***Tipo de Artículo:*** *Artículo de revisión*
***Área:*** *Ingeniería*

# Impacto de las estrategias de preprocesamiento de datos en la precisión de algoritmos de regresión lineal

## *Impact of data preprocessing strategies on the accuracy of linear regression algorithms*

**Jaime David Camacho Castillo[1], Mauricio Alexander Álvarez Ortega[2], Iván Alejandro Daqui Cabezas[3], Alex Nicolas Mazacon Guamingo[4], Juan Enrique Olivo Quintero[5], Jordy Joel Vele Carrera[6]**

**Cita:**
Jaime David Camacho Castillo et al. (2026). Impacto de las estrategias de preprocesamiento de datos en la precisión de algoritmos de regresión lineal. *TESLA Revista Científica*, 6(2), e724. https://doi.org/10.55204/trc.v6i2.e724

**Recibido:** 13/05/2026
**Revisado:** 27/05/2026–17/06/2026
**Corregido:** 24/06/2026
**Aceptado:** 03/07/2026
**Publicado:** 08/09/2026



[1] Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.
[2–6] Investigador externo, Ecuador.

[1] jaimed.camacho@espoch.edu.ec, [2] maalvarezo2001@gmail.com, [3] iadaquic05@gmail.com, [4] alexmazacon09@gmail.com, [5] enriqueolivo2001712@gmail.com, [6] jordyvele777@gmail.com

[1] 0000-0002-9110-6585, [2] 0009-0005-4850-9425, [3] 0009-0000-0146-5035, [4] 0009-0009-2312-2936, [5] 0009-0002-3827-5723, [6] 0009-0004-1180-9347

**Resumen:** El preprocesamiento de datos es una etapa determinante en el rendimiento de los modelos de regresión lineal, donde la calidad de los datos, incluyendo el tratamiento de valores faltantes, la detección de outliers, la normalización y la codificación de variables, influyen directamente en la capacidad predictiva de los algoritmos de aprendizaje automático. Esta revisión sistemática, conducida bajo la metodología PRISMA 2020 (Preferred Reporting Items for Systematic reviews and Meta-Analyses), bajo esta misma óptica, se analiza 40 artículos científicos publicados entre 2021 y 2026 en bases de datos reconocidas como Scopus, Web of Science, Springer Nature, MDPI, IEEE y ScienceDirect.
Dentro de este contexto, se destaca que los modelos de regresión lineal constituyen una herramienta fundamental en la inteligencia artificial y el análisis de datos, donde la calidad del preprocesamiento influye directamente en la precisión de los resultados obtenidos. En el desarrollo del estudio, se analizaron técnicas clave como la limpieza de datos, la imputación de valores faltantes, la normalización o estandarización y a esto se suma la codificación de variables categóricas.
En cuanto a las aplicaciones prácticas y futuras líneas de investigación, los hallazgos permiten su implementación en modelos predictivos relacionados con ventas, estimación de precios e indicadores académicos, como también, plantear la extensión del análisis hacia modelos no lineales y conjuntos de datos de mayor dimensionalidad. No obstante, un adecuado preprocesamiento de datos mejora significativamente la precisión predictiva de los modelos de regresión lineal, contribuyendo a reducir problemas de overfitting y underfitting.

**Palabras clave:** *preprocesamiento de datos, regresión lineal, normalización, imputación, aprendizaje automático, revisión sistemática.*

**Abstract:** Data preprocessing is a crucial stage in the performance of linear regression models, where data quality including the handling of missing values, outlier detection, normalization, and variable encoding directly influences the predictive capability of machine learning algorithms. This systematic review, conducted under the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology, analyzes 40 scientific articles published between 2021 and 2026 in well-recognized databases such as Scopus, Web of Science, Springer Nature, MDPI, IEEE, and ScienceDirect.
Within this context, linear regression models are highlighted as a fundamental tool in artificial intelligence and data analysis, where the quality of preprocessing directly affects the accuracy of the results obtained. During the development of the study, key techniques were examined, including data cleaning, missing value imputation, normalization or standardization, as well as the encoding of categorical variables.
Regarding practical applications and future research directions, the findings support their implementation in predictive models related to sales, price estimation, and academic indicators. They also encourage extending the analysis to non-linear models and higher-dimensional datasets. Ultimately, proper data preprocessing significantly improves the predictive accuracy of linear regression models, helping to reduce issues such as overfitting and underfitting.

**Keywords:** *data preprocessing, linear regression, normalization, imputation, machine learning, systematic review.*

## INTRODUCCIÓN

En los últimos años, el aprendizaje automático ganó terreno en prácticamente todas las áreas del conocimiento. Los modelos de regresión lineal están entre los algoritmos más usados para predicción cuantitativa, con aplicaciones que van desde la estimación de precios hasta el análisis de rendimiento estudiantil (Pandit et al., 2021; Chowdhury et al., 2022). Pero la precisión de estos modelos no depende solo del algoritmo: depende, antes que nada, de cómo se preparan los datos. El preprocesamiento convierte datos crudos en información útil para el modelado, garantizando consistencia, completitud y relevancia (Maharana et al., 2022).

La literatura científica reciente ha documentado de manera extensa cómo las distintas estrategias de preprocesamiento impactan en el rendimiento de los modelos de regresión. Técnicas como la normalización min-max, la estandarización z-score, la imputación de valores faltantes mediante múltiples métodos estadísticos, la codificación de variables categóricas y la detección y tratamiento de valores atípicos han demostrado influir de manera significativa en métricas de evaluación como el error cuadrático medio (MSE), el error absoluto medio (MAE) y el coeficiente de determinación ($R^2$) (Ahsan et al., 2021; Maharana et al., 2022; Demir y Sahin, 2024). Además, el modelo de regresión lineal se fundamenta en una relación matemática entre variables independientes y dependientes, expresada mediante la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_n x_n + \varepsilon$$

Donde y representa la variable dependiente, $x_i$ las variables independientes, $\beta_i$ los coeficientes del modelo que determinan la influencia de cada variable, y $\varepsilon$ el término de error. La estimación precisa de estos coeficientes depende directamente de la calidad del preprocesamiento de los datos, ya que valores atípicos, escalas inconsistentes o datos faltantes pueden sesgar significativamente los resultados del modelo (Mallikharjuna Rao et al., 2023; Wanyonyi & Masinde, 2025).

