ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.981

Comparison of Machine Learning Models for Predicting Environmental Risk Associated with Coastal Waste at a Global Level

Richard Fernando Fernández Vásquez, Doctor[1]
[1]Universidad Nacional de Ingeniería, Perú, *rfernandezv@uni.edu.pe*

Abstract*– Machine learning models are key tools in coastal environmental risk management, as they allow for the identification of critical areas, optimize waste management, and support the development of more effective and sustainable conservation policies globally. This research aimed to compare machine learning models for predicting the environmental risk associated with coastal waste globally, with the goal of identifying the most suitable model and guiding the formulation of coastal conservation policies. The research used a database of 165 countries with varying levels of environmental risk associated with coastal waste. The data were divided into a training sample (80%) and a validation sample (20%). The performance of five Machine Learning models —Random Forest, Gradient Boosting, XGBoost, LightGBM and CatBoost— was evaluated in predicting the probability of environmental risk associated with coastal waste at a global level, with the Random Forest model showing the best performance, with an accuracy of 0.5455, recall of 0.8000, F1-score of 0.6486, area under the ROC curve of 0.6852 and Gini index of 0.3704, demonstrating the greatest capacity for discrimination and predictive accuracy.

Keywords*-- Machine learning, environmental risk, coastal waste, coastal conservation, global sustainability.

Comparación de Modelos de Machine Learning para la Predicción del Riesgo Ambiental Asociado a Residuos Costeros a Nivel Global

Richard Fernando Fernández Vásquez, Doctor[1]
[1]Universidad Nacional de Ingeniería, Perú, *rfernandezv@uni.edu.pe*

Resumen– ***Los modelos de Machine Learning son herramientas clave en la gestión del riesgo ambiental costero, ya que permiten identificar áreas críticas, optimizar la gestión de residuos y respaldar la formulación de políticas de conservación más eficaces y sostenibles a nivel global. Esta investigación tuvo como objetivo comparar modelos de Machine Learning para la predicción del riesgo ambiental asociado a residuos costeros a nivel global, con el propósito de identificar el modelo más adecuado y orientar la formulación de políticas de conservación costera. La investigación utilizó una base de datos compuesta por 165 países con distintos niveles de riesgo ambiental asociado a residuos costeros. Los datos se dividieron en una muestra de entrenamiento (80%) y una muestra de validación (20%). Se evaluó el desempeño de cinco modelos de Machine Learning —Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost— en la predicción de la probabilidad de riesgo ambiental asociado a los residuos costeros a nivel global, siendo el modelo Random Forest con mejor desempeño, con una precisión de 0.5455, recall de 0.8000, F1-score de 0.6486, área bajo la curva ROC de 0.6852 e índice de Gini de 0.3704, evidenciando la mayor capacidad de discriminación y precisión predictiva.***

Palabras clave-- Machine Learning, riesgo ambiental, residuos costeros, conservación costera, sostenibilidad global.

I. Introducción

La contaminación por residuos costeros representa una amenaza creciente para los ecosistemas marinos, la salud pública y las economías locales, tal es así que la Agencia Europea del Medio Ambiente [1], menciona que la acumulación progresiva de desechos plásticos en los océanos constituye una amenaza creciente para la integridad de los ecosistemas marinos, ocasionando daños directos a la fauna debido a la ingestión de plásticos. Asimismo, la salud humana podría verse comprometida, dado que estos materiales se fragmentan en microplásticos capaces de incorporarse a la cadena alimentaria. Estos efectos reflejan solo una parte de las múltiples consecuencias ambientales y sanitarias derivadas de la acumulación de residuos en los ambientes marinos.

Para la Organización de las Naciones Unidas [2], la contaminación por plásticos en los ecosistemas acuáticos ha experimentado un incremento sostenido en los últimos años y se estima que podría duplicarse hacia 2030, representando una amenaza significativa para la salud humana, la economía, la biodiversidad y el equilibrio climático. Se menciona que de los 9200 millones de toneladas de producción de plástico producidas entre 1950 y 2017, alrededor de 7000 millones se han transformado en residuos plásticos, considerándola como una crisis mundial y expresando la urgencia de reducir la producción mundial de plástico y de residuos plásticos en el medio ambiente. Concluye que las estrategias de reciclaje vigentes son insuficientes para abordar de manera efectiva la magnitud del problema.

Por otro lado, el Banco Mundial [3], menciona que la contaminación marina, derivada de la presencia de plásticos y otros subproductos, constituye una amenaza significativa para actividades económicas clave como la pesca y el turismo, de las cuales dependen numerosas economías caribeñas. En esta región, el turismo aporta aproximadamente el 15.2% del producto interno bruto y supera el 50% en países como Granada, lo que evidencia su alta vulnerabilidad frente a los impactos ambientales.

Contreras [4], destaca un informe de la organización Algalita Marine Research and Education, en cual indica que existe una isla de plástico de aproximadamente 2.6 millones de kilómetros cuadrados frente a las costas de Perú y Chile. Según el informe, la insuficiente gestión de residuos representa un riesgo tanto para la integridad de las especies que habitan los ecosistemas litorales como para la salud de la población humana.

Estos residuos, mayoritariamente plásticos, provienen de múltiples fuentes, tanto terrestres como marinas, y su gestión eficaz requiere estrategias de evaluación de riesgo que permitan priorizar zonas y acciones de intervención. En este contexto, los modelos de Machine Learning ofrecen herramientas para analizar datos ambientales, socioeconómicos y relacionados a los residuos costeros. Por ejemplo, Abbasi et al. [5] aplicaron modelos de Machine Learning para predecir la generación de residuos municipales en Teherán, demostrando la precisión de estos algoritmos en este ámbito.

Fang et al. [6], señalaron que el aumento constante de residuos a nivel mundial está causando problemas relacionados con la contaminación, la gestión y el reciclaje. Esto demanda el desarrollo de nuevas estrategias, como el desarrollo de modelos de Machine Learning, que permiten

clasificar residuos con una precisión entre el 72.8% y 99.95%, lo que contribuye en mejorar de gestión de residuos.

Para Pourzangbar et al. [7], los modelos de Machine Learning, cuando son implementados y optimizados adecuadamente, contribuyen significativamente a la sostenibilidad de los ecosistemas marinos y costeros.

