

ESTUDIO DE MODELOS DE PREDICCIÓN EN BIG DATA
STUDY OF PREDICTION MODELS IN BIG DATA

Autores: ¹Mario Alberto Ibarra Martínez, ²Edgar Xavier Mendoza Arce y ³Wilson Javier Romero Berrones.

¹ORCID ID: <https://orcid.org/0009-0004-5500-871X>

²ORCID ID: <https://orcid.org/0009-0001-4368-6941>

³ORCID ID: <https://orcid.org/0009-0007-3791-0525>

¹E-mail de contacto: mibarra@uagraria.edu.ec

²E-mail de contacto: emendozaa9@unemi.edu.ec

³E-mail de contacto: wromero@uagraria.edu.ec

Afiliación: ¹³*Universidad Agraria del Ecuador, (Ecuador). ²*Universidad Estatal de Milagro, (Ecuador).

Artículo recibido: 7 de julio del 2025

Artículo revisado: 8 de julio del 2025

Artículo aprobado: 18 de julio del 2025

¹Ingeniero Industrial, graduado de la Escuela Superior Politécnica del Litoral, (Ecuador). Maestrante en Ingeniería de Software y Sistemas Informáticos de la Universidad Internacional de la Rioja, (España). Docente de la Universidad Agraria del Ecuador, (Ecuador), con años de experiencia laboral.

²Ingeniero Mecánico, graduado de la Escuela Superior Politécnica del Litoral, (Ecuador). Magíster en Global Production Engineering, graduado de Technische Universität, (Alemania). Docente de la Universidad Estatal de Milagro, (Ecuador), con años de experiencia laboral.

³Licenciado en Ciencias de Educación, especialización Informática, graduado de la Universidad de Guayaquil, (Ecuador). Magíster en Gerencia Educativa, (Ecuador), con años de experiencia laboral.

Resumen

La integración de enfoques de IA en la tecnología Big Data ha facilitado la toma de decisiones, gracias a la creciente disponibilidad de datos, es por ello que en el presente estudio se planteó examinar la interrelación entre los modelos predictivos y la tecnología Big Data. Para tal efecto, se realizó una revisión sistemática a partir de la recopilación de datos pertinentes con la metodología PRISMA seleccionando seis documentos científicos que cumplieran con estándares de alta calidad. Los hallazgos ratifican que hay una tendencia mayor hacia la implementación de modelos no lineales, redes neuronales, así como métodos híbridos con datos multitemporales y polifacéticos. En los resultados, se determinó que la sinergia de la tecnología Big Data y modelos predictivos mejora la capacidad de anticipación y adaptación de sectores estratégicos. Se concluye que la efectividad en la predictibilidad no solamente depende de la estructura modelo algorítmica sino de la calidad de la data, de su contexto, y de la capacidad

analítica sobre cómo actuar para intervenir en la estrategia.

Palabras clave: Modelos, Predictivo, Big Data, Algoritmos, Datos.

Abstract

The integration of AI approaches in Big Data technology has facilitated decision making, thanks to the increasing availability of data, which is why this study set out to examine the interrelationship between predictive models and Big Data technology. To this end, a systematic review was carried out based on the collection of relevant data using the PRISMA methodology, selecting six scientific papers that met high quality standards. The findings ratify that there is a major trend towards the implementation of nonlinear models, neural networks, as well as hybrid methods with multitemporal and multifaceted data. In the results, it was determined that the synergy of Big Data technology and predictive models improves the capacity of anticipation and adaptation of strategic sectors. It is concluded that the effectiveness in predictability depends not only on the algorithmic model structure but also on the quality of the data, its context, and

the analytical capacity on how to act in order to intervene in the strategy.

Keywords: Models, Predictive, Big Data, Algorithms, Data.

Sumário

A integração de abordagens de IA na tecnologia de Big Data facilitou a tomada de decisões, graças à crescente disponibilidade de dados, razão pela qual este estudo se propôs a examinar a inter-relação entre modelos preditivos e tecnologia de Big Data. Para o efeito, foi realizada uma revisão sistemática através da recolha de dados relevantes utilizando a metodologia PRISMA e seleccionando seis artigos científicos que cumpriam elevados padrões de qualidade. Os resultados confirmam que existe uma tendência crescente para a implementação de modelos não lineares, redes neurais, bem como métodos híbridos com dados multitemporais e multifacetados. Nos resultados, determinou-se que a sinergia da tecnologia Big Data e dos modelos preditivos melhora a capacidade de antecipação e adaptação dos sectores estratégicos. Conclui-se que a eficácia na previsibilidade depende não só da estrutura do modelo algorítmico mas também da qualidade dos dados, do seu contexto e da capacidade analítica sobre como atuar para intervir na estratégia.

Palavras-chave: Modelos, Preditivos, Big Data, Algoritmos, Dados.

