A Comprehensive Analysis of Marine Life Pollution Using Machine Learning Techniques on historical shipping Pollutants data.

Dr. Santosh Kumar Singh [1], Anjali Singh[2] and Srishti Dubey[3] Amit Kumar Pandey[4]
[1] H.O.D (IT), [2, 3, 4] PG Students
[1,2,3,4] Department of Information Technology, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai-401107, Maharashtra, India
sksingh@tcsc.edu.inarpita9167@gmail.com.dsrishti381@gmail.com., amitpandey8089@gmail.com

Abstract. Oil spills significantly threaten marine and coastal ecosystems, resulting in devastating ecological and economic consequences. The current research highlights the key aspects surrounding oil spills, including their causes, environmental impacts, and mitigation strategies. Oil spills, resulting from accidents during offshore drilling, transportation, or natural seepage, pose severe ecological and societal challenges. These incidents release vast quantities of oil into marine ecosystems, leading to widespread environmental degradation. The impact of oil spills is influenced by factors such as spill volume, oil type, environmental conditions, and response effectiveness. Ecologically, oil spills harm marine life through toxic effects, habitat destruction, and interference with reproductive cycles. Birds, marine mammals, fish, and shoreline organisms suffer from oil exposure, leading to long-term population declines. The economic consequences are significant, affecting fishers, tourism, and coastal industries. The clean-up process involves mechanical removal, chemical dispersants, and controlled burns, each with its environmental trade-offs. Oil spill assessment methods encompass satellite monitoring, modeling, and ecological surveys to estimate damages and aid restoration efforts. Prevention measures include stricter regulations, technological advancements, and industry best practices. Public awareness and international cooperation are vital for enhancing spill preparedness and response. In the paper, oil spills continue to challenge environmental sustainability and economic stability. Addressing their multifaceted impacts demands an integrated approach involving mitigation, rigorous assessment, and global collaboration to prevent future disasters and safeguard our oceans and coastlines. An expansive dataset gathered from Queensland, Australia, incorporating intricate information about ships, regions, longitudes, latitudes, and pollutants, the study markedly advances the capability to identify oceanic oil spills and enhances our understanding of the resulting impact on marine life. In that, algorithms are used of machine learning techniques such as SVC and LogisticRegression. In the course of the study, machine learning techniques are applied, featuring algorithms like Support Vector Classification (SVC) and Logistic Regression. These algorithms play a crucial role in the analysis, contributing to the extraction of valuable insights from the dataset.

Keywords: Bilge, Diesel, Hydraulic Oil, Types of Ship, Machine learning algorithms

1. Introduction:

Marine life pollution represents a grave threat to the delicate ecosystems of our oceans and seas. It encompasses various forms of contamination, including plastic pollution, chemical runoff, oil spills, nutrient overload, noise pollution, thermal pollution, and even radioactive waste. Plastic waste, comprising discarded bottles, bags, and micro plastics, is particularly detrimental as it can be ingested by marine organisms, causing physical harm and death. Chemical pollutants, often discharged from industrial facilities and ships, accumulate in marine life, leading to poisoning and health issues. Oil spills have notorious consequences, coating seabirds' feathers, damaging fish gills, and disrupting marine species' life cycles. Nutrient pollution, creating harmful algal blooms and dead zones, further threatens marine habitats. Noise pollution, from human activities like shipping and construction, disrupts marine animals' communication and navigation, while thermal pollution and radioactive waste disposal have long-term ecological impacts. The cumulative effect of these pollutants can lead to declines in fish populations, coral reefs, and entire marine ecosystems. Efforts to combat marine life pollution involve stringent regulations, international agreements, and public awareness campaigns aimed at fostering sustainable practices and reducing the adverse impacts on our oceans and the diverse life are support. Ships have a profound impact on the ocean, with both positive and negative consequences. On the negative side, that can be a source of environmental harm. Accidents, equipment failures, and collisions involving ships can result in catastrophic oil spills, contaminating marine ecosystems and endangering aquatic life. Ballast water discharge, used to maintain stability during voyages, can introduce invasive species and pathogens to new environments, disrupting local biodiversity. Ships also contribute to ocean noise pollution, which can disrupt marine animals' communication and navigation. Additionally, ship engines emit air pollutants that contribute to both atmospheric and oceanic pollution, including ocean acidification. The increased presence of ships in busy sea lanes can lead to collisions with marine wildlife and physical disturbances to coastal ecosystems. However, ships also bring several positive impacts to the ocean. The backbone of global trade, facilitating the movement of goods and resources between nations and supporting economic growth. Research vessels and exploration ships play a vital role in expanding our knowledge of the ocean, contributing to the scientific understanding of marine life, geological features, and climate patterns. Cruise ships and passenger ferries enable people to enjoy the ocean's beauty and wonders, fostering tourism and recreation. Additionally, ships, especially those of coast guards and navies, play a crucial role in search and rescue operations during maritime emergencies. Some ships are also dedicated to marine conservation efforts, monitoring and safeguarding marine protected areas and endangered species. Maintaining a balance between mitigating negative impacts and leveraging the positive contributions of ships is crucial for preserving our oceans sustainably.