Edeh et al. [8] afirmaron que la contaminación marina por plásticos altera los ecosistemas acuáticos y constituye una seria amenaza para la salud a largo plazo de los océanos, lo que representa una creciente preocupación global. En este contexto, los autores aplicaron modelos de Machine Learning para cuantificar la contaminación marina y submarina, así como para identificar y clasificar residuos plásticos. Con ello desarrollaron y evaluaron un sistema automatizado de detección de plásticos flotantes en el océano, alcanzando una precisión del 98.38% en la identificación, lo que demuestra el potencial de estas técnicas para la mitigación de la contaminación marina.

Binetti et al. [9] señalaron que en la monitorización ambiental se emplean modelos de Machine Learning, siendo el modelo de Random Forest uno de los más utilizados debido a su capacidad para abordar tareas de clasificación y regresión. Este modelo presenta una notable resistencia al sobreajuste, derivada de la generación de múltiples árboles de decisión a partir de subconjuntos aleatorios de datos. Además, su simplicidad favorece la implementación y permite la integración de diferentes conjuntos de datos, mejorando así la eficacia del modelo en la resolución de problemas complejos y aumentando su precisión predictiva.

Sin embargo, existe una limitada evaluación comparativa entre modelos de Machine Learning para la predicción del riesgo ambiental asociado a residuos costeros a nivel global. Aunque existen esfuerzos internacionales para cuantificar y mitigar la contaminación costera, no se dispone de modelos predictivos robustos que integren variables ambientales, sociales y económicas. Dichos modelos son fundamentales como herramientas de apoyo a la toma de decisiones para los gestores ambientales, ya que permiten mejorar la gestión costera al anticipar zonas de alto riesgo y priorizar acciones de mitigación y políticas de conservación.

Para abordar estas limitaciones, el presente estudio se fundamenta en los principios de Machine Learning para comparar los modelos Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost como herramientas en la gestión del riesgo ambiental asociado a residuos costeros a nivel global. La selección del modelo con mayor capacidad de discriminación permite estimar con mayor precisión la probabilidad de riesgo ambiental, contribuyendo al desarrollo de un marco predictivo más robusto y confiable. Desde una perspectiva práctica, esta investigación resulta relevante para la gestión del riesgo ambiental costero, ya que proporciona una herramienta que puede apoyar a los responsables de la formulación de políticas en el diseño de estrategias eficaces de conservación y mitigación. Además, el enfoque metodológico adoptado se respalda en estudios recientes, como el de Binetti et al. [2024], quienes destacan la aplicación de modelos de Machine Learning en la monitorización ambiental.

Por las justificaciones mostradas, el objetivo general de la presente investigación es comparar modelos de Machine Learning para la predicción del riesgo ambiental asociado a residuos costeros a nivel global, con el propósito de identificar el modelo más adecuado y orientar la formulación de políticas de conservación costera basadas en niveles de riesgo ambiental. Asimismo, se busca evaluar el desempeño de los modelos de Machine Learning Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost en la predicción de la probabilidad de riesgo ambiental asociado a residuos costeros a nivel global haciendo uso de los indicadores estadísticos de precisión, recall, F1-score, el valor del área bajo la curva ROC y el índice de Gini para determinar su capacidad predictiva utilizando un conjunto de datos proveniente de fuentes públicas y monitoreos costeros. Finalmente, analizar los factores de mayor influencia en el riesgo ambiental, identificando las variables más significativas para la predicción de la probabilidad de riesgo asociado a los residuos costeros a nivel global.

El resto del artículo se organiza de la siguiente manera: la sección 2 presenta los materiales y métodos empleados, la sección 3 presenta los resultados y se discute los resultados, y en la sección 4 se presenta las conclusiones.

II. Materiales y Métodos

A. Datos

Para la predicción del riesgo ambiental asociado a residuos costeros a nivel global, el conjunto de datos utilizado en este estudio fue obtenido de la plataforma de datos abiertos de Kaggle [10]. Este recurso proporciona información detallada sobre la producción y gestión residuos plásticos a nivel global, abarcando 165 países en el año 2023. Los datos incluyen diversos aspectos de la gestión de residuos plásticos, desde los volúmenes de producción hasta la eficiencia del reciclaje y la evaluación de riesgos ambientales.

Las variables independientes fueron:

- Country: es el país analizado.
- Total_Plastic_Waste_MT: es la producción de residuos plásticos por país en millones de toneladas métricas.
- Main_Sources: son las principales fuentes de residuos plásticos por país.
- Recycling_Rate: es la tasa nacional de reciclaje (%).
- Per_Capita_Waste_KG: es la producción de residuos per cápita (kg/persona).

La variable dependiente fue:

- Risk: es la evaluación de riesgo ambiental de residuos costeros. Toma dos valores, high, si el riesgo ambiental es alto y not high, si el riesgo ambiental no es alto.

B. Preprocesamiento de Datos

El preprocesamiento incluyó la identificación y tratamiento de valores faltantes; dado que no se encontraron datos ausentes, se procedió con la codificación de las variables

independientes cualitativas mediante la creación de variables dummy.

C. *Análisis Estadístico*

Como parte del análisis estadístico, se realizó un análisis univariado. Adicionalmente, en el análisis bivariado se aplicaron las pruebas de Shapiro-Wilk, Chi-cuadrado y Wilcoxon o Mann-Whitney para evaluar los supuestos de normalidad y determinar diferencias significativas entre las variables analizadas. A continuación, se presentan sus principales características y fundamentos teóricos.

i) Prueba de Shapiro-Wilk

Se efectuó un análisis univariado de las variables independientes cuantitativas con el propósito de evaluar su ajuste a una distribución normal. De acuerdo con Royston [11], la prueba de Shapiro-Wilk constituye un método estadístico eficaz para verificar la normalidad de una variable cuantitativa, formulando las siguientes hipótesis:

- Ho: La variable se aproxima a una distribución normal
- Ha: La variable no se aproxima a una distribución normal

Si el pvalor es menor a un nivel de significancia de 5%, se rechaza la Ho, por lo tanto, la variable no se aproxima a una distribución normal.

ii) Prueba Wilcoxon o Mann-Whitney

Para el análisis bivariado de los datos y la identificación de relaciones significativas entre la variable dependiente y las variables independientes cuantitativas que no se ajustan a una distribución normal, se empleó la prueba estadística de Wilcoxon o Mann-Whitney, descrita por Hollander et al. [12], la cual establece las siguientes hipótesis:

- Ho: No existe diferencia entre las distribuciones de los dos grupos
- Ha: Existe diferencia entre las distribuciones de los dos grupos

Si el pvalor es menor a un nivel de significancia de 5%, se rechaza la Ho, por lo tanto, existe diferencia entre las distribuciones de los dos grupos.

iii) Prueba Chi-cuadrado

Para el análisis bivariado de los datos y la identificación de relaciones significativas entre la variable dependiente y las variables independientes cualitativas, se empleó la prueba estadística Chi-cuadrado, propuesta en Agresti [13], la cual establece las siguientes hipótesis:

- Ho: No existe dependencia entre las variables
- Ha: Existe dependencia entre las variables

Si el pvalor es menor a un nivel de significancia de 5%, se rechaza la Ho, por lo tanto, existe dependencia entre las variables.

