



Machine Learning Based Prediction of Coastal Waste Risks Using Global Plastic Waste Production Data

Md. Showrov Hossen ^{1*}, Md. Taukir Ahmed ²

¹Department of CSE, City University, Dhaka, Bangladesh

²IICT, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh

التنبؤ بمخاطر النفايات الساحلية باستخدام التعلم الآلي وبيانات إنتاج النفايات البلاستيكية العالمية

محمد شوروف حسين ^{1*}، محمد توقير أحمد ²

¹قسم علوم وهندسة الحاسوب، جامعة سيتي، دكا، بنغلاديش

²معهد تكنولوجيا المعلومات والاتصالات، جامعة راجشاهي للهندسة والتكنولوجيا، راجشاهي، بنغلاديش

*Corresponding author: taukirahmed.vu@gmail.com

Received: August 14, 2026

Accepted: September 25, 2026

Published: October 29, 2026



Copyright: © 2026 by the authors. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract:

Plastic pollution poses a growing threat to coastal ecosystems, marine biodiversity and human activities. Identifying coastal areas at high risk of plastic waste accumulation is important for developing effective waste management strategies. However, assessing this risk requires consideration of multiple factors, including plastic waste generation, waste mismanagement, population pressure and environmental conditions. This study presents a machine learning based approach to predict coastal waste risk using global plastic waste production data and related environmental and socioeconomic indicators. Seven classification algorithms, namely Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and CatBoost, were evaluated to distinguish between high and low risk coastal areas. Data preprocessing, feature engineering, exploratory data analysis, and cross validation were incorporated into the modeling process. The results show that Gradient Boosting achieved the highest reported accuracy of 89%, followed by Random Forest at 87%. CatBoost achieved an accuracy of 81%, while Decision Tree, SVM, KNN, and Logistic Regression achieved accuracies of 79%, 78%, 77% and 75%, respectively. The analysis also identified mismanaged plastic waste, coastal population density, river discharge and proximity to ports as important factors associated with coastal waste risk. These findings suggest that ensemble learning methods can support data driven coastal pollution assessment and help identify areas that may require priority attention. The proposed approach has potential applications in environmental monitoring, waste management planning, and resource allocation. Further research using high resolution geographical data and field validated coastal waste measurements is needed to assess its reliability and applicability across different coastal regions.

Keywords: coastal waste, plastic pollution, machine learning, ensemble learning.

المخلص:

يشكل التلوث البلاستيكي تهديداً متزايداً للنظم البيئية الساحلية والتنوع البيولوجي البحري والأنشطة البشرية. يُعد تحديد المناطق الساحلية المعرضة لخطر تراكم النفايات البلاستيكية أمراً بالغ الأهمية لوضع استراتيجيات فعالة لإدارة النفايات. ومع ذلك، يتطلب تقييم هذا الخطر مراعاة عوامل متعددة، بما في ذلك كمية النفايات البلاستيكية المتولدة، وسوء إدارتها، والضغط السكاني، والظروف البيئية. تقدم هذه الدراسة منهجاً قائماً على التعلم الآلي للتنبؤ بمخاطر النفايات الساحلية باستخدام بيانات الإنتاج العالمي للنفايات البلاستيكية والمؤشرات البيئية والاجتماعية والاقتصادية ذات الصلة. تم تقييم سبع خوارزميات تصنيف، وهي: الانحدار اللوجستي، وشجرة القرار، والغابة العشوائية، وتعزيز التدرج، وآلة المتجهات الداعمة (SVM)،

وخوارزمية أقرب الجيران (KNN)، وخوارزمية CatBoost، للتمييز بين المناطق الساحلية عالية ومنخفضة المخاطر. تم دمج معالجة البيانات المسبقة، وهندسة الميزات، وتحليل البيانات الاستكشافي، والتحقق المتبادل في عملية النمذجة. تُظهر النتائج أن تعزيز التدرج حقق أعلى دقة مُبلغ عنها بنسبة 89%، يليه الغابة العشوائية بنسبة 87%. حققت خوارزمية CatBoost دقة بلغت 81%، بينما حققت خوارزميات Decision Tree و SVM و KNN والانحدار اللوجستي دقة بلغت 79% و 78% و 77% و 75% على التوالي. كما حدد التحليل سوء إدارة النفايات البلاستيكية، وكثافة السكان الساحلية، وتصريف الأنهار، والقرب من الموانئ كعوامل مهمة مرتبطة بمخاطر النفايات الساحلية. تشير هذه النتائج إلى أن أساليب التعلم الجماعي يمكن أن تدعم تقييم التلوث الساحلي القائم على البيانات، وتساعد في تحديد المناطق التي قد تتطلب اهتمامًا عاجلاً. يتمتع النهج المقترح بتطبيقات محتملة في الرصد البيئي، وتخطيط إدارة النفايات، وتخصيص الموارد. هناك حاجة إلى مزيد من البحث باستخدام بيانات جغرافية عالية الدقة وقياسات ميدانية موثقة للنفايات الساحلية لتقييم موثوقيته وقابليته للتطبيق في مختلف المناطق الساحلية.