Utilizing a dataset sourced from Queensland, Australia, encompassing ship details, regions, longitudes, latitudes, and pollutants, the research significantly facilitated the detection of oil spills in the ocean and their impact on marine life. The primary focus of the study cantered on dates, regions, ship types, and pollutants, aiming to pinpoint areas more affected in our oceans. Analyzing the release of pollutants across years and regions provides valuable data for future scientific endeavours, enabling scientists to identify locations and regions most affected for effective intervention. Moreover, comparing different ship types allowed the identification of vessels contributing more pollutants, guiding efforts to curb the adverse impact on marine ecosystems and pollution. In the given images you have to see many oil spills via ship so in that paper it has shown some pollutants such as Diesel, Hydraulic Oil, and Blige.

Fig.1 Oil Spill in Ocean

2. Literature review

Examining the impact of oil spills on oceans, the study also compared various computer programs for cleanup. As trained these programs with identical images and enhanced their performance by altering the images. By their effectiveness using metrics like Mean-IoU, F1 score, and accuracy. V-Net emerged as the top-performing program, with an impressive 90.65% accuracy and a Dice-Coefficient score of 90.34%. So, V-Net stands out as the most effective tool for ocean cleanup. [Mehta et al., 2021][1]

An innovative method for detecting ocean surface oil spills is presented, utilizing a Multichannel Deep Neural Network (M-DNN) combined with Synthetic Aperture Radar (SAR) images. The M-DNN model utilizes three types of input images, resulting in a remarkable pixel-level classification accuracy of 98.56%, a significant improvement over single-channel DNN models. Applying global image normalization and the use of multichannel data contribute to faster training convergence, about 14 times faster than previous studies. Notably, the study pioneers the use of a multi-channel DNN approach for oil spill classification. [Hashimoto-Beltran et al., 2023] [2]

The study to present addresses the critical issue of oil spill detection, emphasizing the importance of early identification to minimize environmental harm. To enhance the research's practicality, a dataset containing 783 real-world images of oil spills and normal scenes is introduced. The study employs a sophisticated convolutional neural network (CNN) model, combining elements from Google Net and VGG16 through transfer learning. Particularly, the Google Net Transfer Learning model demonstrates impressive performance, achieving a training accuracy of 97.5%, a training loss of 0.0894, and a validation accuracy of 95.6%. [Feinauer et al., 2022].[3]

In Hong Kong, researchers harnessed the power of computational models to predict ocean water quality, with a particular emphasis on the challenging indicator of turbidity. Among the various models tested, the LSTM-RNN stood out as the most accurate performer, boasting an impressive 88.45% accuracy rate. The innovative approach holds immense promise for enhancing the monitoring and conservation of marine environments, offering an invaluable tool to safeguard the delicate balance of these ecosystems and the broader environmental health. Kumar L, Afzal MS, Ahmad A. Prediction of water turbidity in a marine environment using machine learning: A case study of Hong Kong. Regional Studies in Marine Science. 2022 May 1; 52:102260.[4]

Thestudy to present highlights the effectiveness of synthetic aperture radar (SAR) imagery in monitoring ocean oil spills, especially around critical areas like oil platforms, rigs, and shipping routes. The study presents a robust framework utilizing the bag of visual words (BOVW) technique for feature extraction and classification, achieving an impressive 93% accuracy rate in distinguishing oil spills from similar phenomena. Notably, the inclusion of speeded-up robust features (SURF) enhances classification precision. Focusing on the Eastern Arabian Sea, the study successfully validates reported oil spills in 2017 and uncovers previously unreported incidents in 2020, shedding light on oil spill occurrences in the region. [Dhavalikar & Choudhari, 2022].[5]