D. *Preparación de los Datos*

Como parte de la preparación de los datos, se normalizaron las variables independientes para asegurar la comparabilidad y estabilidad durante el entrenamiento de los modelos de Machine Learning. Posteriormente, los datos fueron divididos en conjuntos de entrenamiento (80%) y validación (20%) para evaluar el desempeño de los modelos Machine Learning en la predicción de la probabilidad de riesgo ambiental asociado a residuos costeros a nivel global haciendo uso de los indicadores estadísticos de precisión, recall, F1-Score, el área bajo la curva ROC y el índice de Gini para determinar su capacidad predictiva.

E. *Modelos de Machine Learning*

Se entrenaron y compararon los modelos de Machine Learning Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost. La elección de estos modelos se fundamenta en su capacidad para manejar grandes volúmenes de datos, capturar relaciones no lineales y ofrecer métricas de importancia de variables, lo que facilita la interpretación de los resultados.

i) Modelo de Random Forest

Según Salman et al. [14], el modelo de Machine Learning Random Forest es una técnica ampliamente utilizada en tareas de clasificación. Este modelo se fundamenta en el principio de los árboles de decisión, mediante la construcción de un conjunto de árboles generados a partir de subconjuntos aleatorios del conjunto de datos original. Cada árbol individual se entrena de manera independiente, y sus resultados se combinan posteriormente para producir una predicción final más robusta y precisa. Esta estrategia de agregación reduce la varianza y mejora la capacidad de generalización del modelo. El modelo de Random Forest se considera una de las metodologías más eficientes y versátiles en el campo del Machine Learning, con aplicaciones destacadas en la categorización automática, la predicción de datos y el aprendizaje supervisado.

Los detalles sobre la implementación y documentación del modelo Random Forest se encuentran disponibles en los tutoriales oficiales publicados en el proyecto scikit-learn [15].

ii) Modelo de Gradient Boosting

Según Natekin et al. [16], Gradient Boosting pertenece a una familia de potentes técnicas de Machine Learning que han demostrado un desempeño sobresaliente en una amplia variedad de aplicaciones prácticas. Este enfoque se caracteriza por su alta capacidad de personalización, permitiendo adaptarse a distintas funciones de pérdida y a los requerimientos específicos de cada problema. Además, se distingue por ofrecer resultados de elevada precisión y una notable capacidad de generalización. El método se enmarca dentro de los modelos de Boosting, cuya idea fundamental consiste en construir un conjunto, llamado ensemble, de modelos de forma secuencial. En cada iteración, un nuevo modelo base débil se entrena para corregir los errores cometidos por el conjunto acumulado hasta ese momento. Este proceso iterativo ajusta sucesivamente nuevos modelos con el objetivo de mejorar la estimación de la variable de respuesta. En esencia, cada modelo base se construye de manera que esté altamente correlacionado con el gradiente negativo de la función de pérdida del modelo completo, lo que permite optimizar progresivamente el rendimiento global del sistema.

Los detalles sobre la implementación y documentación del modelo Gradient Boosting Machine se encuentran disponibles en los tutoriales oficiales publicados en el proyecto scikit-learn [17].

iii) Modelo de XGBoost

Según Chen et al. [18], los modelos de Boosting constituyen una de las técnicas de Machine Learning más eficaces y ampliamente utilizadas. Entre ellos destaca el modelo XGBoost (Extreme Gradient Boosting), un sistema escalable de aumento de árboles que ha sido adoptado extensamente por la comunidad científica y por profesionales del análisis de datos debido a su alto rendimiento y eficiencia computacional. El éxito de XGBoost radica principalmente en su alta escalabilidad, lograda mediante una serie de optimizaciones tanto algorítmicas como de sistema. Entre las innovaciones más relevantes se incluyen: un nuevo algoritmo de aprendizaje de árboles diseñado para manejar datos dispersos de manera eficiente, y un procedimiento de boceto de cuantiles ponderados (weighted quantile sketch), teóricamente fundamentado, que permite gestionar los pesos de las instancias en el proceso de aprendizaje aproximado de árboles. Además, XGBoost incorpora mecanismos de computación paralela y distribuida, los cuales aceleran significativamente el entrenamiento y facilitan una exploración más rápida y exhaustiva del espacio de modelos.

Los detalles sobre la implementación y documentación del modelo XGBoost se encuentran disponibles en los tutoriales oficiales publicados en la página de XGBoost [19].

iv) Modelo de LightGBM

Según Ke et al. [20], se propone el modelo de Machine Learning LightGBM, el cual introduce dos técnicas innovadoras: el Muestreo Unilateral Basado en Gradiente (Gradient-based One-Side Sampling, GOSS) y el Agrupamiento Exclusivo de Características (Exclusive Feature Bundling, EFB). Estas metodologías permiten manejar eficientemente grandes volúmenes de datos y un elevado número de variables, respectivamente. A través de análisis teóricos y experimentos empíricos, los autores demostraron que ambas técnicas son coherentes con los fundamentos teóricos del aprendizaje de gradiente y que, en conjunto, proporcionan mejoras sustanciales en el rendimiento computacional del modelo. En particular, los resultados evidencian que LightGBM supera significativamente a XGBoost en términos de velocidad de cálculo y uso de memoria, manteniendo niveles de precisión comparables.

Los detalles sobre la implementación y documentación del modelo LightGBM se encuentran disponibles en los tutoriales oficiales publicados por Microsoft Corporation [21].

v) Modelo de CatBoost

Según Prokhorenkova et. al [22] y respecto a otras implementaciones de aumento de gradiente, el modelo de Machine Learning de CatBoost presenta dos avances algorítmicos cruciales como la implementación del aumento ordenado, una alternativa basada en permutaciones al algoritmo clásico, y un algoritmo innovador para el procesamiento de características categóricas. Ambas técnicas se crearon para combatir un sesgo de predicción causado por un tipo especial de fuga de datos presente en todas las implementaciones actuales de algoritmos de aumento de gradiente.