Introducción

En la actualidad, el uso de modelos de predicción en entornos de Big Data ha adquirido un papel central en la toma de decisiones estratégicas, desde la planificación urbana hasta la gestión de recursos naturales (Bedoya et al., 2024). La rápida digitalización de las actividades humanas ha generado volúmenes masivos de datos cuya complejidad requiere enfoques algorítmicos avanzados para identificar patrones, prever comportamientos y reducir incertidumbres, un ejemplo notable se observa en la predicción de la capacidad de carga ambiental de ríos en Vietnam mediante

redes Long Short-Term Memory (LSTM), alcanzando niveles de error aceptables con RMSE de 22.16 kg/día para nitratos y 38.92 kg/día para demanda biológica de oxígeno (Bui et al., 2024). Estos logros reflejan el potencial de los modelos predictivos en contextos dinámicos, multivariados y no lineales. La convergencia entre algoritmos de aprendizaje automático y tecnologías de Big Data ha permitido avanzar hacia predicciones más precisas en contextos reales, superando las limitaciones de los modelos tradicionales, este fenómeno se traduce en un cambio paradigmático en la investigación científica, que exige una reevaluación metodológica en función de los avances computacionales, estadísticos y sociales de la era digital.

Por lo tanto, el presente estudio aborda un problema de creciente relevancia: la necesidad de mejorar la capacidad predictiva en entornos complejos mediante la integración de Big Data y modelos predictivos avanzados (Cevallos et al., 2024). Esta necesidad responde al reto de extraer información útil de flujos de datos masivos, diversos y en tiempo real, cuya estructura no reconoce distribuciones tradicionales. Zhang et al. (2025) destacan la insuficiencia de los métodos inferenciales clásicos para producir predicciones individuales a partir de muestras probabilísticas, proponiendo un enfoque de inferencia predictiva basada en el diseño que mejora la precisión sin renunciar al rigor estadístico. Cheng y Yue (2024), demuestran que la combinación de redes neuronales convolucionales (CNN) y *long short-term memory* (LSTM) permite reducir el error cuadrático medio (MSE) en predicciones de demanda de bicicletas compartidas desde 0.0084 hasta 0.0066 en modelos híbridos CNN-LSTM.

Por otro lado, Bui et al. (2024) aplican LSTM en series temporales para predecir la capacidad de carga ambiental en sistemas hídricos, obteniendo resultados confiables en 24 tramos de río, lo que permite decisiones ambientales proactivas. Estas evidencias cuantificables refuerzan el alcance de este estudio, centrado en la exploración de enfoques híbridos para potenciar el valor predictivo en contextos masivos y multidimensionales de datos. En este sentido, los modelos predictivos constituyen sistemas computacionales diseñados para estimar valores futuros a partir de datos históricos, que emplean técnicas estadísticas, algoritmos de aprendizaje automático y redes neuronales, orientadas a minimizar el error de predicción (Moreno, 2022). La predicción puede construirse sobre el diseño muestral mediante modelos asistidos, conservando la validez inferencial aún en contextos complejos, lo cual representa un avance respecto a la inferencia clásica que solo estima parámetros poblacionales.

Los modelos se clasifican comúnmente en lineales; como regresión lineal y logística. No lineales; como árboles de decisión y *random forest* y modelos neuronales; como redes LSTM, CNN y modelos híbridos. El uso de aprendizaje por conjunto o *ensemble learning* mostró superioridad en predicciones geotécnicas urbanas al combinar múltiples algoritmos y reducir la varianza del modelo (Cong y Inazumi, 2024). Por su parte, Bui et al. (2024) muestran que el modelo LSTM, especializado en procesar datos secuenciales, puede capturar patrones temporales con elevada precisión, siendo especialmente útil en problemas de series temporales ambientales. Estos modelos funcionan ajustando parámetros internos con base en funciones de error como el Error Cuadrático Medio de la Raíz o *Root Mean Square Error* (RMSE), el Error Absoluto Medio

o *Mean Absolute Error* (MAE) y el Error Porcentual Absoluto Medio o *Mean Absolute Percentage Error* (MAPE). De esta manera, los modelos predictivos aportan beneficios fundamentales en el análisis de datos al permitir anticipar resultados futuros con base en patrones históricos, su capacidad para adaptarse a datos no lineales y multivariantes mejora la precisión en contextos complejos (Jiménez, 2024). Además, facilitan la detección temprana de anomalías, optimizan procesos de planificación y permiten la toma de decisiones proactiva.

Modelos como LSTM, *Random Forest* y CNN-LSTM muestran alto rendimiento en tareas de predicción temporal, clasificación o regresión, su implementación contribuye a reducir la incertidumbre, mejorar la eficiencia en sectores críticos como salud, transporte o energía, y fortalecer la resiliencia frente a eventos inesperados. En cambio, el término Big Data hace referencia a conjuntos de datos caracterizados por su gran volumen, velocidad de generación y variedad estructural, los cuales requieren métodos computacionales avanzados para su procesamiento (Hua, 2025). De acuerdo con Cong y Inazumi (2024), Big Data representa una herramienta estratégica en la planificación urbana, al permitir análisis en tiempo real que integran sensores inteligentes, algoritmos predictivos y sistemas de información geográfica para anticipar fenómenos como la subsidencia del terreno, su funcionamiento se apoya en tecnologías de almacenamiento distribuido, procesamiento paralelo como *Hadoop* y *Spark*, y algoritmos capaces de extraer conocimiento útil en tiempo real.