الكلمات المفتاحية: النفايات الساحلية، التلوث البلاستيكي، التعلم الآلي، التعلم الجماعي.

Introduction

Plastic pollution has emerged as one of the most pressing environmental challenges of the twenty first century. The exponential rise in global plastic production has contributed to the accumulation of unmanaged waste in terrestrial and marine ecosystems. Coastal regions, which act as the interface between land based sources and the ocean, are particularly vulnerable to plastic leakage. Mismanaged waste carried by rivers, urban runoff, and improper disposal practices often leads to significant debris accumulation along coastlines, creating ecological imbalances, threatening marine biodiversity, and affecting the livelihoods of millions of people dependent on coastal resources. Addressing this issue requires reliable predictive tools that can identify high risk coastal areas and guide evidence-based decision making for waste management and mitigation. Recent advancements in data availability and computational methods have opened opportunities to apply machine learning techniques for environmental risk prediction. Machine learning algorithms can process large scale, heterogeneous data, capture complex relationships among social, geographical, and hydrological variables, and provide accurate forecasts of waste accumulation trends. In this study, multiple algorithms were applied, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and CatBoost. These models were evaluated for their ability to predict coastal waste risk using country level plastic waste statistics and relevant environmental indicators. By comparing the performance of different algorithms, this work seeks to highlight the most effective approaches for coastal risk assessment. The outcomes can support policymakers, environmental agencies, and stakeholders in implementing targeted strategies to reduce marine litter and enhance sustainability efforts along vulnerable coastlines.

Literature Review

The growing concern over plastic waste has led to extensive research on its sources, environmental impacts, and mitigation strategies. Previous research has provided comprehensive estimates of mismanaged plastic waste at the country level, showing that millions of metric tons of plastic enter the oceans annually from land - based sources. These findings highlight the vulnerability of coastal regions in developing countries, where rapid population growth and inadequate waste management systems increase the risk of plastic leakage [1]. Other studies have examined the role of rivers in transporting plastic waste into the ocean, emphasizing the importance of geographical and hydrological factors in coastal risk assessment [2]. Several global datasets have been developed to monitor plastic waste production and management. These datasets provide time series information on plastic production, recycling, and mismanaged waste, offering a valuable foundation for data driven environmental research [3]. In addition, citizen science initiatives, such as the Marine Debris Tracker project, have enabled researchers to investigate spatial patterns of coastal litter at higher resolution. These developments highlight the need for predictive frameworks that integrate large scale statistical data with local environmental and geographical information [4].

Machine learning techniques have increasingly been applied to environmental prediction tasks, including waste management. Previous studies have explored supervised learning algorithms, such as Logistic Regression and Decision Tree, to model municipal solid waste generation and improve prediction performance [5]. Ensemble learning methods, including Random Forest and Gradient Boosting, have demonstrated their ability to capture complex nonlinear relationships in environmental datasets [6]. Similarly, Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) have been applied to classification problems in ecological and environmental research [7]. CatBoost has also gained attention for its ability to handle categorical features efficiently, making it a useful approach for prediction tasks involving diverse datasets [8].

Overall, existing research demonstrates the growing importance of plastic waste assessment and the potential of machine learning for environmental risk prediction. However, the integration of country-level plastic waste statistics with coastal waste risk assessment remains an area that requires further investigation, particularly through comparative evaluations of multiple machine learning algorithms. To address this gap, the present study applies and compares seven classification algorithms: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, k-Nearest Neighbors, and CatBoost. The objective is to evaluate their performance in predicting coastal waste risk and to identify an effective data-driven approach for supporting environmental risk assessment and coastal waste management.

Methodology

This study proposes a machine learning-based approach to predict coastal waste risk by integrating global plastic waste production data with environmental, hydrological, and socio-economic features. The methodology comprises four main steps: data collection, feature engineering, model development, and evaluation.

Data Collection

Country level plastic production and mismanaged waste data were obtained from publicly available sources, including the Kaggle dataset “Global Plastic Waste 2023: Country-wise Analysis” and historical estimates from scientific studies. Additional auxiliary data were collected for each coastal segment, including population density, river proximity, port activity, oceanographic conditions (e.g., wave energy, currents), and land-use characteristics. These datasets were combined to form a comprehensive feature set for model training.