Según Ding et. al [23], CatBoost es un modelo de Machine Learning eficaz para pronosticar atributos categóricos, el cual emplea el método de potenciación de gradiente y utiliza árboles de decisión binarios como predictores fundamentales.

Los detalles sobre la implementación y documentación del modelo CatBoost se encuentran disponibles en los tutoriales oficiales publicados por Yandex LLC [24].

F. Evaluación del Desempeño

Cada uno de los modelos fue entrenado mediante la técnica de búsqueda de hiperparámetros GridSearchCV, utilizando validación cruzada de tipo k-fold con k = 5. Este procedimiento permitió minimizar el riesgo de sobreajuste y garantizar una adecuada capacidad de generalización de los modelos. Posteriormente, para evaluar el desempeño predictivo de los modelos de Machine Learning, se comparó su capacidad de discriminación del riesgo ambiental asociado a los residuos costeros, identificando las variables que más contribuyen a dicha probabilidad de riesgo. La evaluación se realizó mediante indicadores estadísticos, tales como la precision, recall, la puntuación F1-Score, el área bajo la curva ROC y el índice de Gini.

i) Matriz de confusión

Según George [25], los indicadores estadísticos como la precisión (precision), la sensibilidad (recall) y la puntuación F1 (F1-score) pueden analizarse mediante la matriz de confusión, una tabla que resume el desempeño de un modelo de clasificación al comparar los valores reales con los valores predichos. En el caso de una clasificación binaria, dicha matriz está compuesta por cuatro categorías: verdaderos positivos (VP), que representan las predicciones correctas de casos positivos; verdaderos negativos (VN), correspondientes a las predicciones correctas de casos negativos; falsos positivos (FP), cuando el modelo predice un valor positivo (1) para un caso que en realidad es negativo (0); y falsos negativos (FN), cuando se predice un valor negativo (0) para un caso que en realidad es positivo (1).

La estructura general de esta matriz se presenta en la Tabla I.

TABLA I
MATRIZ DE CONFUSIÓN

	Predicción	
Real	**Negativo**	**Positivo**
Negativo	Verdaderos Negativos (VN)	Falsos Positivos (FP)
Positivo	Falsos Negativos (FN)	Verdaderos Positivos (VP)

ii) Precision, recall, and F1-Score

De acuerdo con George [25], los indicadores estadísticos de desempeño, como la precisión, la sensibilidad y la puntuación F1-Score, dependen directamente de los componentes de la matriz de confusión: verdaderos positivos (VP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (VN), los cuales permiten evaluar la calidad de las predicciones del modelo.

La precision se define como la proporción de predicciones positivas que fueron correctamente clasificadas, es decir, el número de verdaderos positivos (VP) dividido entre la suma de verdaderos positivos (VP) y falsos positivos (FP). Matemáticamente, se expresa como:

$$\text{Precision}=VP/(VP + FP) \quad (1)$$

La sensibilidad o recuperación (recall) se define como la proporción de casos positivos correctamente identificados por el modelo respecto al total de casos positivos reales. En otras palabras, mide la capacidad del modelo para detectar correctamente las observaciones positivas. Matemáticamente, se expresa como:

$$\text{Recall}=VP/(VP + FN) \quad (2)$$

La puntuación F1-Score es la media armónica de estas dos medidas, que es:

$$\text{F1-Score}=2(\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \quad (3)$$

Finalmente, el modelo de Machine Learning que muestre mayores valores de precision, recall y F1-Score es el modelo con mejor desempeño para el pronóstico de la variable dependiente.

iii) Área bajo la curva ROC

Según George [25], una métrica adicional de gran utilidad en la evaluación de modelos de clasificación es el área bajo la curva ROC (Receiver Operating Characteristic, por sus siglas en inglés). La fig. 1 muestra la curva ROC de un modelo de Machine Learning con un área bajo la curva ROC de 0.92, en comparación con el valor ideal de 1.0, correspondiente a un modelo con desempeño predictivo perfecto.

En dicha representación, el eje X indica la tasa de falsos positivos (TFP o FPR por sus siglas en inglés), calculada como:

$$\text{TFP}=FP/(FP + VN) \quad (4)$$

Mientras que el eje Y muestra la tasa de verdaderos positivos (TVP o TPR por sus siglas en inglés), definida como:

$$\text{TVP}=VP/(VP + FN) \quad (5)$$

La posición de ambas curvas por encima de la línea diagonal discontinua, que representa el comportamiento aleatorio, indica que los modelos evaluados presentan un rendimiento superior al que se obtendría por azar.

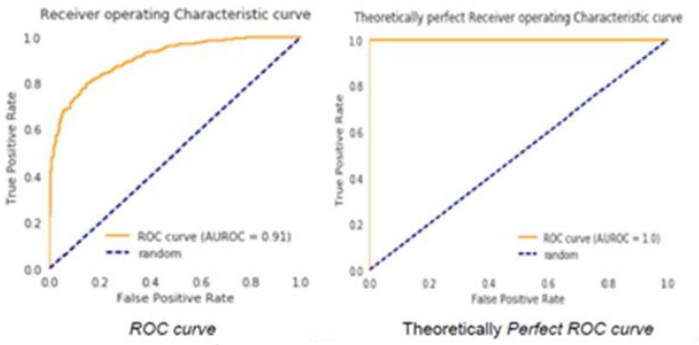

Fig. 1. Área bajo la curva ROC [26]

iv) Índice de GINI

Adams y Hand [27], así como Hand y Anagnostopoulos [28], señalan que el índice de Gini es una métrica empleada para evaluar el desempeño de modelos cuya variable dependiente es dicotómica, constituyendo una alternativa al uso del área bajo la curva ROC. Este índice mide la capacidad discriminativa del modelo, de modo que valores más cercanos a 1 indican un mayor poder predictivo. La expresión matemática para su cálculo se presenta a continuación:

$$\text{GINI}=2(\text{ROC}-0.5) \quad (6)$$

G. Herramienta Computacional

El análisis univariado y bivariado de los datos se realizó utilizando RStudio en su versión 2025.09.2+418, soportado en R en su versión 4.5.2. Para el procesamiento y visualización de la información se emplearon las bibliotecas dplyr y ggplot2, mientras que las pruebas estadísticas se ejecutaron mediante las funciones shapiro.test() y wilcox.test().