La arquitectura de Big Data implica una infraestructura que permita la ingestión, limpieza, integración y análisis de datos provenientes de múltiples fuentes como;

sensores, registros administrativos, redes sociales, etc. Esta infraestructura no solo posibilita el procesamiento de grandes volúmenes, sino que habilita la aplicación efectiva de modelos predictivos en contextos dinámicos (Jamarani et al., 2024). Entre sus beneficios destacan la capacidad para tomar decisiones informadas, anticiparse a eventos futuros, optimizar recursos y personalizar servicios; por ejemplo, en predicciones ambientales, el uso de Big Data permite modelar interacciones complejas entre factores meteorológicos, químicos e hidrológicos, como lo evidencian los estudios sobre capacidad de carga hídrica. En contextos urbanos, su integración con inteligencia artificial permite generar mapas inteligentes para gestionar riesgos geotécnicos (Cisneros et al., 2022). La sinergia entre modelos predictivos y Big Data se configura como un eje estratégico en la evolución de los sistemas inteligentes de toma de decisiones, los modelos predictivos adquieren mayor potencia al operar sobre conjuntos de datos masivos, diversos y multitemporales, ya que ello permite entrenarlos con mayor representatividad y adaptabilidad (Moreno, 2022).

En el estudio de Cheng y Yue (2024), la implementación de un modelo CNN-LSTM demuestra que esta combinación mejora significativamente el rendimiento predictivo al incorporar tanto características espaciales como temporales. En cambio, Cong y Inazumi (2024), documentan que la aplicación de Big Data y aprendizaje por conjunto en predicciones geotécnicas permite no solo detectar capas portantes del suelo, sino anticipar riesgos estructurales con alta precisión. Por su parte, Zhang et al. (2024) explican que la inferencia predictiva basada en diseño aprovecha los marcos probabilísticos del muestreo tradicional integrando modelos algorítmicos con bases de

datos administrativos y sensores, sin necesidad de asumir distribuciones poblacionales específicas. Estos casos exponen que, cuando se integran, los modelos predictivos y la infraestructura Big Data es posible extraer información valiosa en entornos donde la variabilidad, el volumen y la veracidad de los datos superan los límites del análisis convencional, su impacto trasciende el ámbito técnico, afectando positivamente la planificación de políticas públicas, la gestión ambiental y la eficiencia operativa de las organizaciones.

En función a lo expuesto, este estudio adopta un enfoque de revisión bibliográfica de tipo descriptivo y analítico, para lo cual se recopilaron y analizaron artículos científicos revisados por pares que abordan tanto el desarrollo de modelos predictivos como la aplicación de Big Data en distintos contextos. La selección se realizó considerando la relevancia, el rigor metodológico y la actualidad de las publicaciones, el análisis permitió establecer patrones de implementación, métricas de desempeño y escenarios de aplicación para evaluar la efectividad predictiva en entornos complejos. En consecuencia, surge la siguiente formulación del problema: ¿Cómo incide la integración de modelos predictivos en los sistemas de Big Data para mejorar la precisión de las predicciones en contextos complejos? En respuesta a esta interrogante, el objetivo general de esta investigación es examinar la interrelación entre los modelos predictivos y la tecnología Big Data, determinando sus alcances, limitaciones y beneficios conjuntos en procesos de análisis avanzado. Por su parte, la construcción teórica analiza las variables generando el siguiente aporte literario.

Big Data: Fundamentos, arquitectura e implicaciones analíticas

Big Data se ha consolidado como un paradigma técnico y analítico para el tratamiento de volúmenes masivos de datos con alta velocidad, variedad y veracidad, lo cual representa un desafío para los sistemas tradicionales de procesamiento de información (Jamarani et al., 2024). Las dimensiones esenciales de Big Data incluyen: volumen, que alude a la cantidad ingente de datos generados por sensores, redes y plataformas digitales; velocidad, relacionada con la rapidez del flujo de datos; variedad, que contempla formatos estructurados, semiestructurados y no estructurados; y veracidad, que se refiere a la calidad, confiabilidad e integridad de los datos capturados (Ali et al., 2024). La arquitectura de Big Data se apoya en tecnologías distribuidas como Hadoop, Spark y sistemas de almacenamiento escalables, que permiten la paralelización de procesos y la persistencia eficiente de grandes cantidades de información. Apache Spark, por ejemplo, se utiliza para análisis predictivos y aprendizaje automático escalable, siendo capaz de manejar flujos de datos en tiempo real y fuentes heterogéneas (Bergui et al., 2024). Desde una perspectiva analítica, Big Data permite integrar y modelar datos provenientes de múltiples dominios como salud, energía, transporte o redes sociales, facilitando el análisis de patrones ocultos, predicciones de eventos futuros y optimización de decisiones estratégicas. Esta capacidad transforma los modelos descriptivos retrospectivos en sistemas predictivos proactivos, lo que resulta esencial en contextos como las ciudades inteligentes, salud pública y sistemas industriales.