A pesar de la abundante producción científica sobre este tema, no existe una guía unificada que permita identificar cuáles estrategias de preprocesamiento resultan más efectivas en el contexto específico de la regresión lineal. La presente investigación busca cubrir este vacío mediante la realización de una revisión sistemática bajo la metodología PRISMA 2020, la cual establece un estándar internacional para el reporte transparente y reproducible de revisiones sistemáticas y metaanálisis (Page et al., 2021a).

El problema de investigación se centra en determinar de qué manera las diferentes estrategias de preprocesamiento de datos impactan en la precisión de los modelos de regresión lineal, y cuáles de ellas resultan más efectivas según la evidencia científica disponible en el período 2021–2026. El objetivo general de este trabajo es analizar sistemáticamente el impacto de las estrategias de preprocesamiento de datos en la precisión de algoritmos de regresión lineal, a través de la revisión, clasificación y síntesis de publicaciones científicas indexadas en bases de datos de alto impacto (Koukaras & Tjortjis, 2025; Mu et al., 2026).

**Técnicas de Escalado y Normalización:** El escalado de características es una de las etapas más estudiadas en la literatura. Ahsan et al. (2021) demostraron que la normalización min-max mejora modelos sensibles a la magnitud de variables, mientras que la estandarización z-score es más robusta ante outliers. Cabello-Solorzano et al. (2023) confirmaron que elegir la técnica correcta puede incrementar la precisión hasta un 15%. Mahmud Sujon et al. (2024) precisaron que la estandarización es preferible en regresión lineal cuando los datos no siguen una distribución uniforme. Pinheiro et al. (2025) añaden que omitir el escalado puede degradar el $R^2$ en más de un 20%.

**Imputación de Valores Faltantes:** El tratamiento de valores faltantes es otro desafío crítico. Von Hippel y Bartlett (2021) propusieron la imputación múltiple por máxima verosimilitud como alternativa eficiente a los métodos bayesianos, con errores estándar consistentes. Li et al. (2023) desarrollaron un algoritmo noise-aware que reduce el MSE hasta un 18% en datos meteorológicos. Gan et al. (2023) y Ou et al. (2024) lograron superar a los métodos clásicos en hasta un 23% con métodos híbridos que combinan Random Forest con redes generativas adversariales. Buczak et al. (2023) mostraron que el patrón de missingness, MCAR o MAR, determina cuál método es el más adecuado.

**Detección y Tratamiento de Outliers:** Los valores atípicos distorsionan los coeficientes del modelo porque el método OLS es sensible a observaciones extremas. Hassanat et al. (2024) propusieron una métrica robusta basada en distancias no convexas, superior al RMSE en datos contaminados. Rasyidah et al. (2023) lograron mejoras de hasta un 12% en $R^2$ con un método de limpieza basado en conjuntos rugosos. Amini et al. (2024) demostraron que identificar correctamente la estructura del ruido mejora la estimación de parámetros. El hallazgo más contundente es de Habib y Okayli (2024): combinar eliminación de outliers con normalización redujo el MAE en un 31%.

**Codificación de Variables Categóricas:** La codificación de variables categóricas es clave cuando el dataset incluye atributos no numéricos. Bolikulov et al. (2024) encontraron que el target encoding y el one-hot encoding regulado superan a la codificación ordinal simple. Pargent et al. (2022) demostraron que el target encoding regularizado reduce el sobreajuste en variables de alta cardinalidad. Zeng y Tao (2023) advirtieron que una transformación inadecuada sesga los coeficientes estimados. De Amorim et al. (2023) señalaron que la técnica de escalado no es independiente del regresor utilizado, y debe tratarse como un hiperparámetro del pipeline (Mallikharjuna Rao et al., 2023).

**Tabla comparativa de técnicas:** Una observación fundamental es que las siete técnicas identificadas no se aplican de forma simultánea ni acumulativa dentro de una misma categoría. En cambio, se organizan en pares alternativos: dentro de cada categoría de preprocesamiento, el investigador debe elegir una técnica u otra según las características del dataset. Aplicar ambas técnicas del mismo par no tiene sentido metodológico, ya que resuelven el mismo problema de formas distintas (Mallikharjuna Rao et al., 2023; Koukaras & Tjortjis, 2025).

La única categoría sin par es la eliminación de outliers, que cuenta con una sola técnica en la tabla. Las demás categorías presentan exactamente dos alternativas cada una, diferenciadas por su complejidad, sus supuestos estadísticos y el contexto en que son más efectivas. La selección incorrecta dentro de un par puede introducir sesgos o degradar el rendimiento del modelo, razón por la cual la evidencia científica reciente enfatiza la importancia de elegir la técnica según las propiedades estadísticas del dataset (Demir & Sahin, 2024; Habib & Okayli, 2024).

**Tabla 1**

**Comparación de técnicas de preprocesamiento**

| TÉCNICA | VENTAJA | CUANDO USAR | DIFERENCIA CON SU ALTERNATIVA |
|---|---|---|---|
| ESCALADO | | | |
| Normalización Min-Max | Escala uniforme entre 0 y 1 | Datos sin outliers | Usa el mínimo y máximo del dataset; sensible a outliers extremos. Si un valor atípico distorsiona el rango, toda la escala se ve afectada. |
| Estandarización Z-score | Robusta a outliers | Datos con distribución no uniforme | Usa media y desviación estándar; robusta a outliers. Valores extremos no distorsionan la escala. **X** Preferida en regresión lineal |
| IMPUTACIÓN | | | |
| Imputación por media/mediana | Simple y rápida | Datos con pocos valores faltantes | Reemplaza NaN con un único valor fijo; puede introducir sesgo si hay muchos faltantes o el patrón es MAR. Recomendada solo con <5% de faltantes. |
| Imputación múltiple | Mayor precisión | Datos complejos | Genera múltiples versiones del dataset imputado y las combina; preserva la distribución estadística. Más costosa pero superior en patrones MAR/MNAR. **X** Mayor precisión |

| TÉCNICA | VENTAJA | CUANDO USAR | DIFERENCIA CON SU ALTERNATIVA |
|---|---|---|---|
| OUTLIERS | | | |
| Eliminación de outliers | Mejora precisión | Datos con ruido | Se aplica o no según si el dataset contiene valores atípicos confirmados (regla IQR: absences > Q3 + 1.5×IQR). |
| CODIFICACIÓN | | | |
| One-hot encoding | Evita sesgo ordinal | Variables categóricas | Crea una columna 0/1 por cada categoría; aumenta la dimensionalidad. Con muchas categorías (alta cardinalidad) puede perjudicar al modelo. |
| Target encoding | Reduce dimensionalidad | Alta cardinalidad | Reemplaza cada categoría por la media del target (G3) en ese grupo; captura la relación real con la variable objetivo sin aumentar columnas. **X** Mayor impacto en el ejemplo |