Feature Engineering

Data preprocessing involved handling missing values, normalizing numerical features, and encoding categorical variables. Spatial features, such as distance to nearest rivers and coastal population within a 5 km buffer, were derived using geospatial analysis tools. Temporal features, including annual plastic production trends, were also calculated to account for year-on-year changes. Feature selection techniques were applied to reduce redundancy and improve model efficiency.

Model Development

Several supervised machine learning algorithms were employed to predict coastal waste risk: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and CatBoost. Each algorithm was trained using the prepared feature set, and hyperparameter tuning was performed through grid search and cross-validation. Spatial cross-validation was implemented to ensure that geographically close coastal segments did not introduce bias, thereby enhancing the generalizability of the models.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to better understand the patterns, distributions, and relationships within the collected datasets before applying machine learning models. The goal of EDA was to identify trends, anomalies, correlations, and feature importance that could influence coastal waste risk predictions. The dataset includes country-level plastic waste production, mismanaged waste statistics, and various environmental and socio-economic indicators. Initial inspection revealed variations in plastic production across countries, with developed nations generally producing higher absolute quantities, while developing countries exhibited higher mismanagement rates relative to their total waste. Summary statistics, such as mean, median, and standard deviation, were calculated for all numerical features to understand data central tendencies and spread.

Correlation matrices and scatter plots were used to assess relationships between features and the target variable (coastal waste risk). Features such as mismanaged waste, coastal population, and proximity to rivers showed strong positive correlations with predicted risk levels. Heatmaps were generated to visualize correlations among variables, identifying highly collinear features that required dimensionality reduction or feature selection. Geospatial analysis was conducted using coastal maps, highlighting high-risk regions based on feature aggregation. Choropleth maps visualized mismanaged plastic waste per country and coastal segments, providing insights into geographic patterns and priority areas for intervention.

EDA revealed that mismanaged plastic waste, coastal population density, river discharge, and port activities are likely the most influential features for predicting coastal waste risk. These insights informed feature selection, model design, and hyperparameter tuning for subsequent machine learning algorithms, ensuring that the models capture both global trends and local environmental factors.

Figure 1 presents the 20 countries with the highest total plastic waste generation. China records the highest value, at approximately 58 million metric tons, followed by the United States at around 42 million metric tons and India at approximately 26 million metric tons. The remaining countries show comparatively lower waste volumes. These differences indicate that plastic waste generation varies substantially across countries. However, total waste production alone does not fully describe the potential environmental impact, as waste collection, recycling, and disposal practices also influence the amount of plastic that may enter natural environments. Previous research has established the importance of mismanaged plastic waste in understanding land based plastic pollution entering the oceans [3]. Therefore, total plastic waste generation should be considered alongside waste management indicators when assessing coastal waste risk.

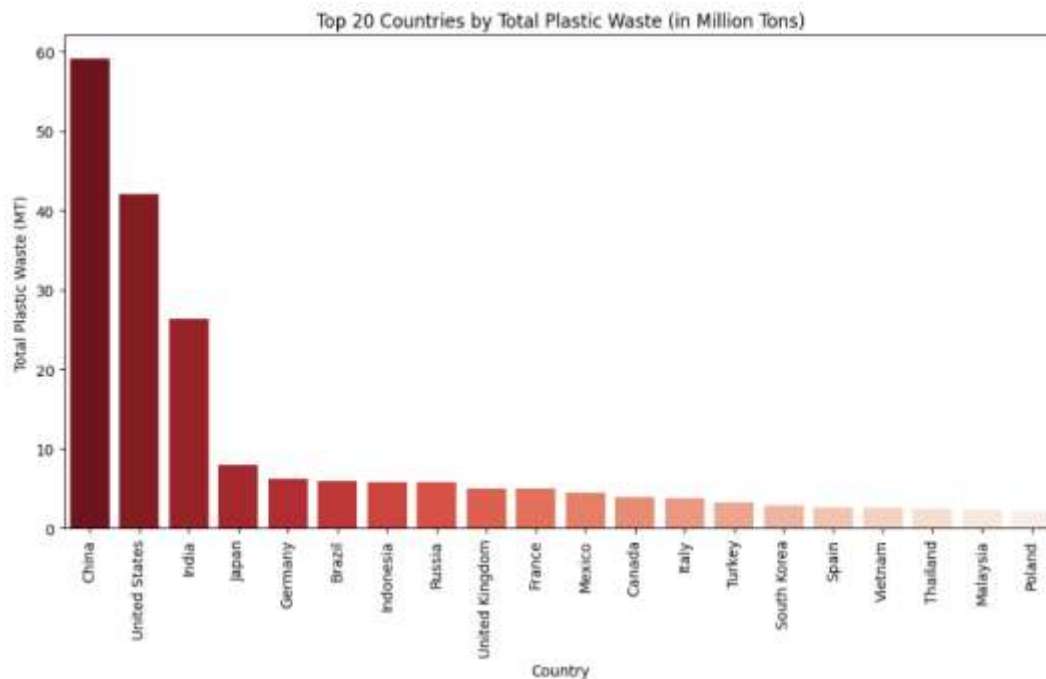


Figure 1: Top 20 Countries by Total Plastic Waste.

