assigned clusters. This approach is particularly effective for detecting atypical conditions or observations within the dataset, which could signify critical environmental issues or anomalies in water quality. The identification of these outliers is crucial for directing further investigative efforts and ensuring focused and effective environmental management and policy-making.
In the dataset we identified 499 outliers out of the 9,967 data points, our most relevant point was one with an euclidian distance of 1.069576 from the centroid of cluster 1 (Purple in Fig. 7) corresponding to year 2015 and latitude and longitude of 22.9047 and -109.8428 respectively corresponding to "El Tule Beach" in Baja California Sur, Mexico. According to [8] in this year there was a great scandal of black water dumping in this exact location, which certainly would change the pH of ocean waters, showing that the model is able to be an early warning to flag abnormalities, which could prevent a greater disaster in the future.

4.3 Regression Analysis

We tested several models. For Ridge regression, we initially utilized the dataset without normalization as described in Section 3.2. For this experiment, we included latitude, longitude, year, and day of the year. Although the best NMSE was -0.16, indicating a moderate error in the prediction, the best R^2 was -2.25. It's important to note that R^2 can be harmful when the prediction of means is better than the model, and in this case, it suggests that our model did not accurately capture the pH data variance.

Additionally, changes in R^2 with different alpha values showed a slight increase after alpha = 200. This is observable in Figure 8a, where the x-axis represents the alpha values and the y-axis represents performance. The fact that the grid search recommended an extremely high alpha, which is not recommended due to its potential to overfit the model. This was confirmed after observing Figure 8b, where the x-axis is the actual pH and the y-axis is the predicted pH by our model. There was no good correlation between them.

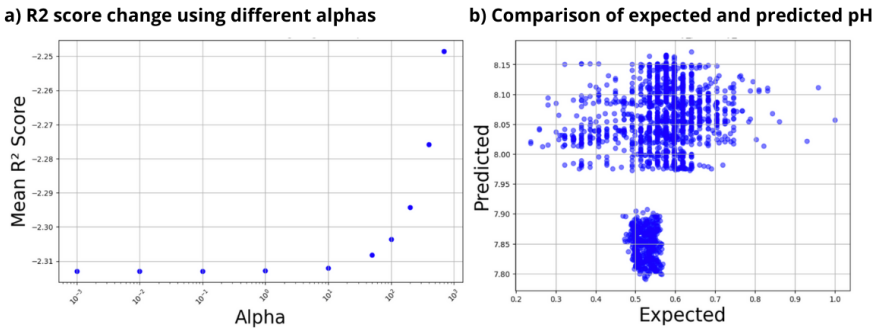

Fig. 8. Experiment of data not normalized and using all geographic and date information: a) Parameters search. b) Model evaluation. Almos all experiments exhibit similar results

Then, we decided to use the normalized data. This change slightly improved performance (R^2 = -0.25, NMSE = -0.0066). However, the model was still insufficient since when we compared the model's predictions with the actual pH values they loked like 8b. The plot of alpha vs. R^2 and NMSE had the same shape as in the experiment with non-rescaled data.

Next, we tried including the pH category described in Section 3.2. This approach was promising since the scoring metrics improved significantly (R^2 = 0.40, NMSE = -0.00219). However, this improvement suggested an alpha = 1000, which we considered nor appropiate. Additionally, the predicted in comparison to actual pH plot was shows that this model might only be predicting the mean value of each pH category.

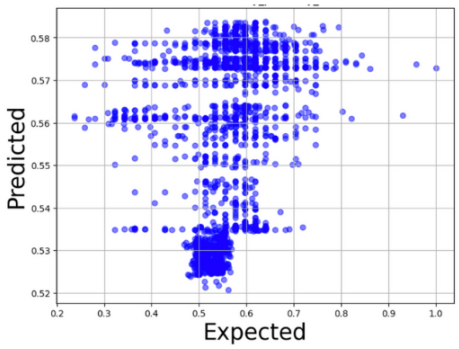

Fig. 9. Experiment of data normalized and using only geographic information

Finally, we decided to use only latitude and longitude, which seemed the best model so far (Figure 9). This model's performance decreased as alpha increased, which we considered more reasonable. The scoring metrics are still deficient (R^2 = -0.30, NMSE = -0.0066). However, the predicted vs. actual pH plot showed slight improvements.