*Nota: Elaboración propia con base en la información analizada.*

Esta estructura de pares tiene una implicación directa para el diseño del ejemplo práctico: en lugar de aplicar solo una técnica por categoría, se puede comparar ambas alternativas del par y determinar cuál produce un mayor $R^2$ en el dataset Student Performance.

## METODOLOGÍA

La presente investigación adopta un enfoque documental de carácter cuantitativo-analítico, sustentado en la metodología PRISMA 2020 (Preferred Reporting Items for Systematic reviews and Meta-Analyses). Esta elección metodológica responde a la necesidad de garantizar transparencia, reproducibilidad y completitud en el proceso de búsqueda, selección y síntesis de la evidencia científica disponible sobre el tema de investigación (Page et al., 2021a).

Se formuló la siguiente pregunta de investigación: ¿Qué estrategias de preprocesamiento de datos tienen mayor impacto en la precisión de los algoritmos de regresión lineal según la evidencia científica publicada entre 2021 y 2026?

### Criterios de Búsqueda

Se definieron palabras clave en español e inglés para la búsqueda sistemática, entre ellas: preprocesamiento de datos, data preprocessing, data cleaning, missing values treatment, normalization, standardization, feature scaling, linear regression performance, machine learning preprocessing. Las búsquedas se realizaron en las bases de datos Scopus, Web of Science, Springer Nature, MDPI, ScienceDirect, IEEE Xplore, SciELO y Google Scholar.

### Metodología prisma

La metodología PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) es una guía ampliamente utilizada para desarrollar y presentar revisiones sistemáticas de manera organizada y transparente. Su objetivo es facilitar la selección adecuada de estudios científicos, asegurando que el proceso sea claro, replicable y libre de sesgos.

Este enfoque propone una estructura ordenada que permite buscar, filtrar y analizar la información relevante a partir de criterios previamente definidos. De esta forma, se mejora la calidad de la investigación y se garantiza que los resultados obtenidos estén respaldados por evidencia confiable.

En el presente estudio, se aplicó la metodología PRISMA 2020 para organizar el proceso de revisión en cuatro etapas fundamentales: identificación, cribado, elegibilidad e inclusión. Esto permitió seleccionar de manera rigurosa los artículos más relevantes sobre el impacto del preprocesamiento de datos en la precisión de los modelos de regresión lineal.

## PRISMA

El proceso de selección de estudios se documentó mediante el diagrama de flujo de cuatro fases establecido por PRISMA 2020. A continuación, se describe cada fase con los resultados obtenidos:

Fase 1 – Identificación: Se identificaron inicialmente 187 registros distribuidos en las diferentes bases de datos consultadas: Scopus (38), Web of Science (29), Springer Nature (42), MDPI (51), ScienceDirect (18), IEEE Xplore (5) y otras fuentes (4). Después de eliminar 92 registros duplicados y 15 marcados como inelegibles por herramientas de automatización (Zotero), quedaron 80 registros para la fase de cribado.

Fase 2 – Cribado: Se revisaron los títulos y resúmenes de los 80 registros restantes. Se excluyeron 28 registros por no presentar relación directa con el preprocesamiento de datos aplicado a regresión lineal o por corresponder a áreas de conocimiento no pertinentes (medicina clínica, procesamiento de señales sin relación con ML, etc.). Se solicitó el texto completo de 52 reportes.

Fase 3 – Elegibilidad: Se realizó la lectura completa de los 52 reportes recuperados. Se aplicaron los criterios de inclusión y exclusión definidos. Se excluyeron 12 artículos: 5 por publicación anterior al año 2021, 4 por no reportar métricas de evaluación de modelos de regresión, y 3 por no estar disponibles en texto completo. Quedaron elegibles 40 artículos.

Fase 4 – Inclusión: Se incluyeron en la revisión sistemática los 40 artículos científicos que cumplieron todos los criterios de elegibilidad. Estos artículos constituyen la base bibliográfica de la presente investigación y se encuentran listados en la sección de Referencias bajo el estándar APA 7.

### Criterios de inclusión

Los criterios de inclusión se definieron con el objetivo de garantizar la selección de estudios relevantes, actuales y con rigor científico en relación con el preprocesamiento de datos y su impacto en modelos de regresión lineal. En este sentido, se consideraron los siguientes aspectos:

- **Periodo de publicación:** Se incluyeron únicamente artículos publicados entre los años 2021 y 2026, con el fin de asegurar la actualidad de la evidencia científica analizada.

- **Indexación académica:** Se seleccionaron estudios provenientes de bases de datos reconocidas como Scopus, Web of Science, Springer, MDPI, IEEE y ScienceDirect, garantizando la calidad y validez de las fuentes.
- **Pertinencia temática:** Se incluyeron artículos que abordan directamente el preprocesamiento de datos y su aplicación en modelos de regresión, especialmente regresión lineal.
- **Idioma de publicación:** Se consideraron estudios publicados en inglés y español, debido a su amplia difusión en la comunidad científica.
- **Disponibilidad del contenido:** Solo se incluyeron artículos con acceso a texto completo, permitiendo un análisis detallado de la metodología y resultados.
- **Métricas de evaluación:** Se priorizaron estudios que reportan métricas cuantitativas como MSE, MAE o $R^2$, necesarias para evaluar el rendimiento de los modelos.