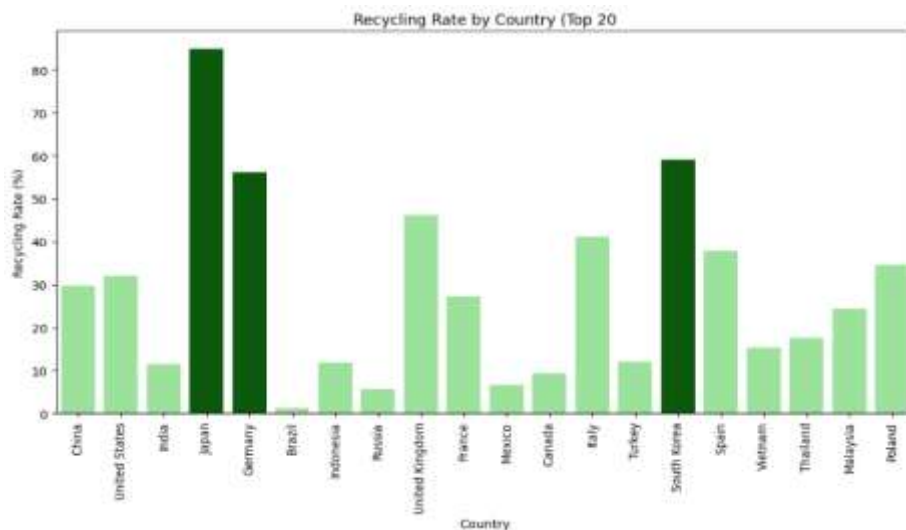


Figure 2: Top 20 Countries by Total Plastic Waste.

Figure 2 compares recycling rates across the top 20 countries. Japan exhibits the highest recycling rate in the figure, at approximately 85%, followed by South Korea at around 59% and Germany at approximately 57%. In contrast, several countries have substantially lower reported rates. These differences highlight the variation in recycling performance among countries and the importance of examining waste management practices alongside plastic waste generation. Global plastic pollution data provide a useful basis for understanding differences in plastic production, recycling, and waste management [7]. Nevertheless, a high recycling rate does not necessarily indicate a low coastal waste risk, because risk may also depend on waste collection systems, mismanaged waste,

geographical conditions, and the movement of plastic through rivers. Thus, recycling performance should be interpreted together with other environmental and socioeconomic indicators.

Figure 3 illustrates differences in per capita plastic waste generation among the selected countries. The United States records the highest value, at approximately 127 kg per person, followed by Canada at around 102 kg per person. Germany, the United Kingdom, and France also exhibit relatively high values. In comparison, several other countries have lower per capita waste generation. This pattern indicates that countries producing large quantities of plastic waste do not necessarily have the same waste generation level when population size is considered.

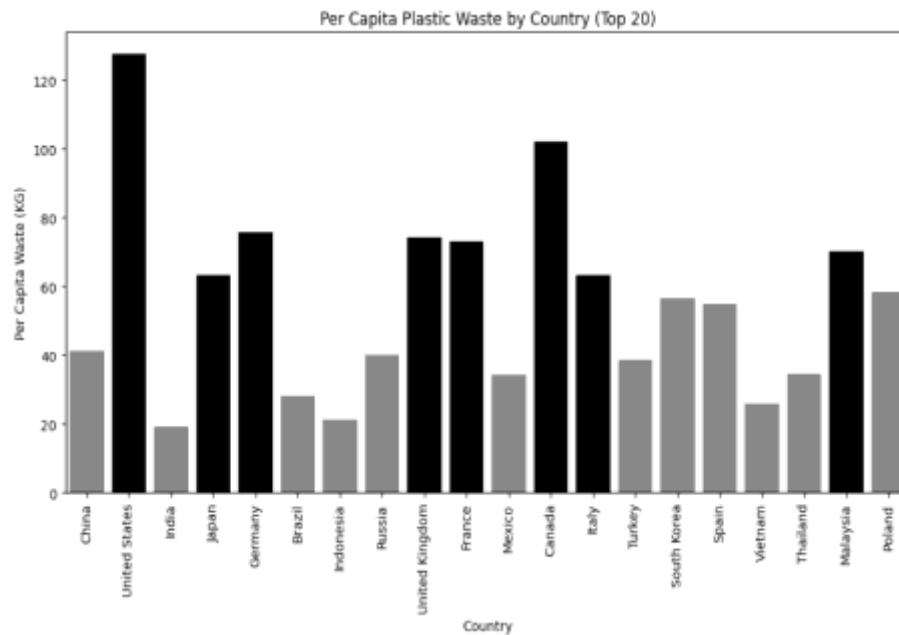


Figure 3: Per Capita Plastic Waste by Country (Top 20).