Por otro lado, el procesamiento, análisis de datos e implementación de los modelos de Machine Learning se realizaron utilizando Python en su versión 3.13 sobre Jupyter Notebook, distribuido por Anaconda. Para el procesamiento y análisis de datos se emplearon las bibliotecas Pandas, NumPy, Matplotlib y Seaborn, mientras que el desarrollo y entrenamiento de los modelos de Machine Learning se llevó a cabo mediante las bibliotecas Scikit-learn, XGBoost, LightGBM y CatBoost.

En cuanto al hardware, los modelos de Machine Learning fueron entrenados en una laptop con procesador Intel Core i7-1255U de 12ª generación (1.70 GHz) y 24 GB de memoria RAM.

III. Resultados y Discusión

A. Análisis Estadístico univariado

En la figura 2 se observa que China (59 MT), Estados Unidos (42 MT) e India (26 MT) son los principales generadores de residuos plásticos a nivel mundial, superando ampliamente al resto de los países. La generación de residuos disminuye de manera considerable después de los tres primeros, con Japón, Alemania, Brasil, Indonesia y Rusia produciendo aproximadamente 6 millones de toneladas cada uno. Estos resultados indican que los países desarrollados e industrializados continúan siendo los mayores contribuyentes, lo que resalta la necesidad de fortalecer las políticas globales de gestión de residuos.

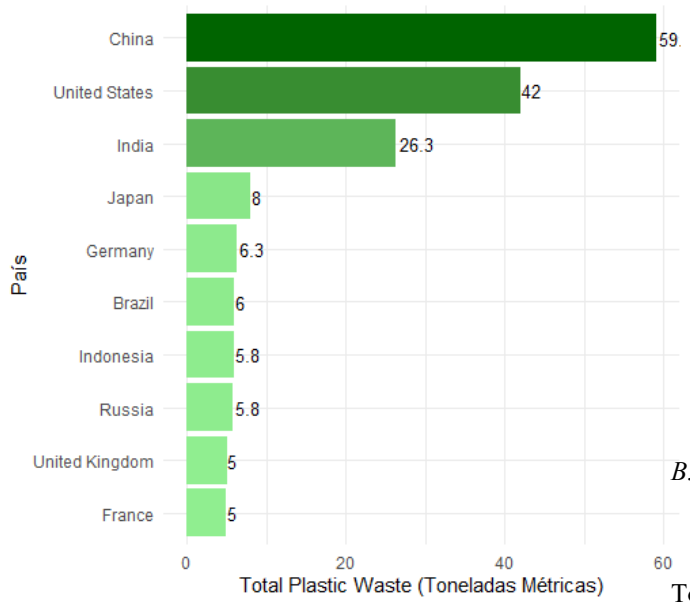

Fig. 2. Top 10 de países con mayor producción de residuos plásticos en millones de toneladas métricas

En la figura 3, se aprecia que el embalaje de consumo (Consumer Packaging) es la principal fuente de residuos plásticos en la mayoría de los países.

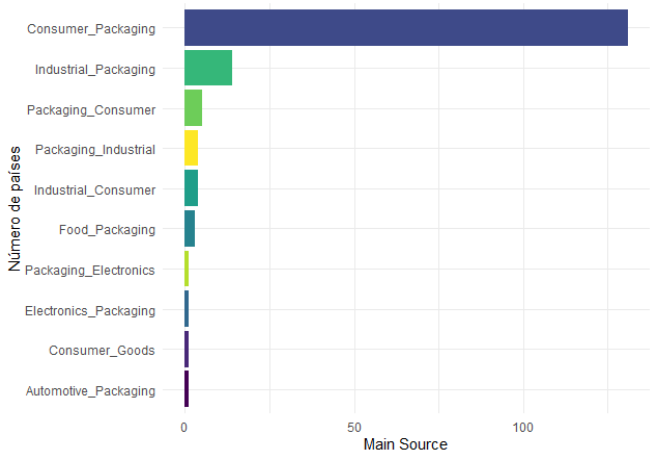

Fig. 3. Principales fuentes de residuos plásticos según cantidad de países

En la figura 4, se aprecia que Japón lidera a nivel mundial con una tasa de reciclaje del 84.8%, seguido de Singapur (59.8%) y Corea del Sur (59.1%), lo que demuestra la eficacia de sus políticas de gestión de residuos. Los países europeos que dominan este top son Alemania, Austria y los Países Bajos, pues reciclan más del 55% de sus residuos plásticos.

A pesar de ser uno de los principales productores de residuos plásticos, Estados Unidos no figura en la lista, lo que evidencia una brecha en la eficiencia del reciclaje.

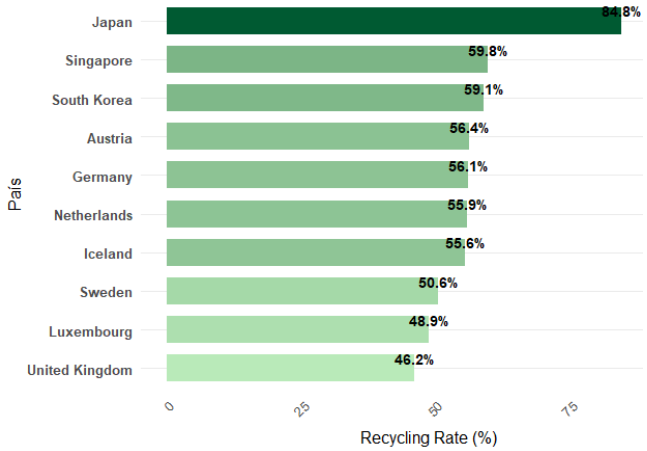

Fig. 4. Top 10 de países con mayor tasa de reciclaje

B. Análisis Estadístico bivariado

i) Prueba de Shapiro-Wilk

Para evaluar si las variables independientes cuantitativas Total_Plastic_Waste_MT, Recycling_Rate y Per_Capita_Waste_KG se ajustan a una distribución normal, se aplicó la prueba estadística de Shapiro-Wilk. Los resultados se presentan en la Tabla II. Dado que el p-value obtenido es menor que el nivel de significancia del 5 %, se rechaza la hipótesis nula (H_0), lo que indica que las variables no se aproximan a una distribución normal.