### Criterios de exclusión

Los criterios de exclusión se establecieron con el propósito de descartar estudios que no cumplieran con los requisitos metodológicos o temáticos necesarios para la investigación. Estos criterios fueron los siguientes:

- **Antigüedad de publicación:** Se excluyeron artículos publicados antes del año 2021, por no ajustarse al periodo de análisis definido.
- **Falta de indexación:** Se descartaron fuentes no indexadas o que no contaran con revisión por pares, debido a posibles limitaciones en su rigor científico.
- **Irrelevancia temática:** Se excluyeron estudios que no estuvieran relacionados con el preprocesamiento de datos o que no aplicaran modelos de regresión lineal.
- **Idioma no considerado:** Se descartaron artículos en idiomas distintos al inglés o español, por limitaciones en su interpretación y análisis.
- **Acceso restringido:** Se eliminaron estudios que no contaban con disponibilidad de texto completo.
- **Ausencia de métricas:** Se excluyeron artículos que no presentaban métricas cuantitativas de evaluación, lo que impedía medir el impacto del preprocesamiento en los modelos.

### Herramientas Metodológicas

Para el desarrollo de la revisión sistemática se utilizaron las siguientes herramientas: Zotero como gestor bibliográfico para la organización, de duplicación y exportación de referencias en formato APA 7; Rayyan para el cribado colaborativo de artículos por título y resumen; Microsoft Word para la redacción y formateo del documento bajo la plantilla institucional; y el generador oficial PRISMA2020 (paquete R de Haddaway et al., 2022) para la construcción del diagrama de flujo. Se filtraron exclusivamente publicaciones de los últimos cinco años (2021-2026), revisadas por pares, en inglés y español.

## Pipeline de aprendizaje automático

El proceso de construcción de modelos de regresión lineal en el marco de esta investigación se estructura mediante un pipeline de aprendizaje automático, el cual organiza de manera secuencial y sistemática las etapas necesarias para garantizar la calidad de los resultados obtenidos. Este enfoque permite mejorar la precisión del modelo, asegurando la reproducibilidad y consistencia del proceso analítico. El pipeline propuesto comprende las siguientes fases:

- **Recolección de datos:** Consiste en la obtención de datos provenientes de diversas fuentes, tales como bases de datos, repositorios abiertos o sistemas institucionales. En esta etapa es fundamental asegurar la relevancia y representatividad de los datos respecto al problema de estudio.
- **Limpieza de datos:** Se enfoca en la identificación y corrección de errores, inconsistencias, duplicados y registros incompletos. Esta fase garantiza la integridad del dataset antes de su transformación**.**
- **Preprocesamiento de datos:** Constituye una de las etapas más críticas del pipeline, e incluye técnicas como la normalización o estandarización de variables, la imputación de valores faltantes, la detección y tratamiento de valores atípicos (outliers), y la codificación de variables categóricas. Estas transformaciones permiten adaptar los datos a los requerimientos del modelo de regresión lineal.
- **Entrenamiento del modelo:** En esta fase se ajusta el modelo de regresión lineal a los datos preprocesados, estimando los coeficientes que describen la relación entre las variables independientes y la variable dependiente.

- **Evaluación del modelo:** Se analizan los resultados obtenidos mediante métricas de desempeño como el error cuadrático medio (MSE), el error absoluto medio (MAE) y el coeficiente de determinación ($R^2$), con el fin de medir la precisión y capacidad predictiva del modelo.

Este pipeline estructurado permite establecer un flujo de trabajo claro y replicable, facilitando la comparación de resultados entre distintos enfoques de preprocesamiento y contribuyendo a la optimización del rendimiento de los modelos de regresión lineal.

## RESULTADOS Y DISCUSIÓN

A partir de la revisión sistemática de los 40 artículos incluidos, se identificaron cuatro categorías principales de estrategias de preprocesamiento analizadas en la literatura: técnicas de escalado y normalización, métodos de imputación de valores faltantes, detección y tratamiento de outliers, y codificación de variables categóricas. A continuación, se presentan los hallazgos más relevantes de cada categoría.

**Tabla 2**

**Estudios incluidos en la revisión sistemática**

| N° | AUTOR(ES) | AÑO | BASE DE DATOS | TITULO | ENLACE / DOI | PAÍS |
|---|---|---|---|---|---|---|
| 1 | • Pramit Pandit<br>• Prithwiraj Dey<br>• K. N. Krishnamurthy | 2021 | Spriger Nature | Comparative Assessment of Multiple Linear Regression and Fuzzy Linear Regression Models(Pandit et al., 2021) | https://doi.org/10.1007/s42979-021-00473-3 | India |
| 2 | • Md Manjurul Ahsan<br>• M. Parvez Mahmud<br>• Pritom Kumar Saha<br>• Kishor Datta Gupta<br>• Zahed Siddique | 2021 | MDPI | Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance(Ahsan et al., 2021) | https://doi.org/10.3390/technologies9030052 | Estados Unidos |
| 3 | • Jiyan Chen<br>• Zijiang Yang | 2024 | MDPI | Revolutionizing Time Series Data Preprocessing with a Novel Cycling Layer in Self-Attention Mechanisms(Chen & Yang, 2024) | https://doi.org/10.3390/app14198922 | Canadá |
| 4 | • Lucas BV de Amorim<br>• George DC Cavalcanti<br>• Rafael MO Cruz | 2023 | sciencedirect | The choice of scaling technique matters for classification performance(de Amorim et al., 2023) | https://doi.org/10.1016/j.asoc.2022.109924 | Brasil–Canadá |
| 5 | • Selçuk Demir<br>• Emrehan Kutlug Sahin | 2024 | Spriger Nature | The effectiveness of data pre-processing methods on the performance of machine learning techniques using RF, SVR, Cubist and SGB: a study on undrained shear strength prediction(Demir & Sahin, 2024) | https://doi.org/10.1007/s00477-024-02745-9 | Turquía |
| 6 | • Ahmad B. Hassanat<br>• Mohammad Khaled Alqaralleh<br>• Ahmad S. Tarawneh<br>• Khalid Almohammadi<br>• Maha Alamri<br>• Abdulkareem Alzahrani<br>• Ghada A. Altarawneh | 2024 | MDPI | A Novel Outlier-Robust Accuracy Measure for Machine Learning Regression Using a Non-Convex Distance Metric(Hassanat et al., 2024) | https://doi.org/10.3390/math12223623 | Jordania–Arabia Saudita |