Modelos predictivos

Los modelos predictivos constituyen herramientas matemáticas y computacionales

que permiten anticipar eventos futuros mediante el análisis de patrones presentes en datos históricos, su principal función es generar inferencias cuantitativas sobre variables objetivo a partir de variables explicativas, permitiendo orientar decisiones estratégicas en condiciones de incertidumbre. En contextos complejos, como salud pública, energía, logística o redes sociales, estos modelos son fundamentales debido a la alta dimensionalidad, no linealidad y estacionalidad de los datos (Zhang et al., 2025). Existen diversos tipos de modelos predictivos; la regresión lineal, utilizada para relaciones continuas simples; la regresión logística, para clasificaciones binarias; árboles de decisión y bosques aleatorios (random forest), que capturan interacciones no lineales; y modelos de redes neuronales profundas como LSTM (Long Short-Term Memory) y CNN (Convolutional Neural Networks), adecuados para series temporales y datos espaciales complejos (Bui et al., 2024; Cheng y Yue, 2024). La evaluación del rendimiento de estos modelos se basa en métricas cuantitativas como el Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el Porcentaje de Error Absoluto Medio (MAPE), las cuales permiten comparar la precisión y robustez de los algoritmos bajo distintas condiciones (Azeez et al., 2024). La selección del modelo adecuado implica un balance entre interpretabilidad, eficiencia computacional y capacidad de generalización, siendo común la implementación de modelos híbridos que combinan varios enfoques para mejorar la precisión y adaptabilidad en sistemas dinámicos.

Integración de algoritmos de aprendizaje automático en Big Data

La integración de algoritmos de aprendizaje automático (machine learning) en entornos de Big Data permite transformar grandes

volúmenes de datos en conocimiento procesable, optimizando la toma de decisiones automatizadas en contextos dinámicos. El aprendizaje automático se clasifica en supervisado y no supervisado, el primero requiere datos etiquetados para predecir variables de salida, mientras que el segundo identifica patrones ocultos sin conocimiento previo (Jamarani et al., 2024). En Big Data, la escalabilidad es esencial, técnicas ensemble como boosting, bagging y stacking mejoran la precisión al combinar múltiples modelos débiles. Por ejemplo, random forest (bagging) y XGBoost (boosting) son ampliamente utilizados por su eficiencia y precisión. Modelos secuenciales como LSTM y redes convolucionales (CNN) son aplicados en series temporales multivariadas, particularmente en predicciones de calidad ambiental, movilidad urbana y salud pública (Bui et al., 2024; Cheng y Yue, 2024). Apache Spark y Hadoop proporcionan la infraestructura distribuida necesaria para entrenar estos algoritmos en grandes volúmenes de datos. Spark MLlib, por ejemplo, permite la integración de pipelines de aprendizaje automático para tareas de clasificación, regresión y clustering sobre datos en streaming o batch (Ali et al., 2024). La implementación exitosa depende del diseño de arquitecturas que consideren el preprocesamiento de datos, la selección de características y la evaluación continua del rendimiento del modelo, garantizando una adaptación robusta ante cambios contextuales y concept drift.

Ventajas estratégicas de la convergencia entre modelos predictivos y Big Data

La convergencia entre Big Data y modelos predictivos proporciona una ventaja estratégica sustancial al permitir la planificación anticipada basada en patrones emergentes. Esta sinergia posibilita la optimización de recursos, la toma

de decisiones informadas y la personalización de servicios en tiempo real (Kumar et al., 2024). Permite, además, la detección temprana de anomalías y eventos críticos, lo cual resulta esencial para la mitigación de riesgos en sectores como salud, energía y logística. La capacidad de modelar comportamientos dinámicos mejora la resiliencia organizacional y potencia sistemas de alerta y predicción con alta capacidad de adaptación.

Materiales y Métodos

El presente trabajo corresponde a una revisión sistemática de carácter descriptivo y analítico. Este tipo de estudio permitió compilar, evaluar y sintetizar de forma rigurosa la literatura científica disponible sobre la aplicación de modelos de predicción en entornos de Big Data (Silador, 2023). La finalidad fue identificar patrones metodológicos, métricas de desempeño y áreas de aplicación relevantes, con el propósito de establecer el grado de efectividad predictiva en contextos caracterizados por grandes volúmenes de datos heterogéneos, multitemporales y dinámicos. En el proceso metodológico de la investigación, se siguieron los lineamientos establecidos por el protocolo PRISMA 2020 Preferred Reporting Items for Systematic Reviews and Meta-Analyses, el cual proporcionó una estructura formal para el desarrollo de cada fase de la revisión (Yáñez et al., 2021).