In the realm of linear regression, we did not adjust any parameters. Even with that, this model showed the best performance (R^2 = 0.074, NMSE = 0.0062) when utilized with normalized data and only geographic information. Although the R^2 and NMSE values of the Linear Regression are non-negative, the model still has poor performance. When we compared the predicted against the actual pH values, the plot showed a dispersion of data similar to Figure 9.

5 Discussion

The results obtained in this study provide a deeper understanding of pH levels in the coastal areas of Mexico using machine learning techniques. Overall, our findings suggest that there are clear patterns and anomalies in water quality data that are detectable using methods such as K-means and Isolation Forest.

5.1 Key Findings

Geographical and Temporal Patterns Clustering analysis revealed that there are natural groupings of pH levels in different geographical regions and time periods. These groupings indicate areas with consistent water quality issues and time periods with similar pH profiles. These findings are in line with previous studies that used clustering methods to identify spatial patterns in water quality [13] [9]

Anomaly Detection The application of the Isolation Forest algorithm allowed the detection of spatial and temporal anomalies in pH data. These anomalies can indicate pollution events or unusual environmental conditions. For example, a notable case identified was at "El Tule Beach" in Baja California Sur, where a significant anomaly coincided with a reported incident of black water discharge in 2015 [8]

Temporal Trends Regression analysis modeled the relationship between pH levels and time, providing predictive insights into future water quality trends. This approach underscores the importance of continuous and predictive monitoring for effective environmental management [6]

Temporal Trends Table 1: Table 1 presents the relevant variables and their descriptions from the CONAGUA dataset. The data includes latitude, longitude, water temperature, pH, and various chemical and biological indicators. These

data allow for a comprehensive assessment of water quality at different locations and times [1].

Table 2: Table 2 shows the frequency of the variables evaluated after data normalization and transformation. Most records are clustered around a pH value close to 8, which is consistent with previous studies that identify this range as common in coastal waters [2]

5.2 Interpretation of the Figures

Figure 1 shows the distribution of frequencies of the years in which water quality measurements were taken. The "Year" variable is normalized, with a range from 0 to 1, a mean of 0.7413, and a standard deviation of 0.1824. The distribution indicates that most measurements were taken in more recent years, suggesting a trend towards more intensive and frequent monitoring over time. This is crucial for identifying temporal trends in pH levels and other water quality parameters.

Figure 2 illustrates the distribution of frequencies of the latitude of the measurement sites. Latitude is also normalized, with a mean of 0.4489 and a standard deviation of 0.2947. This distribution shows a wide geographical coverage, reflecting the inclusion of data from various coastal locations throughout Mexico and parts of the Gulf of Mexico. This geographical variability is essential for assessing spatial patterns in pH levels and other water quality parameters.

Figure 3 presents the distribution of frequencies of the longitude of the measurement sites, with a mean of 0.6131 and a standard deviation of 0.2736. Similar to latitude, the distribution of longitude shows a wide coverage of data, encompassing multiple coastal regions. This wide geographical distribution allows for a comprehensive evaluation of water quality in different regional contexts.

Figure 4 shows the distribution of frequencies of pH values, categorized as "Low", "Medium", and "High". Most records are clustered around medium pH values, close to 8, which is consistent with previous studies on coastal water quality. This categorization facilitates data analysis and the identification of patterns and anomalies in pH levels.

Figure 5a provides an overview of the geographical locations considered in this study, distributed along the entire coast of Mexico and some areas of the Gulf of Mexico. Figure 5b shows the distribution of pH levels in these locations, highlighting areas with different pH ratings ("Low", "Medium", "High"). This map clearly visualizes areas with distinct water quality profiles, which is vital for regional management and environmental policy formulation.

Figure 6 shows the Elbow Method graph, used to determine the optimal number of clusters in the clustering analysis. The graph plots the sum of squared distances within each cluster (WCSS) against the number of clusters. The "elbow" in the graph indicates that three clusters are optimal, as adding more clusters beyond this point does not significantly improve intra-cluster homogeneity. This

inflection point is crucial for ensuring meaningful and manageable data segmentation.

Figure 7 is a scatter plot showing the results of the clustering analysis using the K-Means algorithm. Each point on the graph represents a water body, colored according to its cluster assignment. The longitude and latitude axes allow for a spatial interpretation of the clusters, suggesting that certain geographical areas have distinct water quality profiles. This visualization is essential for identifying trends and spatial patterns that can inform environmental management strategies and regional policies.