| N° | AUTOR(ES) | AÑO | BASE DE DATOS | TITULO | ENLACE / DOI | PAÍS |
|---|---|---|---|---|---|---|
| 7 | • Rauf I. Rauf<br>• Masad A. Alrasheedi<br>• Rasheedah Sadiq<br>• Abdulrahman M. A. Aldawsari | 2024 | MDPI | Evaluating Predictive Accuracy of Regression Models with First-Order Autoregressive Disturbances: A Comparative Approach Using Artificial Neural Networks and Classical Estimators(Rauf et al., 2024) | https://doi.org/10.3390/math12243966 | Nigeria-Arabia Saudita |
| 8 | • Guoping Zeng<br>• sha tao | 2023 | MDPI | A Generalized Linear Transformation and Its Effects on Logistic Regression(Zeng & Tao, 2023) | https://doi.org/10.3390/math11020467 | Estados Unidos-Reino Unido |
| 9 | • Furkat Bolikulov<br>• Rashid Nasimov<br>• Akbar Rashidov<br>• Farkhod Akhmedov<br>• Young-Im Cho | 2024 | MDPI | Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms(Bolikulov et al., 2024) | https://doi.org/10.3390/math12162553 | Corea del Sur-Uzbekistán |
| 10 | • Cong Li<br>• Xupeng Ren<br>• Guohui Zhao | 2023 | MDPI | Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data(C. Li et al., 2023) | https://doi.org/10.3390/a16090422 | China |
| 11 | • Fangfang Li<br>• Hui Sun<br>• Yu Gu<br>• Ge Yu | 2022 | MDPI | A Noise-Aware Multiple Imputation Algorithm for Missing Data(F. Li et al., 2023) | https://doi.org/10.3390/math11010073 | China |
| 12 | • Philip Buczak<br>• Jian-Jia Chen<br>• Markus Pauly | 2023 | MDPI | Analyzing the Effect of Imputation on Classification Performance under MCAR and MAR Missing Mechanisms(Buczak et al., 2023) | https://doi.org/10.3390/e25030521 | Alemania |
| 13 | • Debora Anelli<br>• Pierluigi Morano<br>• Francesco Tajani<br>• Maria Rosaria Guarini | 2025 | MDPI | The Interpretative Effects of Normalization Techniques on Complex Regression Modeling: An Application to Real Estate Values Using Machine Learning(Anelli et al., 2025) | https://doi.org/10.3390/info16060486 | Italia |
| 14 | • Florian Pargent<br>• Florian Pfisterer<br>• Janek Thomas<br>• Bernd Bischl | 2022 | Spriger Nature | Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features(Pargent et al., 2022) | https://doi.org/10.1007/s00180-022-01207-6 | Alemania |
| 15 | • Kelsy Cabello-Solorzano<br>• Isabela Ortigosa de Araujo<br>• Marco Peña<br>• Luís Correia<br>• Antonio J. Tallón | 2023 | Spriger Nature | The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis (Cabello-Solorzano et al., 2023) | https://doi.org/10.1007/978-3-031-42536-3_33 | España-Portugal |
| 16 | • Kiran Maharana<br>• Surajit Mondal<br>• Bhushankumar Nemade | 2022 | sciencedirect | A review: Data preprocessing and data augmentation techniques(Maharana et al., 2022) | https://doi.org/10.1016/j.gltp.2022.04.020 | India |

| N° | AUTOR(ES) | AÑO | BASE DE DATOS | TITULO | ENLACE / DOI | PAÍS |
|---|---|---|---|---|---|---|
| 17 | • Paraskevas Koukaras<br>• Christos Tjortjis | 2025 | MDPI | Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices(Koukaras & Tjortjis, 2025) | https://doi.org/10.3390/ai6100257 | Grecia |
| 18 | • Weixian Mu<br>• Riccardo Cardelli<br>• Simone Ferrari | 2026 | MDPI | Data Preprocessing Techniques for Machine Learning Towards Improving Building Energy Performance: A Systematic Review(Mu et al., 2026) | https://doi.org/10.3390/en19061561 | Italia |
| 19 | • João Manoel Herrera Pinheiro<br>• Suzana Vilas Boas de Oliveira<br>• Thiago Henrique Segreto Silva<br>• Pedro Antonio Rabelo Saraiva<br>• Enzo Ferreira de Souza<br>• Marcelo Becker | 2025 | IEEE | The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks(Pinheiro et al., 2025) | https://doi.org/10.1109/ACCESS.2025.3635541 | Brasil |
| 20 | • Esra'a Alshdaifat<br>• Doa'a Alshdaifat<br>• Ayoub Alsarhan<br>• Fairouz Hussein | 2021 | MDPI | The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance(Alshdaifat et al., 2021) | https://doi.org/10.3390/data6020011 | Jordania |
| 21 | • Kanjana G<br>• Sheeja MK | 2022 | sciencedirect | Secure storage and matching of latent fingerprints using phase shifting digital holography(G & Mk, 2022) | https://doi.org/10.1016/j.patrec.2021.10.017 | India |
| 22 | • Paul T. von Hippel<br>• Jonathan W. Bartlett | 2021 | Project Euclid | Maximum Likelihood Multiple Imputation: Faster Imputations and Consistent Standard Errors Without Posterior Draws(Hippel & Bartlett, 2021) | https://doi.org/10.1214/20-STS793 | Estados Unidos |
| 23 | • Dongying Zheng<br>• Xinyu Hao<br>• Muhammad Khan<br>• Lixia Wang | 2022 | frontiers | Comparison of machine learning and logistic regression as predictive models for adverse maternal and neonatal outcomes of preeclampsia: A retrospective study(Zheng et al., 2022) | https://doi.org/10.3389/fcvm.2022.959649 | China |
| 24 | • Giuseppe A. Veltri | 2025 | Taylor & Francis | The Effects of Data Preprocessing Choices on Behavioral RCT Outcomes: A Multiverse Analysis(Veltri, 2025) | https://doi.org/10.1080/00273171.2025.2575399 | Singapur |
| 25 | • Khaled Mahmud Sujon<br>• Rohayanti Binti Hassan<br>• Zeba Tusnia Towshi<br>• Manal A. Othman | 2024 | IEEE | When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI(Mahmud Sujon et al., 2024) | https://doi.org/10.1109/ACCESS.2024.3462434 | Malasia-India-Arabia Saudita |
| 26 | • Maan Habib<br>• Maan Okayli | 2024 | Spriger Nature | Evaluating the Sensitivity of Machine Learning Models to Data Preprocessing Technique in Concrete Compressive Strength Estimation(Habib & Okayli, 2024) | https://doi.org/10.1007/s13369-024-08776-2 | Siria-Arabia Saudita |