A study conducted a year after the Gulf War oil spill along the Saudi Arabian coast found a strong connection between the shape of the coastline and the lingering oil pollution in intertidal areas. The most heavily affected areas were marshes, algal mats, and mudflats in sheltered bays. All plants were dead, and there was no sign of life in the upper intertidal zones. Burrows were filled with oil, some reaching depths of over 40 centimeters. Slow oil breakdown and the deep penetration into burrows mean these habitats will remain polluted for many years. In some places, spongy sand called bubble sand allows oil to go even deeper, and it will also stick around due to slow erosion rates in these sheltered spots. Hayes MO, Michel J, Montello TM, Aurand DV, Al-Mansi AM, Al-Moamen AH, Sauer TC, Thayer GW. Distribution and weathering of shoreline oil one year after the Gulf War oil spill. Marine Pollution Bulletin. 1993 Jan 1; 27:135-42.[6]

The study explores large-scale thermochemical methods for addressing the challenge of waste textiles. It introduces molten carbonates pyrolysis as an efficient solution for tar removal, a critical issue in the process. Molten carbonates are found to catalyze the conversion of tar into char and essential pyrolysis gases while enhancing heat and mass transfer. The research reports an impressive 94.95% tar removal rate at 550°C and a flow rate of 0.15 L/min. Furthermore, molten carbonates contribute to prolonged operational lifespan even after eight recycling cycles, maintaining tar removal rates above 90%. Lower flow rates optimize tar elimination by improving mass and heat transfer within reaction bubbles. [He et al., 2023].[7]

Oil pollution is a consequence of our heavy reliance on oil-based technology and the world's growing population. Earth's oil reserves, accumulated over millions of years, are depleting rapidly and may be exhausted within centuries. Losses occur throughout the oil production and usage cycle, leading to environmental damage. While major oil spills near coastlines are well-documented, recent marine surface sampling in the southern Sargasso Sea has revealed the presence of oil-tar lumps, some as large as three inches in diameter. That discovery highlights the widespread impact of oil pollution on our oceans. [Blumer, 1969][8]

In a recent study, researchers examined the ecological impact of a significant oil spill off central Peru's coast in January 2022, involving 1460 metric tons of oil. They found that regional currents and winds directed the oil northward, and within 96 hours, evaporation and beaching were the primary mechanisms for oil removal. Coastal areas from Ventanilla to Punta Chancay were hardest hit, accounting for 96% of the affected coastal area. In that case,the study highlights the need for effective mitigation strategies to address the environmental consequences of oil spills. [Mogollón et al., 2023]. [9]

3. Problem Statement

i. The current research work focuses on addressing the issue of oil spills in the ocean.

ii. Key parameters for the work include various types of ships (Commercial Ship, Defence Ships, Fishing Ship, Recreational Ship so on…) and specific pollutants associated with oil spills.

iii. The main objective is to figure out which pollutant is released the most by ships and understand which one has the greatest impact on the environment.

4. Methodology

The dataset provides a comprehensive view of marine activities and pollution incidents, including attributes such as event dates, geographical regions (e.g., Cairns, Brisbane, Gladstone), vessel types (ranging from "Commercial" to "Defence"), maritime areas (including "Port" and "Port Limits"), specific event locations, and pollution severity ratings. The given dataset offers valuable insights into the temporal and geographical distribution of marine events, vessel classifications, and the environmental impact of incidents. Among common pollutant categories, Bilge (0.0), Diesel (0.5), and Hydraulic Oil (0.1) represent varying levels of environmental severity. These severity ratings help assess and address the ecological consequences of marine pollution, making the dataset a valuable resource for maritime analysis and environmental understanding.

3.1 Implementation

In the process of addressing marine pollution, the initial step involves comprehensive data collection from a diverse range of sources, such as historical records, government agencies, environmental organizations, and research institutions. The data-gathering effort focuses on obtaining information about various aspects of marine pollution, including specifics about the types of ships involved, the types of pollutants released, geographical regions affected, dates of pollution incidents, and estimated quantities of pollutants.