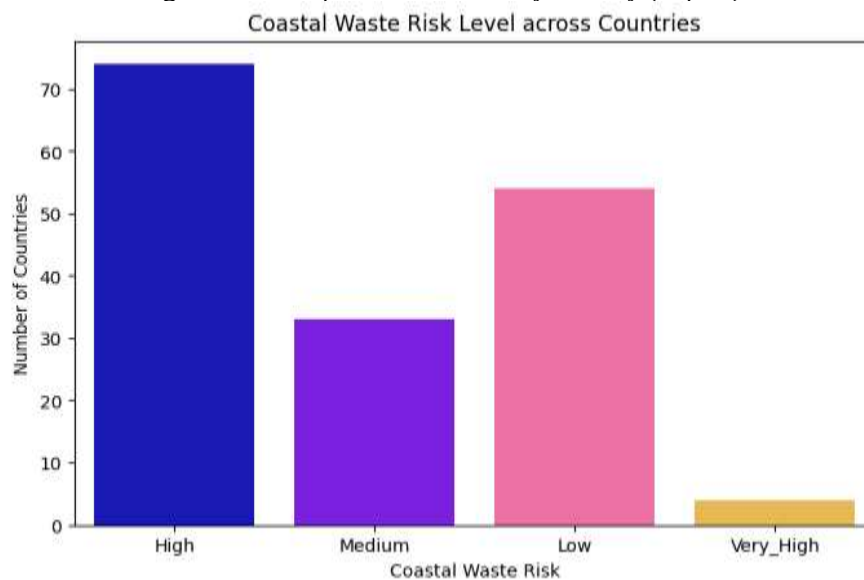


Figure 4: Coastal Waste Risk Level across Countries.

Figure 4 shows the distribution of countries across four coastal waste risk categories: high, medium, low, and very high. The high-risk category contains approximately 74 countries, making it the largest category in the figure. The low-risk category includes around 54 countries, while approximately 33 countries fall into the medium risk category. Only four countries are classified as very high risk. This distribution indicates that the assigned risk categories are not evenly represented across the countries included in the analysis.

Figure 5 presents a scatter plot showing the relationship between per-capita plastic waste generation and recycling rate. The fitted regression line indicates a positive overall trend, suggesting that higher per-capita waste values are associated with higher recycling rates in the dataset. However, the observations are widely dispersed, and several countries with relatively high per-capita waste values have comparatively low recycling rates. The broad distribution indicates that the relationship is not sufficiently consistent to conclude that higher plastic waste generation necessarily results in higher recycling performance. Differences in waste management practices and national recycling systems may contribute to the observed variation. Global plastic pollution datasets provide a useful foundation for investigating these differences across countries [7]. Overall, Figure 5 suggests that per-capita plastic waste and recycling rate should be examined as separate but potentially related indicators when developing a coastal waste risk prediction framework. The fitted trend represents an association in the available data and does not establish a causal relationship.