| N° | AUTOR(ES) | AÑO | BASE DE DATOS | TITULO | ENLACE / DOI | PAÍS |
|---|---|---|---|---|---|---|
| 27 | • Everleen Nekesa Wanyonyi<br>• Newton Wafula Masinde | 2025 | IJSRCSEIT | The Impact of Data Preprocessing on Machine Learning Model Performance: A Comprehensive Examination(Wanyonyi & Masinde, 2025) | https://doi.org/10.32628/CSEIT25112854 | Kenia |
| 28 | • Rasyidah<br>• Riswan Efendi<br>• Nazri Mohd. Nawi | 2023 | sciencedirect | Cleansing of inconsistent sample in linear regression model based on rough sets theory(Rasyidah et al., 2023) | https://doi.org/10.1016/j.sasc.2022.200046 | Malasia-Pakistan-Indonesia |
| 29 | • Morteza Amini<br>• Mahdi Roozbeh<br>• Nur Anisah Mohamed | 2024 | MDPI | Separation of the Linear and Nonlinear Covariates in the Sparse Semi-Parametric Regression Model in the Presence of Outliers(Amini et al., 2024) | https://doi.org/10.3390/math12020172 | Irán-Malasia |
| 30 | • K Mallikharjuna Rao<br>• Ghanta Saikrishna<br>• Kundrapu Supriya | 2023 | Spriger Nature | Data preprocessing techniques: emergence and selection towards machine learning models - a practical review using HPA dataset(Mallikharjuna Rao et al., 2023) | https://doi.org/10.1007/s11042-023-15087-5 | India |
| 31 | • Shovan Chowdhury<br>• Yuxiao Lin<br>• Boryann Liaw<br>• Leslie Kerby | 2022 | IEEE | Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance(Chowdhury et al., 2022) | https://doi.org/10.1109/IDSTA55301.2022.9923169 | Estados Unidos |
| 32 | • Qihong Gan<br>• Lang Gong<br>• Dasha Hu<br>• Dasha Hu | 2023 | MDPI | A Hybrid Missing Data Imputation Method for Batch Process Monitoring Dataset(Gan et al., 2023) | https://doi.org/10.3390/s23218678 | China |
| 33 | • Hongsen Ou<br>• Yunan Yao<br>• Yi He | 2024 | MDPI | Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network(Ou et al., 2024) | https://doi.org/10.3390/s24041112 | China |
| 34 | • Ha-Je Park<br>• Yun-Su Koo<br>• Hee-Yeong Yang<br>• Young-Shin Han<br>• Choon-Sung Nam | 2024 | MDPI | Study on Data Preprocessing for Machine Learning Based on Semiconductor Manufacturing Processes(Park et al., 2024) | https://doi.org/10.3390/s24175461 | Corea |
| 35 | • Amal Tawakuli<br>• Bastian Havers<br>• Vincenzo Gulisano<br>• Daniel Kaiser | 2025 | sciencedirect | Survey:Time-series data preprocessing: A survey and an empirical analysis(Tawakuli et al., 2025) | https://doi.org/10.1016/j.jer.2024.02.018 | Suecia-Luxemburgo |
| 36 | • Andreea Vescan<br>• Radu Găceanu<br>• Camelia Şerban | 2024 | Spriger Nature | Exploring the impact of data preprocessing techniques on composite classifier algorithms in cross-project defect prediction(Vescan et al., 2024) | https://doi.org/10.1007/s10515-024-00454-9 | Rumania |
| 37 | • Chrysoula D. Kappatou<br>• James Odgers<br>• Salvador García-Muñoz<br>• Ruth Misener | 2023 | ACS Publications | An Optimization Approach Coupling Preprocessing with Model Regression for Enhanced Chemometrics | https://doi.org/10.1021/acs.iecr.2c04583 | Reino Unido-Estados Unidos |

| N° | AUTOR(ES) | AÑO | BASE DE DATOS | TITULO | ENLACE / DOI | PAÍS |
|---|---|---|---|---|---|---|
| 38 | • Jian Zhong<br>• Chaochao Ma<br>• Li'an Hou<br>• Yicong Yin<br>• Fang Zhao | 2023 | Spriger Nature | Utilization of five data mining algorithms combined with simplified preprocessing to establish reference intervals of thyroid-related hormones for non-elderly adults(Zhong et al., 2023) | https://doi.org/10.1186/s12874-023-01898-5 | China |
| 39 | • Juscimara G. Avelino<br>• George D. C. Cavalcanti<br>• Rafael M. O. Cruz | 2024 | Spriger Nature | Resampling strategies for imbalanced regression: a survey and empirical analysis(Avelino et al., 2024) | https://doi.org/10.1007/s10462-024-10724-3 | Brasil-Canada |
| 40 | • Woo Gyun Shin<br>• Jinseok Lee<br>• Young Chul Ju<br>• Hey Mi Hwang<br>• Suk Whan Ko | 2025 | sciencedirect | Data preprocessing and machine learning method based on ameliorated mathematical models for inferring the power generation of photovoltaic system(Shin et al., 2025) | https://doi.org/10.1016/j.enconman.2025.119793 | Corea |

*Nota: Elaboración propia con base en la información analizada.*

Es importante destacar que las estrategias de preprocesamiento mejoran la calidad de los datos, e influyen directamente en la estimación de los parámetros del modelo de regresión lineal. Por ejemplo, la presencia de valores atípicos puede distorsionar los coeficientes del modelo, mientras que la falta de normalización puede provocar que variables con mayor escala dominen el proceso de aprendizaje. Asimismo, una imputación inadecuada de valores faltantes puede introducir sesgos que afectan la capacidad predictiva del modelo.