El uso de este marco metodológico facilitó la transparencia del proceso, desde la identificación inicial de artículos hasta la síntesis final de resultados. La búsqueda de fuentes se llevó a cabo en bases de datos académicas de alto impacto: Scopus, ScienceDirect, Dialnet y Scielo, además se estableció como criterio temporal de inclusión el periodo comprendido entre mayo y junio de 2025 y solo se consideraron artículos escritos en inglés o español, revisados por pares, con

acceso completo al texto y que cumplieran con estándares metodológicos verificables. La estrategia de búsqueda se diseñó con base en descriptores normalizados y operadores booleanos, los algoritmos utilizados fueron: ("predictive models" AND "Big Data") ("large scale data" AND "Big Data") ("predictive analytics" AND "Big Data"). Dentro de los criterios de inclusión, se encuentran:

- Artículos empíricos sobre el diseño, evaluación o aplicación de modelos de predicción en entornos Big Data.
- Documentos que integren tecnologías como redes neuronales, árboles de decisión, aprendizaje automático o modelos híbridos o asistidos.
- Publicaciones revisadas por pares y de acceso abierto.
- Estudios realizados entre el 2020 hasta el 2025.
- Estudios desarrollados en inglés y español.

Por otro lado, dentro de los criterios de exclusión, se encuentran:

- Trabajos centrados en modelos sin vínculo con infraestructuras de Big Data.
- Revisiones narrativas sin sistematización documental.
- Publicaciones no accesibles en texto completo o en idiomas distintos al inglés y español.
- Documentos duplicados o redundantes entre bases de datos.
- Artículos con insuficiente reporte metodológico o sin indicadores de validez interna.

La búsqueda bibliográfica se ejecutó entre el 1 de mayo y el 30 de junio de 2025, inicialmente, se aplicaron filtros automáticos por año de publicación, idioma y tipo de documento; posteriormente, se eliminaron duplicados

mediante el gestor bibliográfico Mendeley. La revisión de títulos y resúmenes fue realizada por un evaluador independiente, quien aplicó los criterios de inclusión y exclusión de forma paralela. En los casos con discrepancias, se generó una comparación entre pares para lograr un consenso. Los artículos seleccionados fueron leídos para confirmar su elegibilidad, cada fase fue documentada conforme al diagrama de flujo PRISMA, asegurando transparencia, reproducibilidad y trazabilidad del proceso. El análisis de rigor metodológico se basó en los ítems del Check List PRISMA 2020 ajustados a la necesidad investigativa (Prisma, 2020). Este instrumento permitió verificar el cumplimiento de estándares en aspectos como claridad del objetivo, definición de criterios de selección, transparencia en el proceso de búsqueda y selección, evaluación del sesgo y consistencia en la presentación de resultados. Cada artículo fue sometido a una revisión crítica independiente. Se excluyeron aquellos que no justificaron adecuadamente su diseño u omitieron el reporte de métricas, esta depuración garantizó la validez interna de la revisión y la confiabilidad de las conclusiones derivadas del análisis.

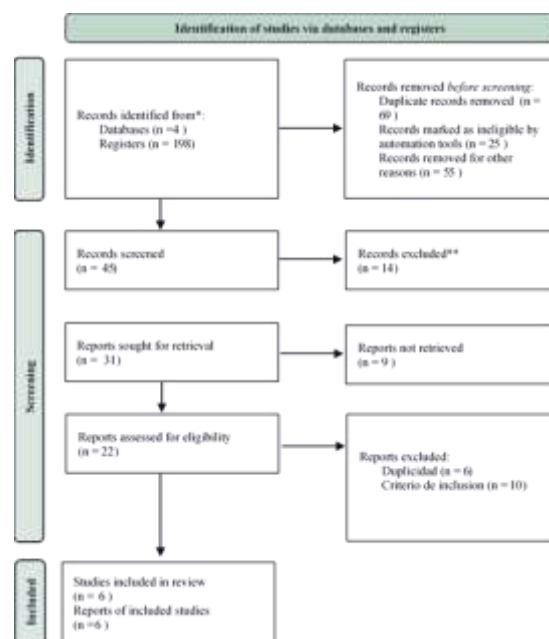


Figura 1. Diagrama de Flujo - PRISMA

Resultados y Discusión

Los hallazgos recopilados evidencian la evolución sincrónica entre los modelos predictivos y la tecnología Big Data en diversos contextos complejos, donde la precisión analítica se ve fortalecida por la capacidad de procesamiento de grandes volúmenes de datos. Las investigaciones analizadas destacan cómo algoritmos avanzados, como redes neuronales,

árboles de decisión y modelos híbridos, aprovechan infraestructuras de datos masivos para generar predicciones altamente confiables. Esta convergencia metodológica no solo potencia la eficiencia predictiva, sino que redefine los alcances, beneficios y desafíos del análisis avanzado orientado a la toma de decisiones estratégicas. (ver tabla 1):