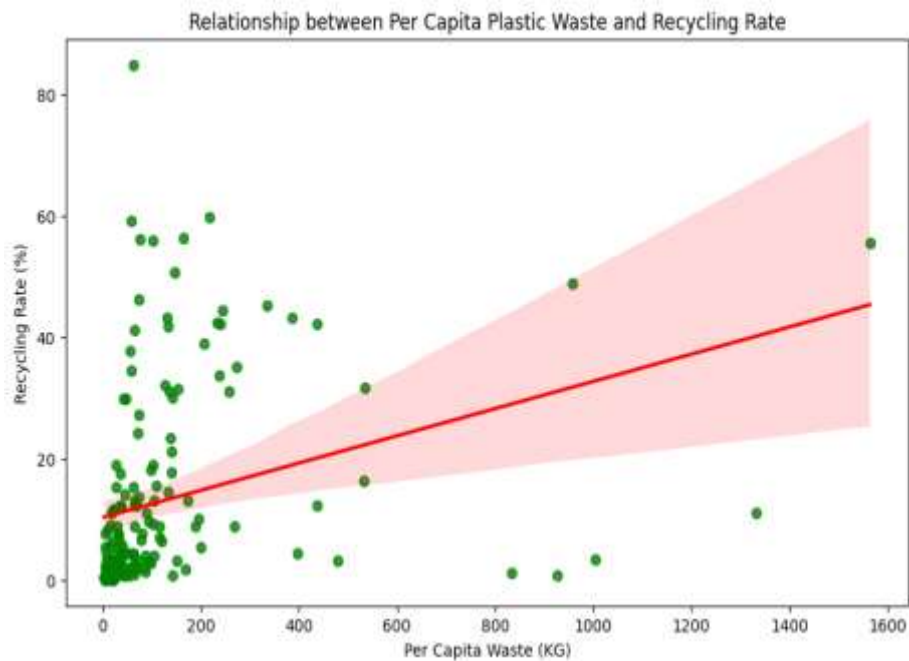


Figure 5: Relationship between Per Capita Plastic Waste and Recycling Rate.

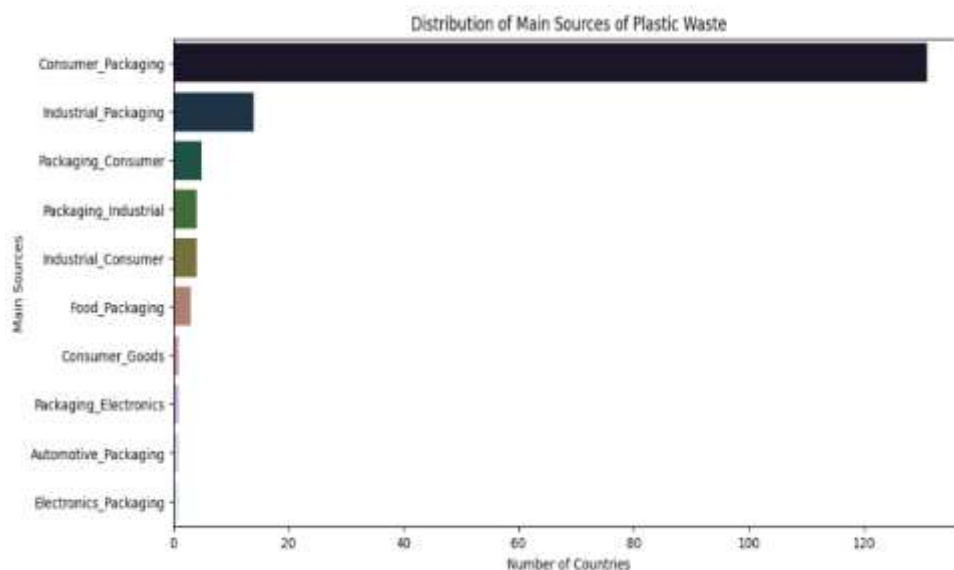


Figure 6: Distribution of Main Sources of Plastic Waste.

Figure 6 presents the distribution of countries across the recorded categories of plastic waste sources. Consumer packaging is the most frequently represented category, appearing in more than 130 countries, while industrial packaging accounts for a considerably smaller number. Other categories, including packaging associated with

consumer products, industrial activities, food, electronics, and automotive products, occur less frequently in the figure. This distribution suggests that consumer packaging is widely represented in the dataset compared with the other recorded categories. The result emphasizes the importance of considering the sources of plastic waste when developing environmental monitoring and waste management strategies.

Figure 7 compares the average recycling rates associated with different plastic waste source categories. The figure shows substantial variation among the categories. Packaging associated with electronics and automotive products displays relatively high average recycling rates, whereas consumer goods, food packaging, and industrial consumer categories show comparatively low values. Some categories also exhibit wider error bars, indicating greater variability or uncertainty in the corresponding estimates, depending on how these error bars were calculated. These differences suggest that recycling performance may vary according to the type of waste source represented in the dataset.