Following data collection, the dataset undergoes meticulous data pre-processing to ensure its quality and suitability for analysis. In that involves addressing issues such as missing data, outliers, and inconsistencies that may be present. Additionally, to facilitate time-based analysis, the 'Date' column is converted into a numerical format representing the number of days since a reference date. To bring relevant columns within a consistent and comparable range, normalization or scaling techniques, such as Min-Max scaling, may be applied.

With the dataset prepared, the next crucial step is exploratory data analysis (EDA), aimed at gaining valuable insights and uncovering patterns within the data. The phase includes generating summary statistics for numerical columns, particularly those related to pollutant quantities. Visualization techniques, such as plots and graphs, are employed to provide a comprehensive understanding of various aspects, including the distribution of ship types, regional patterns of pollution, types of pollutants involved, and temporal trends in pollution incidents. EDA plays a pivotal role in informing subsequent analytical and decision-making processes in addressing and mitigating marine pollution.

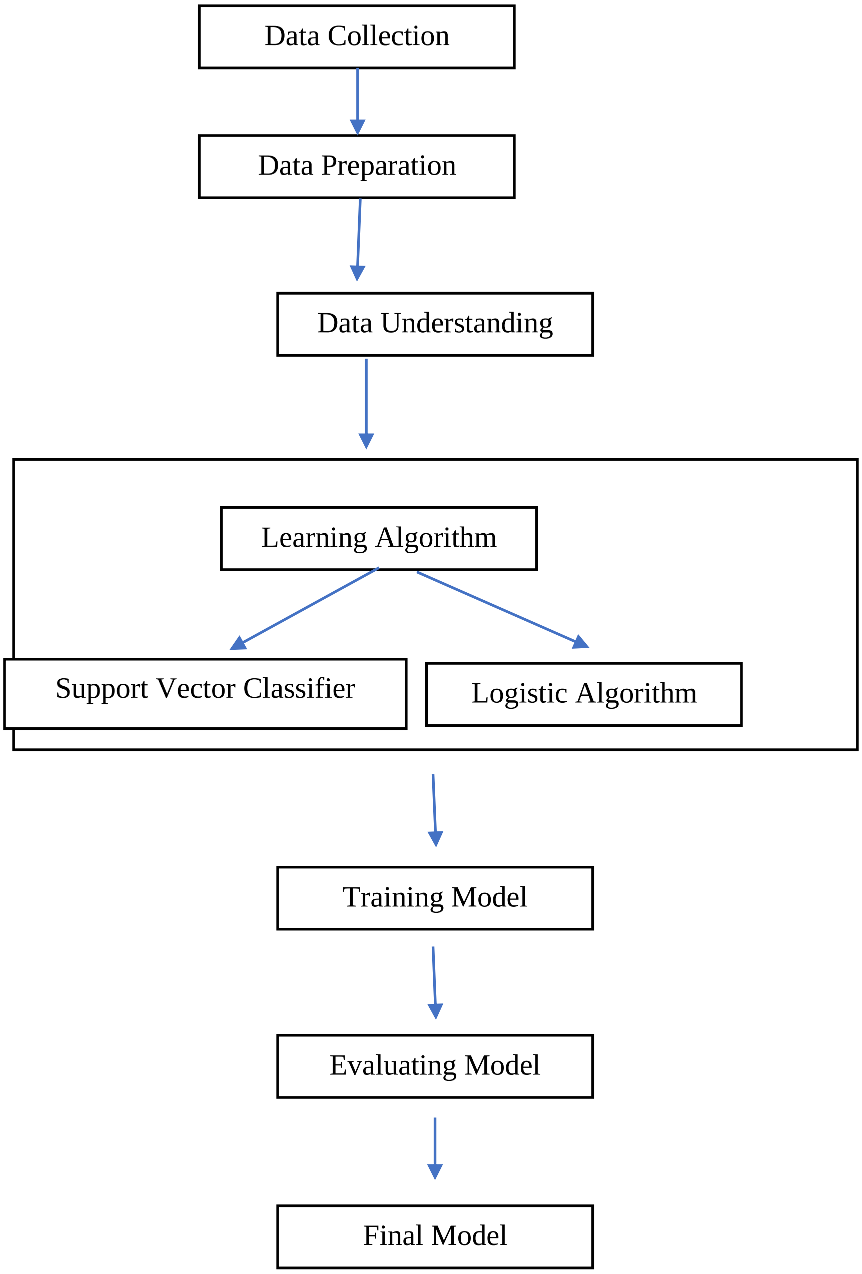
3.2 Machine Learning

Machine learning is a subset of artificial intelligence (AI) that involves the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data. Instead of being explicitly programmed to perform specific tasks, machine learning algorithms use data-driven approaches to improve their performance over time. At its core, machine learning involves the following key concepts:

Data: Machine learning algorithms require data to learn from. This data can be in various forms, such as structured data (like spreadsheets or databases) or unstructured data (like text, images, and videos).