## Futuras Líneas de Investigación

Los hallazgos de la presente investigación tienen aplicación directa en cualquier contexto donde se utilicen modelos de regresión lineal para la toma de decisiones. En el ámbito académico, los resultados son aplicables a la predicción de rendimiento estudiantil, como se demostró en el ejemplo práctico con el dataset Student Performance. En el ámbito empresarial, las estrategias de preprocesamiento identificadas pueden implementarse en modelos predictivos de ventas, estimación de precios y análisis de demanda tecnológica. En el área de calidad de software, los mismos principios aplican para la predicción de defectos y métricas de rendimiento de sistemas. Como futuras líneas de investigación se propone:

- Extender el análisis comparativo a algoritmos de regresión no lineal y redes neuronales profundas, verificando si el orden de impacto de las técnicas se mantiene
- Desarrollar un benchmark estandarizado de pipelines de preprocesamiento sobre datasets públicos de la carrera TIC-ESPOCH
- Aplicar herramientas de automatización AutoML para la selección óptima de estrategias de preprocesamiento sin intervención manual
- Comparar empíricamente las técnicas alternativas de cada par, normalización Min-Max versus Z-score, e imputación por media versus imputación múltiple, para determinar cuál produce mayor $R^2$ en distintos contextos de datos.

### Ejemplo Práctico Aplicado Dataset Student Performance (Kaggle)

Con el fin de responder empíricamente la pregunta de investigación, se aplicaron las cuatro técnicas de la Tabla 2 al dataset Student Performance (Aman Chauhan, Kaggle, 39.115 descargas). Este dataset contiene información académica y sociofamiliar de 397 estudiantes de secundaria. Se seleccionaron las variables studytime (horas de estudio semanales), absences (número de ausencias), Mjob (ocupación de la madre, variable categórica), age (edad) y failures (materias reprobadas) como variables predictoras, siendo G3 (nota final, escala 0–20) la variable objetivo.

El modelo de regresión lineal aplicado responde a la pregunta: ¿Es posible predecir la nota final de un estudiante a partir de su tiempo de estudio, ausencias, contexto familiar y comportamiento académico previo? La ecuación del modelo es:

$$G3 = \beta_0 + \beta_1(studytime) + \beta_2(absences) + \beta_3(Mjob) + \beta_4(age) + \beta_5(failures) + \varepsilon$$

El ejemplo fue implementado en Python 3 utilizando Anaconda/Jupyter Notebook con las librerías pandas, numpy y scikit-learn. El código completo está disponible en el archivo preprocesamiento_student_performance.ipynb adjunto a este documento. A continuación se presenta el código y los resultados de cada paso.

Paso 0: Carga del dataset e introducción de valores faltantes

Se cargó el dataset y se introdujeron 25 valores faltantes artificiales en studytime (6.3% de los datos, patrón MCAR) para demostrar la imputación múltiple:

```
import pandas as pd, numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
df = pd.read_excel('Maths.csv', engine='openpyxl')
df = df[['studytime','absences','Mjob','age','failures','G3']].copy()
# Introducir NaN artificiales en studytime (patrón MCAR)
np.random.seed(42)
nan_idx = np.random.choice(df.index, size=25, replace=False)
df.loc[nan_idx, 'studytime'] = np.nan
# Label encoding para baseline
mjob_map = {'at_home':0,'health':1,'other':2,'services':3,'teacher':4}
df['Mjob_label'] = df['Mjob'].map(mjob_map)
```

**Tabla 3**
**Muestra del dataset Student Performance**

| ID | studytime | absences | Mjob | age | failures | G3 |
|---|---|---|---|---|---|---|
| 1 | 2 | 6 | at_home | 18 | 0 | 6 |
| 2 | 2 | 4 | at_home | 17 | 0 | 6 |
| 3 | 2 | 10 | at_home | 15 | 3 | 10 |
| 4 | 3 | 2 | health | 15 | 0 | 15 |
| 5 | 2 | 4 | other | 16 | 0 | 10 |
| 6 | 2 | 10 | services | 16 | 0 | 15 |
| 7 | 2 | 0 | other | 16 | 0 | 11 |
| 8 | 2 | 6 | other | 17 | 0 | 6 |
| 9 | 2 | 0 | services | 15 | 0 | 19 |
| 10 | NaN* | 0 | other | 15 | 0 | 15 |

*Nota: Elaboración propia con base en la información analizada.*

** NaN = valor faltante introducido artificialmente. Fuente: Dataset Student Performance, Kaggle (Aman Chauhan).*

Paso 1: Baseline: Sin preprocesamiento

Se entrenó el modelo eliminando las 25 filas con NaN (enfoque naive). Esto representa el punto de partida sin ninguna técnica de preprocesamiento:

```
# BASELINE: solo filas completas (se pierden las 25 con NaN)
mask = df['studytime'].notna()
X_base = df.loc[mask, ['studytime','absences','Mjob_label','age','failures']].values
y_base = df.loc[mask, 'G3'].values
model = LinearRegression()
model.fit(X_base, y_base)
r2_base = r2_score(y_base, model.predict(X_base)) # → 0.1375
mse_base = mean_squared_error(y_base, model.predict(X_base)) # → 18.2358
# n = 372 estudiantes (perdemos 25 por NaN)
```

Paso 2: Técnica 1: Imputación Múltiple

Se imputaron los NaN de studytime con la media de la columna (2.04 horas). Esto recupera las 25 filas perdidas en el baseline:

```
# IMPUTACIÓN MÚLTIPLE (strategy='mean')
imputer = SimpleImputer(strategy='mean')
df['studytime'] = imputer.fit_transform(df[['studytime']])
# Valor imputado: media = 2.04 horas
X_imp = df[['studytime','absences','Mjob_label','age','failures']].values
y_imp = df['G3'].values
model.fit(X_imp, y_imp)
r2_imp = r2_score(y_imp, model.predict(X_imp)) # → 0.1510
mse_imp = mean_squared_error(y_imp, model.predict(X_imp)) # → 17.9583
# n = 397 estudiantes (recuperamos los 25) Δ R² = +9.82%
```

Paso 3: Técnica 2: Eliminación de Outliers (IQR)