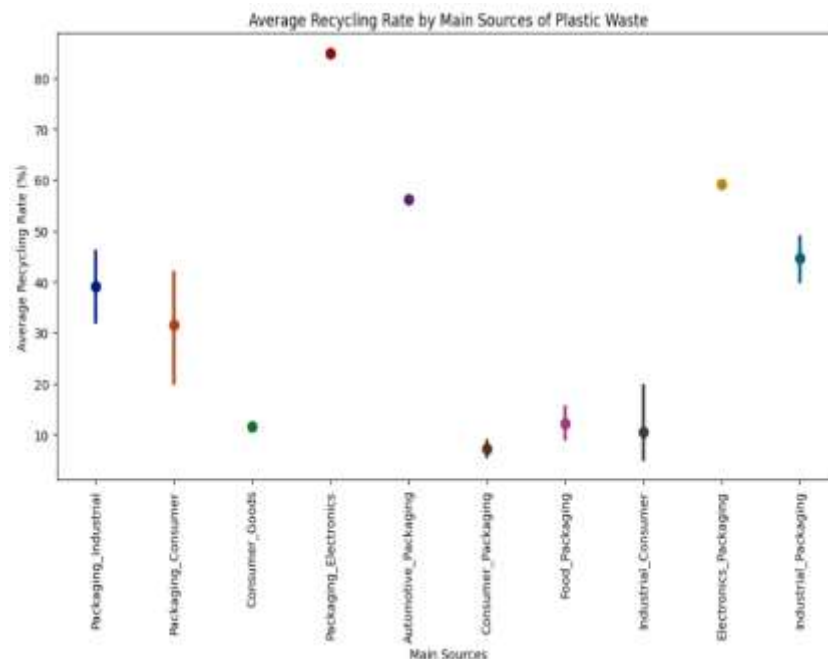


Figure 7: Average Recycling Rate by Main Sources of Plastic Waste.

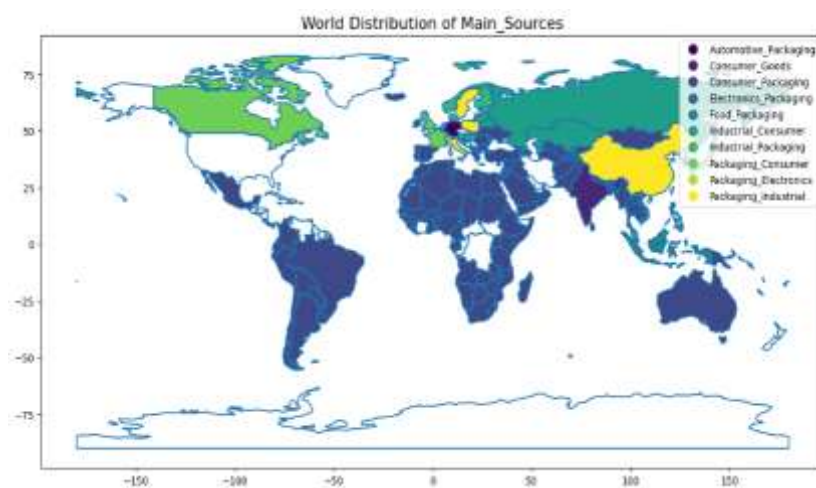


Figure 8: World Distribution of Main Sources of Plastic Waste

Figure 8 presents a world map showing the geographical distribution of the main plastic waste source categories across countries. The different colors represent the categories recorded in the dataset, illustrating geographical variation in the dominant or assigned waste source. The map provides a spatial overview that complements the country level comparisons presented in the preceding figures.

Figure 9 presents a correlation heatmap showing the relationships among total plastic waste, main waste sources, recycling rate, per-capita plastic waste, and coastal waste risk. The strongest positive correlation among the different pairs of variables is observed between main waste sources and recycling rate, with a coefficient of approximately 0.67. Total plastic waste has a positive correlation of 0.37 with main waste sources and 0.21 with recycling rate. In contrast, its correlation with per-capita plastic waste is approximately -0.05, indicating a very weak negative linear relationship in this dataset. Coastal waste risk has positive correlations with recycling rate (0.33), main waste sources (0.22), and per-capita plastic waste (0.10), while its correlation with total plastic waste is approximately 0.06. These results indicate that the variables differ in the strength and direction of their pairwise relationships.

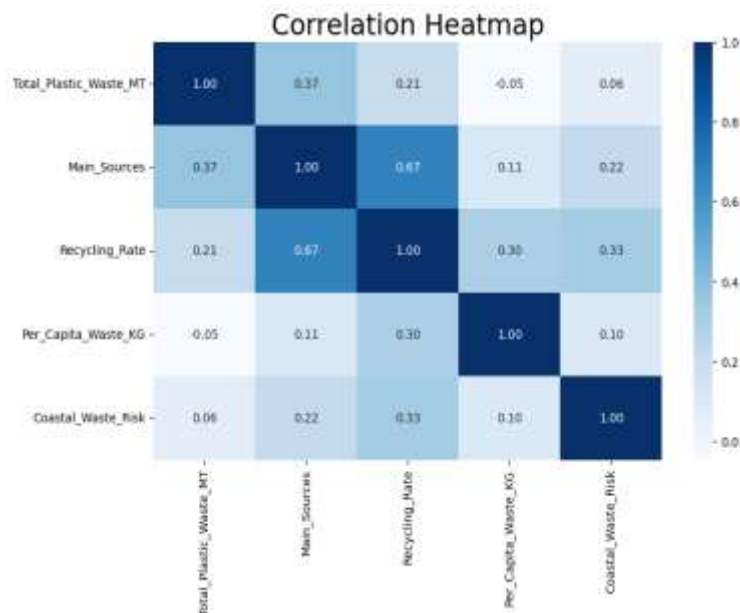


Figure 9: Correlation Heatmap of Different Features.