TABLA II
PRUEBA SHAPIRO-WILK

Variables independientes	W	p-value
Total_Plastic_Waste_MT	0.24034	2.20E-16
Recycling_Rate	0.75851	3.45E-15
Per_Capita_Waste_KG	0.51493	2.20E-16

ii) Prueba Wilcoxon o Mann-Whitney

Para evaluar la significancia entre la variable dependiente Risk y las variables independientes cuantitativas Total_Plastic_Waste_MT, Recycling_Rate y Per_Capita_Waste_KG, se aplicó la prueba estadística de Wilcoxon o Mann–Whitney, esto debido a que, según los resultados de la prueba de normalidad Shapiro-Wilk, las variables independientes no se aproximan a una distribución normal. Los resultados se presentan en la Tabla III. Dado que

el p-value obtenido es menor que el nivel de significancia del 5 %, se rechaza la hipótesis nula (H_0), lo que indica que existen diferencias significativas entre las distribuciones de las variables independientes mencionadas en los dos grupos de la variable dependiente (High y Not_High).

TABLA III
PRUEBA WILCOXON O MANN-WHITNEY

Variables independientes	W	p-value
Total_Plastic_Waste_MT by Risk	2109.5	2.81E-05
Recycling_Rate by Risk	1997.5	5.26E-06
Per_Capita_Waste_KG by Risk	2481.5	0.002945

En las figuras 5, 6 y 7 se aprecian, de forma exploratoria y visual, las diferencias en las distribuciones de las variables independientes según los dos grupos de la variable dependiente (High y Not_High).

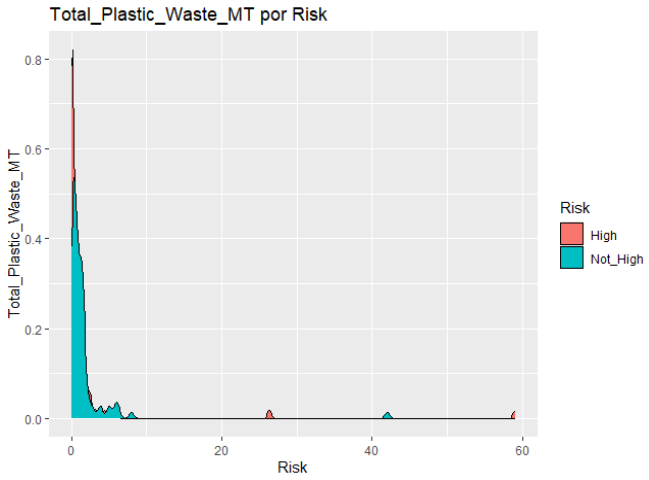

Fig. 5. Producción de residuos plásticos según el nivel de riesgo ambiental asociado a residuos costeros a nivel global

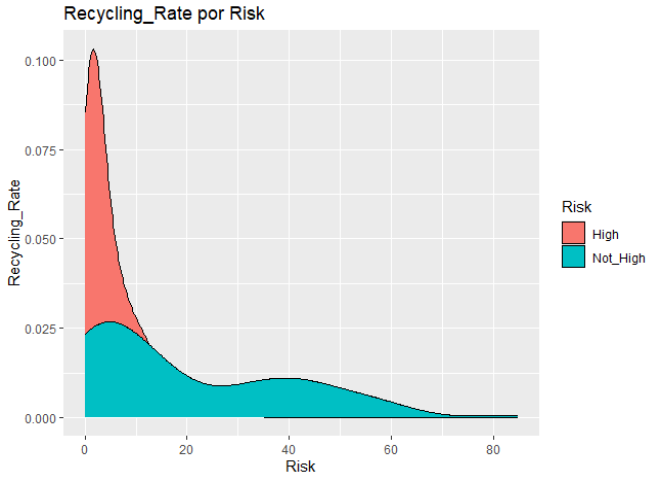

Fig. 6. Tasa nacional de reciclaje según el nivel de riesgo ambiental asociado a residuos costeros a nivel global

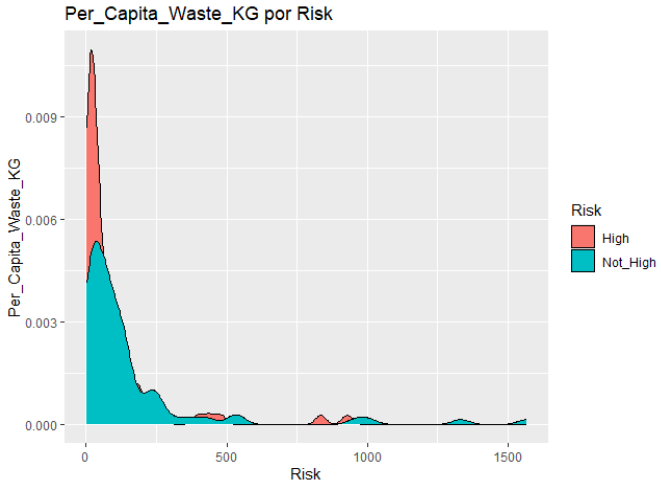

Fig. 7 Producción de residuos per cápita según el nivel de riesgo ambiental asociado a residuos costeros a nivel global

iii) Prueba Chi-cuadrado

Para evaluar la significancia entre la variable dependiente Risk y la variable independiente cualitativa Main_Sources, se aplicó la prueba estadística de Chi-cuadrado. Los resultados se presentan en la Tabla IV. Dado que el p-value obtenido es menor que el nivel de significancia del 5 %, se rechaza la hipótesis nula (H_0), lo que indica que existen dependencia entre la variable independiente y la variable dependiente.

TABLA IV
PRUEBA CHI-CUADRADO

Variable independiente	X-squared	p-value
Main_Sources by Risk	19.885	1.86E-02

C. *Modelos de Machine Learning*

La base de datos, que incluye un total de 165 países, se dividió en una muestra de entrenamiento de 132 países (80 %) y una muestra de validación de 33 países (20 %), cuyos detalles se presentan en la Tabla V.

TABLA V
DATOS DE ENTRENAMIENTO Y VALIDACIÓN

Datos	Paises	% Paises
Entrenamiento	132	80%
Validación	33	20%

Con la muestra de entrenamiento se generaron los modelos de Machine Learning: Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost. Posteriormente, con la muestra de validación, se evaluó el desempeño de estos modelos en la predicción de la probabilidad de riesgo ambiental asociado a residuos costeros a nivel global,

utilizando los indicadores estadísticos de precisión, recall, F1-Score, área bajo la curva ROC y índice de Gini, cuyos resultados se presentan en la Tabla VI.