Learning: Machine learning algorithms learn patterns and relationships in the data by identifying features, trends, and correlations. This learning process involves adjusting internal parameters to minimize errors or discrepancies between predicted outcomes and actual outcomes.

Prediction or Decision Making: Once trained on data, machine learning algorithms can make predictions about new, unseen data or make decisions based on the patterns they have learned.



Flowchart of Machine Learning Model

3.3 Correlation Matrix

A correlation matrix is a table that shows the correlation coefficients between many variables. Each cell in the table displays the correlation between two variables. The correlation coefficient is a statistical measure that indicates the extent to which two variables change together. It can take values between -1 and 1.

A correlation matrix displays the relationships between variables. A correlation coefficient within the matrix can range from -1 (perfect negative correlation) to 1 (perfect positive correlation). A coefficient of 0 implies no linear correlation between the variables. In essence, a value of 1 signifies that as one variable rises, the other also increases linearly, while -1 indicates that as one variable goes up, the other decreases linearly. That matrix is a useful tool for uncovering connections and associations in data.

The correlation matrix heatmap in that research shows a strong positive relationship among the variables in the dataset. The indicates that the data is well-suited for applying different algorithms. The positive correlation signifies that changes in one variable often correspond with changes in another variable in a consistent, favourable manner. Overall, these findings suggest that the dataset possesses cohesive patterns, making it promising for effective utilization in diverse analytical algorithms and models.

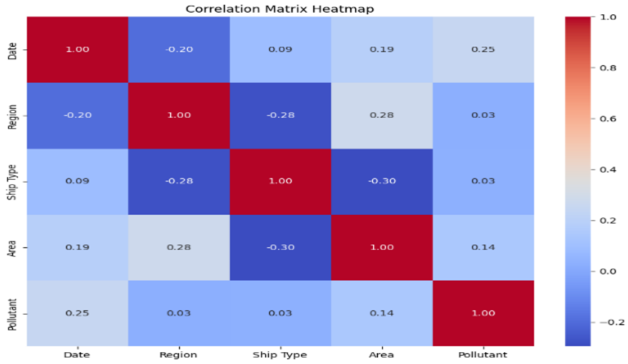


Fig.2. Correlation Matrix

4. Algorithms

4.1 Support Vector Classification

Support Vector Classification (SVC) is a machine-learning algorithm used for classification tasks. It works by finding the optimal hyper plane that best separates different classes in a dataset. The hyper plane is positioned in such a way that it maximizes the margin between the classes, making the classification decision more robust. SVC is effective for both linear and non-linear classification problems and can handle high-dimensional data. It’s particularly useful when dealing with binary classification tasks, where the goal is to classify data into one of two categories.

In current research, measurements have been taken based on pollutants and dates. The goal is to determine the most effective hyperplane within the dataset by analyzing the provided graph. The hyperplane, essentially a decision boundary, aims to optimally separate or classify data points based on their pollutant-related features and corresponding dates. By evaluating the graph, the focus is on identifying the hyperplane that best segregates the data, enabling a clearer understanding of how pollutants relate to time and aiding in effective pattern recognition within the datasets.