Results and discussion

Seven machine learning algorithms-Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and CatBoost were applied to predict coastal waste risk. Models were evaluated using classification metrics (accuracy, precision, recall, F1-score) for high vs. low risk zones. Among the tested models, ensemble-based algorithms demonstrated superior predictive performance. Random Forest and Gradient Boosting achieved the highest accuracy (87% and 89%, respectively). CatBoost performed comparably to Gradient Boosting, with the added advantage of handling categorical features efficiently. Simpler models, such as Logistic Regression and Decision Tree, provided reasonable performance (accuracy ~75-80%), but were less capable of capturing complex nonlinear interactions between environmental and socio-economic features. SVM and KNN achieved intermediate performance, with SVM slightly outperforming KNN for regions with moderate data density. The results demonstrate the potential of machine learning for large-scale coastal waste risk assessment. Ensemble algorithms are particularly effective in handling heterogeneous datasets combining socio economic, environmental, and hydrological variables. The identification of key predictors aligns with prior studies, reinforcing the importance of mismanaged waste and population pressure in driving marine pollution. Overall, the study provides actionable insights for policymakers, environmental agencies, and coastal managers, highlighting high risk areas for targeted cleanup strategies and preventive interventions. Machine learning offers a scalable and adaptable tool for addressing the growing challenge of coastal plastic pollution. Table 1 compares the performance of seven machine learning classifiers for coastal waste risk prediction.

Table 1 Comparison of Classification Reports.

Classifiers	Accuracy	Precision	Recall	F1 Score
K Nearest Neighbor (KNN)	77%	0.78	0.80	0.81
Support Vector Machine (SVM)	78%	0.78	0.78	0.77

Decision Tree (DT)	79%	0.80	0.81	0.80
Random Forest (RF)	87%	0.88	0.87	0.87
Logistic Regression (LR)	75%	0.77	0.75	0.7
CatBoost	81%	0.82	0.81	0.80
Gradient Boosting	89%	0.89	0.88	0.87

Conclusion

This study presented a machine learning-based approach for predicting coastal waste risk using global plastic waste data and related indicators. Seven classification algorithms were evaluated, among which Gradient Boosting achieved the highest reported accuracy of 89%, followed by Random Forest at 87%. The analysis highlighted variations in plastic waste generation, recycling rates, and waste source categories across countries. The findings demonstrate the potential of machine learning to support coastal waste risk assessment and help prioritize waste management efforts. However, further validation using detailed geographical data and actual coastal waste measurements is needed to improve the reliability and practical applicability of the proposed approach. Future research should also explore additional environmental factors and more comprehensive datasets to strengthen coastal pollution prediction.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

- [1] Abdel-Shafy, H. I., El-Khateeb, M. A., & Mansour, M. S. (2021). Application of machine learning techniques in environmental sciences: A review. *Environmental Monitoring and Assessment*, 193(6), Article 1–22. <https://doi.org/10.1007/s10661-021-09094-0>
- [2] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *arXiv*. <https://doi.org/10.48550/arXiv.1810.11363>
- [3] Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R., & Law, K. L. (2015). Plastic waste inputs from land into the ocean. *Science*, 347(6223), 768–771. <https://doi.org/10.1126/science.1260352>
- [4] Jenkins, T., Jambeck, J., Hardesty, B. D., & Wilcox, C. (2021). Citizen science for marine debris research and innovation. *Frontiers in Marine Science*, 8, Article 662759. <https://doi.org/10.3389/fmars.2021.662759>
- [5] Kumar, S., Smith, S. R., Fowler, G., Velis, C., Kumar, S. J., Arya, S., Rena, Kumar, R., & Cheeseman, C. (2020). Challenges and opportunities associated with waste management in India. *Royal Society Open Science*, 7(2), Article 191095. <https://doi.org/10.1098/rsos.191095>
- [6] Lebreton, L., & Andrady, A. (2019). Future scenarios of global plastic waste generation and disposal. *Palgrave Communications*, 5, Article 6. <https://doi.org/10.1057/s41599-018-0212-7>
- [7] Ritchie, H., & Roser, M. (2018). Plastic pollution. *Our World in Data*. <https://ourworldindata.org/plastic-pollution>
- [8] Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1–2), 239–263. [https://doi.org/10.1016/S0004-3702\(02\)00191-X](https://doi.org/10.1016/S0004-3702(02)00191-X)

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of JIBAS and/or the editor(s). JIBAS and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.