TABLA VI
DESEMPEÑO DE MODELOS DE MACHINE LEARNING

Modelos	Precision	Recall	F1-score	ROC	GINI
Random Forest	0.5455	0.8000	0.6486	0.6852	0.3704
Gradient Boosting	0.4667	0.4667	0.4667	0.6593	0.3185
XGBoost	0.4286	0.4000	0.4138	0.6111	0.2222
LightGBM	0.5455	0.8000	0.6486	0.6630	0.3259
CatBoost	0.5455	0.8000	0.6486	0.6444	0.2889

Respecto al indicador estadístico de precision, los modelos Random Forest, LightGBM y CatBoost tienen el mayor valor (0.5455), mientras que el modelo XGBoost el valor más bajo (0.4286).

Respecto al indicador estadístico de recall, los modelos Random Forest, LightGBM y CatBoost logran un valor alto (0.8), lo que significa que identifican correctamente la mayoría de países con un riesgo ambiental asociado a residuos costeros a nivel global. El modelo de XGBoost tiene el valor más bajo (0.4).

Respecto al indicador estadístico de F1-score, los modelos Random Forest, LightGBM y CatBoost logran el valor más alto (0.6486), mientras que el modelo XGBoost presenta el valor más bajo (0.4138).

Respecto al indicador estadístico del área bajo la curva ROC, el modelo de Random Forest tiene el valor más alto (0.6852), indicando la mejor discriminación entre los países con un riesgo ambiental alto y bajo asociado a residuos costeros a nivel global, mientras que el modelo de XGBoost presenta es el valor más bajo (0.6111).

Respecto al indicador estadístico de GINI, el modelo de Random Forest presenta la mejor capacidad de discriminación (0.3704), mientras que el modelo de XGBoost presenta es el valor más bajo de desempeño (0.2222).

Finalmente, en los resultados comparativos de los modelos de Machine Learning se observa que Random Forest, LightGBM y CatBoost presentan un desempeño similar, con valores de precisión, recall y F1-score, destacando Random Forest por su mayor capacidad de discriminación en la predicción del riesgo ambiental asociado a residuos costeros a nivel global, reflejada en el área bajo la curva ROC y el índice de Gini. El modelo de Gradient Boosting muestra un desempeño intermedio, mientras que XGBoost alcanza los valores más bajos en todas las métricas evaluadas, indicando un menor poder predictivo para la variable de riesgo ambiental asociada a residuos costeros.

D. Discusión

Los resultados obtenidos en esta investigación confirman la utilidad de los modelos de Machine Learning como herramientas clave para la gestión del riesgo ambiental asociado a residuos costeros a nivel global. Estos hallazgos son consistentes con estudios previos, como los de Abbasi et al. [5], Fang et al. [6], Pourzangbar et al. [7] y Edeh et al. [8], quienes destacan que la implementación y optimización adecuada de modelos de Machine Learning contribuye significativamente a reducir la probabilidad de riesgos ambientales.

El análisis comparativo de los modelos de Machine Learning de Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost, permitió identificar diferencias significativas en sus capacidades predictivas. Según los resultados obtenidos, Random Forest presentó el mejor desempeño, destacando por su elevada capacidad de discriminación en la predicción del riesgo ambiental asociado a residuos costeros a nivel global, evidenciada en el área bajo la curva ROC y el índice de Gini. Estos hallazgos coinciden con lo reportado por Binetti et al. [9], quienes resaltan que Random Forest es uno de los modelos más utilizados en la monitorización ambiental, debido a su efectividad en tareas de clasificación y regresión.

Una de las limitaciones de esta investigación radica en su diseño transversal, que impide observar cambios en el comportamiento de los países a lo largo del tiempo. Esto es relevante, ya que el riesgo ambiental asociado a residuos costeros a nivel global es dinámico y puede estar influenciado por factores no considerados en el presente estudio.

IV. CONCLUSIONES

Se evaluó el desempeño de los modelos de Machine Learning Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost en la predicción de la probabilidad de riesgo ambiental asociado a residuos costeros a nivel global. Para el modelo de Random Forest, se obtuvieron los siguientes resultados: precision de 0.5455, recall de 0.8000, F1-Score de 0.6486, área bajo la curva ROC de 0.6852 e índice de Gini de 0.3704. Para el modelo de Gradient Boosting, se obtuvieron los siguientes resultados: precision de 0.4667, recall de 0.4667, F1-Score de 0.4667, área bajo la curva ROC de 0.6593 e índice de Gini de 0.3185. Para el modelo de XGBoost, se obtuvieron los siguientes resultados: precision de 0.4286, recall de 0.4000, F1-Score de 0.4138, área bajo la curva ROC de 0.6111 e índice de Gini de 0.2222. Para el modelo de LightGBM, se obtuvieron los siguientes resultados: precision de 0.5455, recall de 0.8000, F1-Score de 0.6486, área bajo la curva ROC de 0.6630 e índice de Gini de 0.3259. Para el modelo de CatBoost, se obtuvieron los siguientes resultados: precision de 0.5455, recall de 0.8000, F1-Score de 0.6486, área bajo la curva ROC de 0.6444 e índice de Gini de 0.2889.

La presente investigación permitió identificar los factores más influyentes en el riesgo ambiental asociado a residuos costeros a nivel global. Entre las variables cuantitativas, la producción de residuos plásticos por país (Total_Plastic_Waste_MT), la tasa nacional de reciclaje (Recycling_Rate) y la producción de residuos per cápita (Per_Capita_Waste_KG) mostraron una contribución

estadísticamente significativa en la predicción de la probabilidad de riesgo ambiental, con un nivel de confianza del 95%. Asimismo, la variable cualitativa principales fuentes de residuos plásticos por país (Main_Sources) también presentó una influencia significativa en la predicción del riesgo, confirmando su relevancia para la evaluación global del impacto ambiental de los residuos costeros.

El análisis comparativo de los modelos de Machine Learning permitió evaluar su desempeño en la predicción del riesgo ambiental asociado a residuos costeros a nivel global. Entre los modelos evaluados se usaron Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost. El modelo de Random Forest se destacó por su mayor capacidad de discriminación y precisión predictiva, evidenciada en métricas como el área bajo la curva ROC y el índice de Gini, con valores de 0.682 y 0.3704, respectivamente. Estos hallazgos sugieren que la implementación de Random Forest puede ser particularmente útil para orientar la formulación de políticas de conservación costera, permitiendo priorizar acciones en áreas de mayor riesgo ambiental y optimizando la gestión de residuos costeros a nivel global.