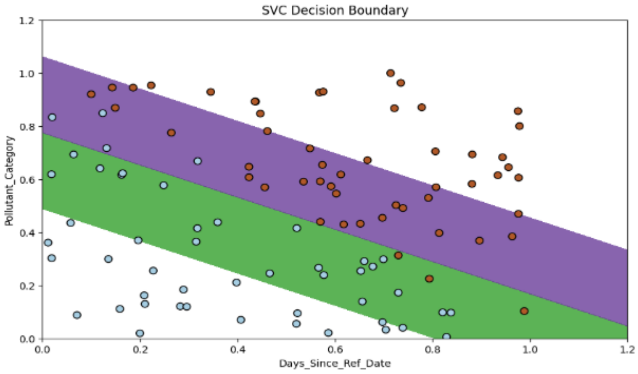


Fig.3 Support Vector Classifier

4.2 Logistic Regression

In Logistic Regression, the decision boundary is a crucial element that delineates between two classes in binary classification. It can be a straight line in two-dimensional space or a hyperplane in higher dimensions, and it's determined by the model's learned coefficients. The primary purpose of the decision boundary is to calculate the probability that an instance belongs to the positive class (class 1). By applying the logistic function (sigmoid) to the linear combination of input features and coefficients, the model transforms that into a probability between 0 and 1. If the probability is greater than or equal to 0.5, the instance is classified as class 1; otherwise, it's classified as class 0. The threshold-based classification forms the essence of Logistic Regression's predictive power, making it a powerful tool for binary classification tasks.

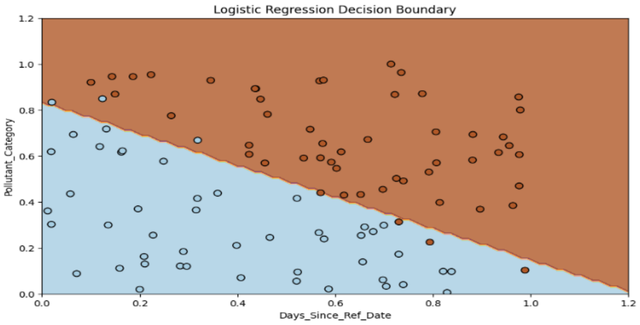


Fig.4 Logistic Regression

5. Result and discussion:

In that comparison of classification models, evaluate the accuracy of two prominent algorithms: Support Vector Classifier (SVC) and Logistic Regression.

Support Vector Classifier (SVC) is a supervised machine learning technique designed for classification tasks. It operates by identifying the optimal hyperplane that effectively separates different classes within the feature space. The Support Vector Classifier (SVC) demonstrates a commendable performance with an overall accuracy of 90%. When predicting positive outcomes, it achieves a precision of approximately 89%, indicating its accuracy in correctly identifying true positives. Moreover, the SVC shows a robust recall of about 94%, successfully capturing a high percentage of actual positive cases. Its balanced performance between precision and recall results in an F1 Score of around 92%.

Logistic Regression is a widely used classification algorithm suitable for both binary and multi-class classification scenarios. It models the probability of a binary target variable, making it particularly useful in such contexts. On the other hand, the Logistic Regression model exhibits a slightly superior accuracy at approximately 93.3%. Remarkably, Logistic Regression achieves perfect precision, ensuring that all positive predictions made by the model are accurate. However, its recall rate of around 87.5% signifies that it might miss identifying some actual positive cases. The trade-off between precision and recall is reflected in the model's F1 Score, which aligns closely with its accuracy at about 93.3%.

Model	Accuracy	Precision	Recall	F1 Score
SVC	0.90	0.8947	0.944	0.9189
Logistic Regression	0.933	1.0	0.875	0.9333

Table. Model’s Accuracy

In the current research work paper ships and related pollutants has been analysed. It has been observed the trend in pollutant levels over the years. It was observed that in 2004 and 2005 pollutant levels were high, but they decreased in 2006. From 2008 to 2016, there were fluctuations, and then, from 2018 to 2020, pollutants consistently remained high. Applying the logistic regression algorithm seems promising for future predictions.