Tabla 1. Resultados obtenidos

No.	Autor y Año	País	Tema	Precisión de las predicciones en contextos complejos	Interrelación entre los modelos predictivos y la tecnología Big Data	Alcance, limitaciones y beneficios
1	(Waghodekar & Bewoor, 2025)	India	<i>Big Data and Blockchain Synergies in Healthcare: Opportunities for Enhanced Decision Making and Operational Management</i>	La integración de redes neuronales y aprendizaje federado optimiza la detección temprana de enfermedades y la asignación de recursos clínicos con una precisión del 90%	Los algoritmos predictivos se nutren de flujos masivos descentralizados a través de <i>Blockchain</i> , mejorando trazabilidad y seguridad	Alcance: predicción diagnóstica. Limitaciones: interoperabilidad regulatoria. Beneficios: reducción del fraude, mejora en tiempos de respuesta y seguridad de datos.
2	(Zhang et al., 2025)	Reino Unido	<i>Design-Based Predictive Inference</i>	Introduce estimaciones predictivas individualizadas con validez inferencial en poblaciones finitas usando diseño probabilístico	Integra modelos de aprendizaje con muestreo probabilístico para garantizar consistencia en escenarios con datos incompletos	Alcance: transporte urbano sostenible. Limitaciones: complejidad de calibración del modelo. Beneficios: reducción de congestión y diseño de rutas óptimas.
3	(Kumar et al., 2024)	India	<i>Predictive Analytics for Mortality: FSRNCA-FLANN Modeling Using Public Health Inventory Records</i>	Aplica redes neuronales adaptativas sobre registros públicos de salud, logrando precisión superior al 95% en escenarios multivariados	Big Data aporta secuencias históricas robustas para alimentar el modelo FLANN, maximizando la capacidad predictiva	Alcance: monitoreo poblacional. Limitaciones: escalabilidad computacional. Beneficios: predicción preventiva y reducción de carga.
4	(Azeez et al., 2024)	Nigeria	<i>A Predictive Model for Benchmarking the Performance of Algorithms for Fake and Counterfeit News Classification in Global Networks</i>	Los modelos de ensamblaje mejoran la precisión de clasificación hasta en un 93% sobre <i>datasets</i> de gran escala	Big Data provee el corpus masivo multilingüe y diverso sobre el cual los modelos incrementan su capacidad de generalización	Alcance: análisis multirregional. Limitaciones: sesgo en calidad de datos. Beneficios: mejora en inferencia predictiva y reducción de incertidumbre.
5	(Boudali et al., 2024)	Francia	<i>A predictive approach for myocardial infarction risk assessment using machine learning and big clinical data</i>	La combinación de <i>LightGBM</i> con ingeniería de características logró identificar factores de riesgo relevantes y anticipar el infarto con alta precisión sobre datos clínicos desequilibrados	Big Data clínico proporciona secuencias históricas estructuradas que optimizan el aprendizaje automático en escenarios sanitarios reales	Alcance: prevención poblacional. Limitación: desbalance de clases. Beneficio: predicción robusta desde múltiples atributos.
6	(Rezaul et al., 2025)	Estados Unidos	<i>A Comparative Study of Predictive Analysis Using Machine Learning Techniques: Performance Evaluation of Manual and AutoML Algorithms</i>	Redes neuronales multicapa aplicadas a registros ambientales permiten clasificar riesgos ecológicos con alta sensibilidad	La plataforma Big Data soporta integración de imágenes satelitales, sensores ambientales y datos históricos para alimentar los modelos	Alcance: planificación hidráulica. Limitación: falta de etiquetas en tiempo real. Beneficio: respuesta temprana a eventos críticos.

Fuente: elaboración propia

Para analizar los estudios, se pueden demostrar varias similitudes metodológicas, como el uso de modelos predictivos con aprendizaje automático, redes neuronales y algoritmos de conjunto, todos destinados a maximizar la

precisión en contextos de alta complejidad. Esta convergencia de hallazgos define una tendencia general hacia el uso de marcos no lineales y multivariantes para resolver problemas heterogéneos como la predicción de

enfermedades, la previsión del consumo de agua, el control del tráfico aéreo y terrestre, el comportamiento del mercado financiero y el cambio climático, entre otros. No obstante, se destacan contrastes significativos en cuanto a los dominios de aplicación, la estructura de datos utilizados y las condiciones de validación. Mientras Kumar et al. (2024) y Boudali et al. (2024) trabajan con datos clínicos estructurados y homogéneos, Jiménez (2024) y Fournier et al. (2024), operan en entornos institucionales, donde la variabilidad y diversidad de las fuentes demanda enfoques de integración más robustos. Asimismo, Waghodekar y Bewoor (2025) articulan modelos predictivos con Blockchain, diferenciándose del resto al enfatizar la descentralización y trazabilidad como pilares para aumentar la precisión y confianza en las decisiones clínicas. Por otro lado, Hua (2025) y Kok et al. (2024), integran variables exógenas, revelando cómo la predicción efectiva depende no solo del modelo utilizado sino también del grado de contextualización que se logra al incorporar factores externos.