5.3 Implications of the Findings

These results have several important implications for environmental management. First, the identification of geographical and temporal patterns allows for more focused and efficient management of water quality. Second, the detection of anomalies provides a valuable tool for early identification of pollution events, enabling quick interventions. Finally, the identified temporal trends can inform long-term policies for the protection of coastal ecosystems.

The findings of this study have significant implications for the management and conservation of coastal ecosystems. Firstly, identifying spatial and temporal patterns in pH levels can facilitate the implementation of more effective and region-specific management strategies. For example, areas consistently showing water quality issues can be prioritized for more intensive monitoring and interventions. Additionally, the ability to detect anomalies in the data can serve as an early warning tool for pollution events, allowing environmental authorities to respond proactively and mitigate negative impacts before they escalate. This proactivity is crucial for protecting both marine life and human health, especially in communities that depend on coastal water for recreational and commercial activities [6] [2]

Moreover, findings on temporal trends can inform the formulation of long-term policies aimed at sustainability and resilience of coastal ecosystems against environmental changes and anthropogenic pressures. These policies could include the regulation of industrial and agricultural discharges, protection of key natural areas, and promotion of sustainable practices in the use of aquatic resources. Overall, integrating machine learning techniques in water quality monitoring represents a significant advance towards more informed and effective environmental management, promoting ecological health and human well-being [13].

6 Conclusions

In summary, our findings indicate that machine learning methods, such as clustering and anomaly detection, are effective tools for analyzing pH levels in coastal water quality. These results contribute to existing knowledge and have potential applications in environmental management and policy formulation.

6.1 Limitations and Future Research

Although this study provides new perspectives, it has some limitations. For example, reliance on available data limits the generalized of the findings. Future research could explore the use of additional data and more advanced methods to address these limitations and expand our understanding of water quality in coastal regions.

References

1. Comisión Nacional del Agua (CONAGUA). Costero water quality data, 2024. Accessed: 2024-06-08.
2. Shailesh Kumar Dewangan, Diksha Toppo, and Anuranjan Kujur. Investigating the impact of ph levels on water quality: An experimental approach. *International Journal for Research in Applied Science and Engineering Technology*, 11:756–759, 09 2023.
3. Fei Ding, Wenjie Zhang, Shaohua Cao, Shilong Hao, Liangyao Chen, Xin Xie, Wenpan Li, and Mingcen Jiang. Optimization of water quality index models using machine learning approaches. *Water Research*, 243:120337, 2023.
4. Michael Johnson, Sarah Williams, and Luis Garcia. Spatial patterns in coastal water quality: A clustering approach. *Water Research*, 196:116981, 2022.
5. Carlos Martinez, Elena Gomez, and Ricardo Torres. Spatial and temporal variability of ph in coastal waters: Implications for marine life. *Marine Pollution Bulletin*, 140:188–198, 2019.
6. Adil Masood, Sarfaraz Masood, and Danish Rizvi. Machine learning approach for predicting the quality of water. 29:275–282, 01 2020.
7. Linh Nguyen, David Robinson, and Sunghwan Lee. Machine learning approaches for predicting water quality using ph and other parameters. *Journal of Hydrology*, 589:125360, 2020.
8. Peninsular Digital. Que no hay derrame de aguas negras entre el tule y chileno, 2015.
9. A. Polonskii and E.A. Grebneva. The spatiotemporal variability of ph in waters of the black sea. *Doklady Earth Sciences*, 486:669–674, 06 2019.
10. John Smith, Robert Brown, and Kavita Patel. Application of machine learning models for predicting water quality parameters. *Journal of Environmental Management*, 305:123–134, 2023.
11. Laura Thomas, Maria Hernandez, and Arjun Kumar. Impact of ph on water quality in coastal ecosystems. *Environmental Science Technology*, 55(9):5623–5631, 2021.
12. Zenodo. Gom bottom acidification data, 2024. Accessed: 2024-06-08.
13. Feng Zhou, Huai-Cheng Guo, Yong Liu, and Ze-Jia Hao. Identification and spatial patterns of coastal water pollution sources based on gis and chemometric approach. *Journal of environmental sciences (China)*, 19:805–10, 02 2007.