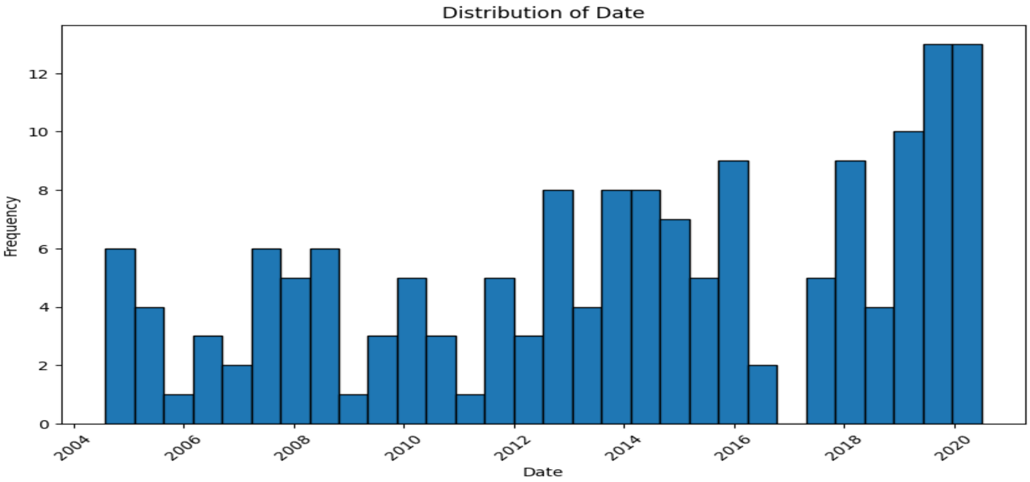


Fig.4 Distribution of Date

6. Conclusion

The comprehensive analysis has shed light on the intricate challenges and multifaceted aspects of marine pollution, particularly focusing on oil spills and their devastating ecological and economic consequences. Oil spills, whether resulting from offshore drilling accidents, transportation mishaps, or natural seepage, pose severe threats to marine ecosystems and societies. Their impact encompasses a wide range of factors, including spill volume, oil type, environmental conditions, and response effectiveness. Ecologically, these incidents harm marine life through toxic effects, habitat destruction, and interference with reproductive cycles, affecting diverse species from birds to marine mammals and fish.

The comparison of classification models, Support Vector Classifier (SVC) and Logistic Regression, reveals nuanced performance distinctions. The SVC demonstrates a commendable overall performance, balancing accuracy, precision, and recall, as reflected in its high F1 Score. Conversely, Logistic Regression exhibits a slightly superior accuracy and perfect precision but at the cost of a lower recall. The choice between these models depends on the specific priorities of a given context, emphasizing the trade-offs between minimizing false positives or false negatives. Shifting the focus to pollutant level analysis, the observed fluctuations and sustained high levels from 2018 to 2020 underscore an ongoing environmental challenge. Recommending the application of the logistic regression algorithm for future predictions suggests its potential efficacy in addressing and forecasting environmental concerns. This study provides valuable insights for decision-makers in both classification tasks and environmental monitoring efforts.

References

1. Mehta N, Shah P, Gajjar P. Oil spill detection over ocean surface using deep learning: a comparative study. Marine Systems & Ocean Technology. 2021 Dec; 16:213-20.

2. Hasimoto-Beltran R, Canul-Ku M, Méndez GM, Ocampo-Torres FJ, Esquivel-Trava B. Ocean oil spill detection from SAR images based on multi-channel deep learning semantic segmentation. Marine Pollution Bulletin. 2023 Mar 1;188:114651

3. Feinauer DM, Latif G, Alenazy AM, Tayem N, Alghazo J, Alzubaidi L. Oil Spill Identification using Deep Convolutional Neural Networks. In2022 14th International Conference on Computational Intelligence and Communication Networks (CICN) 2022 Dec 4 (pp. 240-245). IEEE.

4. Kumar L, Afzal MS, Ahmad A. Prediction of water turbidity in a marine environment using machine learning: A case study of Hong

5. Kong. Regional Studies in Marine Science. 2022 May 1; 52:102260.

6. Dhavalikar AS, Choudhari PC. Detection and Quantification of Daily Marine Oil Pollution Using Remote Sensing. Water, Air, & Soil Pollution. 2022 Aug;233(8):336.

7. Hayes MO, Michel J, Montello TM, Aurand DV, Al-Mansi AM, Al-Moamen AH, Sauer TC, Thayer GW. Distribution and weathering of shoreline oil one year after the Gulf War oil spill. Marine Pollution Bulletin. 1993 Jan 1; 27:135-42.

8. He Y, Hou Y, Wang C, Wang S, Wei Y. Removal of tar from waste textiles by molten carbonates pyrolysis in bubbling reactor. Fuel. 2023 Oct 15; 350:128823.

9. Blumer M. Oil pollution of the ocean. InOil on the Sea: Proceedings of a symposium on the scientific and engineering aspects of oil pollution of the sea, sponsored by Massachusetts Institute of Technology and Woods Hole Oceanographic Institution and held at Cambridge, Massachusetts, May 16, 1969, 1969 May (pp. 5-13). Boston, MA: Springer US.

10. Mogollón R, Arellano C, Villegas P, Espinoza-Morriberón D, Tam J. REPSOL oil spill off Central Perú in January 2022: A modeling case study. Marine Pollution Bulletin. 2023 Sep 1; 194:115282.