A diferencia de esto, Zhang et al. (2025) optan por un enfoque de inferencia predictiva bajo diseño probabilístico, privilegiando la validez inferencial antes que la robustez empírica, lo cual representa un contraste metodológico con el resto de estudios que emplean aprendizaje supervisado o automático. Un elemento transversal a todos los estudios es la sinergia entre modelos predictivos y tecnología Big Data, que posibilita no solo el acceso a grandes volúmenes de datos históricos, sino también el procesamiento en tiempo real y la escalabilidad de los modelos, no obstante, el grado de integración y la sofisticación de las arquitecturas empleadas varía ampliamente, desde entornos distribuidos hasta plataformas colaborativas en la nube como Google Colab como lo exponen Cevallos et al. (2024) y Sariyer

et al. (2024), quienes han identificado diferencias en accesibilidad y requerimientos computacionales. Por último, la reflexión que surge de tal contraste es que la efectividad de los modelos predictivos está condicionada por su calidad, diversidad y estructura de los datos, así como por las decisiones estratégicas del investigador al determinar el alcance, los algoritmos y la infraestructura tecnológica. Esto ilustrará al lector con el hecho de que entender la lógica del modelo, el contexto de los datos y la evaluación a realizar es tan importante como la exactitud de los cálculos que se han efectuados.

Conclusiones

En consecuencia, se observa que la integración del modelado predictivo y la tecnología de Big Data no es simplemente una innovación técnica, más bien, representa una profunda transformación en los procesos de análisis y toma de decisiones en entornos altamente complejos. La capacidad de prever comportamientos, identificar patrones emergentes y adaptar respuestas ha pasado de ser posibilidades teóricas a necesidades operativas en áreas como la salud, el medio ambiente, el transporte y las finanzas. Esta convergencia no puede evaluarse únicamente desde la perspectiva de sus métricas de rendimiento, necesita ser entendida como un nuevo paradigma que altera la interacción entre datos, conocimiento y acción. Esta investigación demuestra que la efectividad de un modelo no está definida exclusivamente por su arquitectura algorítmica, sino más bien, la estrategia depende de un diseño algorítmico balanceado, comprensión contextual, calidad de los datos y selección de métodos. Se identifican dimensiones críticas como la automatización, adaptabilidad y escalabilidad; así como la gobernanza sobre la infraestructura y los flujos de datos propulsores. La precisión de un modelo

no solo es una función de su sofisticación, sino también de qué tan bien se adapta a las necesidades reales, capacidades institucionales y marcos éticos robustos. Al integrar varios enfoques metodológicos y contextos de aplicación, queda claro que el potencial predictivo se ve reforzado cuando se incluye el componente humano, a saber, la capacidad del analista para interpretar grandes datos, tomar decisiones y ajustarse a múltiples factores. Este estudio no concluye con un solo modelo o una técnica ideal, sino más bien con una invitación a repensar el papel del conocimiento aplicado en entornos inciertos, donde una combinación de tecnología, teoría y juicio experto ofrece valor en la era del análisis de datos. De esta manera, el conocimiento se transforma en acción inteligente y la capacidad de anticipar se convierte en una herramienta estratégica para crear futuros posibles y sostenibles.

Referencias Bibliográficas

- Ali, H., Alham, M., & Ibrahim, D. (2024). Big data resolving using Apache Spark for load forecasting and demand response in smart grid: a case study of Low Carbon London Project. *Journal of Big Data*, 11. <https://doi.org/10.1186/s40537-024-00909-6>
- Azeez, N., Misra, S., Ogaraku, D., & Abidoye, A. (2024). A Predictive Model for Benchmarking the Performance of Algorithms for Fake and Counterfeit News Classification in Global Networks. *Sensors*, 24(17). <https://doi.org/10.3390/s24175817>
- Bedoya, J., Góngora, C., García, M., & Ruiz, L. (2024). Usos del Big Data en las empresas: Un instrumento de prestigio y de supervivencia hoy. *Dominio de las Ciencias*, 10(2), 1024–1042. <https://doi.org/10.23857/dc.v10i2.3843>
- Bergui, M., Hourri, S., Najah, S., & Nikolov, N. S. (2024). Predictive modelling of MapReduce job performance in cloud environments using machine learning techniques. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-00964-z>
- Boudali, I., Chebaane, S., & Zitouni, Y. (2024). A predictive approach for myocardial infarction risk assessment using machine learning and big clinical data. *Healthcare Analytics*, 5, 100319. <https://doi.org/10.1016/j.health.2024.100319>
- Bui, L., Tran, D., & Nguyen, D. (2024). Prediction of the river water environment carrying capacity using LSTM networks. *Water Supply*, 27(7), 2518–2533. <https://doi.org/10.2166/ws.2024.138>
- Cevallos, E., Jacho, A., & Córdova, A. (2024). Big Data y Analítica Predictiva en la Toma de Decisiones Empresariales. *Revista Ingenio Global*, 3(2), 55–72. <https://doi.org/10.62943/rig.v3n2.2024.103>
- Cheng, C., & Yue, W. (2024). Construction of a Green and Low-Carbon Travel Order Prediction Model Based on Shared Bicycle Big Data. *International Journal of Computational Intelligence Systems*, 17(1). <https://doi.org/10.1007/s44196-024-00519-3>
- Cisneros, A., Guevara, A., Urdánigo, J., & Garcés, J. (2022). Técnicas e Instrumentos para la Recolección de Datos que Apoyan a la Investigación Científica en Tiempo de Pandemia. *Dominio de las Ciencias*, 8(1), 1165–1185. <https://doi.org/10.23857/dc.v8i1.2546>
- Cong, Y., & Inazumi, S. (2024). Integration of Smart City Technologies with Advanced Predictive Analytics for Geotechnical Investigations. *Smart Cities*, 7(3), 1089–1108. <https://doi.org/10.3390/smartcities7030046>
- Fournier, C., Fernández, R., Cirés, S., López, J. Besada, E., & Quesada, A. (2024). LSTM networks provide efficient cyanobacterial blooms forecasting even with incomplete spatio-temporal data. *Water Research*, 267. <https://doi.org/10.1016/j.watres.2024.122553>
- Hua, Z. (2025). Data Visualization and Prediction Model Analysis. *ITM Web of Conferences*, 03003(11), 1–11.