Por lo tanto, se concluye que el modelo de Machine Learning más adecuado para la predicción del riesgo ambiental asociado a residuos costeros a nivel global es el modelo de Random Forest.

REFERENCIAS

[1] European Environment Agency, "Marine litter – a growing threat worldwide," EEA Highlights, 07 Ago. 2013. eea.europa.eu. https://www.eea.europa.eu/highlights/marine-litter-2013-a-growing [Consultado: 08 Oct. 2025].

[2] United Nations, "El plástico, que ya ha atragantado nuestros océanos, terminará por asfixiarnos a todos si no actuamos rápidamente," Noticias ONU, 21 Oct. 2021. un.org. https://www.un.org/es/sustainabledevelopment/es/2021/10/el-plastico-que-ya-ha-atragantado-nuestros-oceanos-terminara-por-asfixiarnos-a-todos-si-no-actuamos-rapidamente/ [Consultado: 08 Oct. 2025].

[3] Banco Mundial, "Abordar la contaminación marina generada por plásticos en América Latina y el Caribe," Brief - Región América Latina y el Caribe, 31 May. 2023. bancomundial.org. https://www.bancomundial.org/es/region/lac/brief/addressing-marine-plastics-in-latin-america-and-the-caribbean [Consultado: 08 Oct. 2025].

[4] D. Contreras Zuloaga, "El plástico invade al mar peruano, amenazando un ecosistema entero," Instituto de la Naturaleza, Tierra y Energía – PUCP, Lima, Perú, 18 Ago. 2023. inte.pucp.edu.pe. https://inte.pucp.edu.pe/noticias-y-eventos/noticias/el-plastico-invade-al-mar-peruano-amenazando-un-ecosistema-entero/. [Consultado: 08 Oct. 2025].

[5] M. Abbasi, M. Abduli, B. Omidvar, and A. Baghvand, "Results uncertainty of support vector machine and hybrid of wavelet transform-support vector machine models for solid waste generation forecasting," Environmental Progress & Sustainable Energy, vol. 33, pp. 220–228, 4 2014.

[6] Fang, B., Yu, J., Chen, Z., Osman, A. I., Farghali, M., Ihara, I., Hamza, E. H., Rooney, D. W., & Yap, P. S. (2023). Artificial intelligence for waste management in smart cities: a review. Environmental chemistry letters, 1–31. Advance online publication. https://doi.org/10.1007/s10311-023-01604-3

[7] Pourzangbar A, Jalali M and Brocchini M (2023), Machine learning application in modelling marine and coastal phenomena: a critical review. Front. Environ. Eng. 2:1235557. doi: https://doi.org/10.3389/fenve.2023.1235557

[8] Edeh, M. O., Dalal, S., Alhussein, M., Aurangzeb, K., Seth, B., & Kumar, K. (2024). A novel deep learning model for predicting marine pollution for sustainable ocean management. PeerJ. Computer science, 10, e2482. https://doi.org/10.7717/peerj-cs.2482

[9] Binetti, M. S., Massarelli, C., & Uricchio, V. F. (2024). Machine Learning in Geosciences: A Review of Complex Environmental Monitoring Applications. Machine Learning and Knowledge Extraction, 6(2), 1263-1280. https://doi.org/10.3390/make6020059

[10] P. Dongre. "Global Plastic Waste 2023: Country-wise Analysis," kaggle.com. https://www.kaggle.com/datasets/prajwaldongre/global-plastic-waste-2023-a-country-wise-analysis/data [Consultado: 08 Oct. 2025].

[11] J. P. Royston, "An extension of Shapiro and Wilk's W test for normality to large samples," Journal of the Royal Statistical Society Series C: Applied Statistics, vol. 31, no. 2, pp. 115-124, Jun. 1982. doi:10.2307/2347973

[12] M. Hollander y D. A. Wolfe, Nonparametric Statistical Methods, 2.ª ed., New York, NY, USA: John Wiley & Sons, 1999.

[13] A. Agresti, An Introduction to Categorical Data Analysis, 2nd ed., New York, NY, USA: John Wiley & Sons, 2007.

[14] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," Babylonian Journal of Machine Learning, vol. 2024, no. 69-79, Jun. 2024, doi: 10.58496/BJML/2024/007

[15] Scikit Learn, "sklearn.ensemble.RandomForestClassifier - scikit-learn documentation," scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [Consultado: 08 Oct. 2025].

[16] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," Frontiers in Neurorobotics, vol. 7, Art. 21, Dec. 2013, doi: 10.3389/fnbot.2013.00021.

[17] Scikit Learn, "sklearn.ensemble. GradientBoosting Classifier - scikit-learn documentation," scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html [Consultado: 08 Oct. 2025].

[18] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, Aug. 2016.

[19] XGBoost, "XGBoost Documentation," xgboost.readthedocs.io. https://xgboost.readthedocs.io/en/stable/ [Consultado: 08 Oct. 2025].

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[21] Microsoft Corporation, "LightGBM — documentation," lightgbm.readthedocs.io. https://lightgbm.readthedocs.io/en/stable/ [Consultado: 08 Oct. 2025].

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, vol. 31, pp. 6639-6649, Dec. 2018.

[23] H. Ding, X. Yang, and T. Qi, "Hybrid Machine Learning Models Based on CATBoost Classifier for Assessing Students' Academic Performance," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 7, 2024, doi: 10.14569/IJACSA.2024.0150709.

[24] Yandex LLC, "CatBoost — Documentation," catboost.ai. https://catboost.ai/docs/en/ [Consultado: 08 Oct. 2025].

[25] N. George, Practical Data Science with Python: Learn Tools and Techniques from Hands-On Examples to Extract Insights from Data, Packt Publishing, 2021.

[26] Fonseca & Lopes (2017) – "Calibration of Machine Learning Classifiers for Probability of Default Modelling"

[27] Adams, N. M., & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. Pattern Recognition, 32(7), 1139–1147. https://doi.org/10.1016/S0031-3203(98)00154-X

[28] Hand, D. J., & Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognition Letters, 34(5), 492–495. https://doi.org/10.1016/j.patrec.2012.12.004