- <https://doi.org/10.1051/itmconf/20257003003>
Jamarani, A., Haddadi, S., Sarvzadeh, R., Haghi Kashani, M., Akbari, M., & Moradi, S. (2024). Big data and predictive analytics: A sytematic review of applications. In *Artificial Intelligence Review* 57 (7). Springer Netherlands.
<https://doi.org/10.1007/s10462-024-10811-5>
- Jiménez, P. (2024). Modelos híbridos de predicción para Big Data Streaming. Universidad Pablode Olavide.
<https://www.educacion.gob.es/teseo/imprimirFicheroTesis.do?idFichero=0I5spK2nYcs%3D>
- Kok, C. L., Ho, C., Chen, L., Koh, Y., & Tian, B. (2024). A Novel Predictive Modeling for Student Attrition Utilizing Machine Learning and Sustainable Big Data Analytics. *Applied Sciences (Switzerland)*, 14(21).
<https://doi.org/10.3390/app14219633>
- Kumar, M., Kumar Singh, S., & Kim, S. (2024). Predictive Analytics for Mortality: FSRNCA-FLANN Modeling Using Public Health Inventory Records. *IEEE Access*, 12(May), 81252–81264.
<https://doi.org/10.1109/ACCESS.2024.3411162>
- Moreno, S. (2022). *Big data e inteligencia artificial, dos herramientas clave en procesos administrativos para instituciones de educación superior*. Universidad Santo Tomás – Seccional Tunja, 1–18.
<https://repository.usta.edu.co/server/api/core/bitstreams/1a99432f-3dd3-43ad-988d-33687cbeac69/content>
- Prisma. (2020). Elementos de Reporte Preferidos para Revisiones Sistemáticas y Metaanálisis. <https://www.prisma-statement.org/>
- Rezaul, K., Jewel, M., Sudhan, A., Khan, M., Fernando, M., Siddiquee, K., Jannat, T., Rahman, M., & Islam, M. (2025). A Comparative Study of Predictive Analysis Using Machine Learning Techniques: Performance Evaluation of Manual and AutoML Algorithms. *International Journal of Advanced Computer Science and Applications*, 16(1), 12–31.
<https://doi.org/10.14569/IJACSA.2025.0160102>
- Sariyer, G., Kumar Mangla, S., Chowdhury, S., Erkan Sozen, M., & Kazancoglu, Y. (2024). Predictive and prescriptive analytics for ESG performance evaluation: A case of Fortune 500 companies. *Journal of Business Research*, 181, 114742.
<https://doi.org/10.1016/j.jbusres.2024.114742>
- Silador, R. (2023). Manual de Investigación. Instituto Superior Tecnológico Universitario.
<https://tecnologicolezaeta.edu.ec/wp-content/uploads/2023/09/MANUAL-DE-INVESTIGACION-2023-1.pdf>
- Waghodekar, P., & Bewoor, M. S. (2025). Big Data and Blockchain Synergies in Healthcare: Opportunities for Enhanced Decision-Making and Operational Management. *Journal of Neonatal Surgery*, 14(4), 255–265.
<https://doi.org/10.52783/jns.v14.1780>
- Yáñez, C., Aguilera, R., Fuentes, H., & Videla, Á. (2021). Importancia de la directriz PRIMAS. *Nutrición Hospitalaria*, 38(3), 677–678. <https://doi.org/10.20960/nh.04616>
- Zhang, L., Sanguiao, L., & Lee, D. (2025). Design-Based Predictive Inference. *Journal of Official Statistics*, 41(1), 404–432.
<https://doi.org/10.1177/0282423X241277719>



Esta obra está bajo una licencia de Creative Commons Reconocimiento-No Comercial 4.0 Internacional. Copyright © Mario Alberto Ibarra Martínez, Edgar Xavier Mendoza Arce y Wilson Javier Romero Berrones.