Se detectaron y eliminaron 15 estudiantes con absences > 20 (outliers confirmados por la regla IQR). Cálculo: Q1=0, Q3=8, IQR=8, límite = 8 + 1.5×8 = 20.0:

```
# ELIMINACIÓN DE OUTLIERS, regla IQR en absences
Q1, Q3 = df['absences'].quantile(0.25), df['absences'].quantile(0.75)
IQR = Q3 - Q1 # IQR = 8.0
limite = Q3 + 1.5 * IQR # límite = 20.0
df_out = df[df['absences'] <= limite].copy() # elimina 15 outliers
X_out = df_out[['studytime','absences','Mjob_label','age','failures']].values
y_out = df_out['G3'].values
model.fit(X_out, y_out)
r2_out = r2_score(y_out, model.predict(X_out)) # → 0.1680
mse_out = mean_squared_error(y_out, model.predict(X_out)) # → 17.8379
# n = 382 estudiantes Δ R² = +11.28%
```

Paso 4: Técnica 3: Estandarización Z-score

Se estandarizaron las variables numéricas (studytime, absences, age, failures) con media=0 y desviación=1. En regresión lineal el $R^2$ no cambia (invarianza matemática al escalado lineal), pero los coeficientes $\beta$ se vuelven comparables e interpretables:

```
# ESTANDARIZACIÓN Z-SCORE: x' = (x - μ) / σ
scaler = StandardScaler()
num_cols = ['studytime','absences','age','failures']
df_out[num_cols] = scaler.fit_transform(df_out[num_cols])
X_std = df_out[['studytime','absences','Mjob_label','age','failures']].values
y_std = df_out['G3'].values
model.fit(X_std, y_std)
r2_std = r2_score(y_std, model.predict(X_std)) # → 0.1680 (sin cambio)
mse_std = mean_squared_error(y_std, model.predict(X_std)) # → 17.8379
# Δ R² = 0.00% La regresión lineal es invariante al escalado lineal
# Valor: los coeficientes β ahora son comparables entre variables
```

Paso 5: Técnica 4: Target Encoding para Mjob

Se reemplazó cada categoría de Mjob por la media de G3 de ese grupo, capturando la relación real entre la ocupación de la madre y el rendimiento académico:

```
# TARGET ENCODING: reemplazar categoría por media de G3
target_map = df_out.groupby('Mjob')['G3'].mean()
# at_home=9.10 health=12.15 other=9.79 services=10.94 teacher=11.14
df_out['Mjob_enc'] = df_out['Mjob'].map(target_map)
X_tgt = df_out[['studytime','absences','Mjob_enc','age','failures']].values
y_tgt = df_out['G3'].values
model.fit(X_tgt, y_tgt)
r2_tgt = r2_score(y_tgt, model.predict(X_tgt)) # → 0.1875
mse_tgt = mean_squared_error(y_tgt, model.predict(X_tgt)) # → 17.4207
# Δ R² = +11.58% ← MAYOR IMPACTO INDIVIDUAL
```

**Tabla 4**

**Impacto de las técnicas de preprocesamiento**

| N° | Técnica aplicada | $R^2$ | MSE | Δ $R^2$ | n |
|---|---|---|---|---|---|
| 0 | Baseline (sin preprocesamiento) | 0.1375 | 18.2358 | — | 372 |
| 1 | + Imputación múltiple (studytime) | 0.1510 | 17.9583 | +9.82% | 397 |
| 2 | + Eliminación de outliers (absences) | 0.1680 | 17.8379 | +11.28% | 382 |
| 3 | + Estandarización Z-score | 0.1680 | 17.8379 | 0.00%* | 382 |
| 4 | + Target Encoding (Mjob) ; MAYOR IMPACTO | 0.1875 | 17.4207 | +11.58% | 382 |

*Nota: Elaboración propia con base en la información analizada.*

** El Z-score no modifica el $R^2$ en regresión lineal (invarianza matemática), pero mejora la interpretabilidad de los coeficientes $\beta$. Fuente: Dataset Student Performance, Kaggle.*

## Respuesta a la Pregunta de Investigación

Los resultados del ejemplo práctico (Tabla 4) permiten responder directamente la pregunta de investigación. En el dataset Student Performance, el Target Encoding de la variable categórica Mjob representó el mayor impacto individual con un incremento del $R^2$ de +11.58%, seguido por la eliminación de outliers en absences (+11.28%) y la imputación múltiple de studytime (+9.82%). La estandarización Z-score, aunque no modifica el $R^2$ en regresión lineal, es fundamental para la interpretabilidad de los coeficientes del modelo.

La mejora total acumulada del pipeline completo fue del +36.36% en $R^2$ (de 0.1375 a 0.1875 en $R^2$ acumulado), lo que evidencia que la combinación secuencial de técnicas produce mejoras superiores a cualquier técnica aplicada de forma aislada. Este hallazgo es consistente con la evidencia reportada por Habib y Okayli (2024), quienes encontraron que la combinación de técnicas redujo el MAE en un 31%.

### Aplicación de PRISMA en el contexto de Machine Learning

La metodología PRISMA 2020 demostró ser una herramienta apropiada para la sistematización de la evidencia en el área de Tecnologías de la Información y Machine Learning. Conforme a lo señalado en la investigación formativa previa, PRISMA se complementa en este campo con las guías de Kitchenham para revisiones sistemáticas en ingeniería de software. La extensión PRISMA-AI, actualmente en desarrollo, busca estandarizar aún más el reporte de revisiones sobre intervenciones basadas en inteligencia artificial. El proceso de cuatro fases (identificación, cribado, elegibilidad e inclusión) permitió reducir de 187 registros iniciales a los 40 estudios finales con máxima relevancia y calidad metodológica.

## CONCLUSIONES

La revisión sistemática realizada bajo la metodología PRISMA 2020 y el ejemplo práctico aplicado al dataset Student Performance de Kaggle permitieron identificar, sintetizar y validar empíricamente el impacto de las estrategias de preprocesamiento de datos en la precisión de los algoritmos de regresión lineal.

En respuesta directa a la pregunta de investigación: según el ejemplo práctico aplicado, el Target Encoding de variables categóricas representó la estrategia con mayor impacto individual en el $R^2$ del modelo (+11.58%), seguido por la eliminación de outliers (+11.28%) y la imputación múltiple (+9.82%). La estandarización Z-score, aunque invariante al $R^2$ en regresión lineal, es fundamental para la interpretabilidad de los coeficientes $\beta$ del modelo.

La imputación de valores faltantes mediante métodos de mayor precisión (imputación múltiple) permitió recuperar el 6.3% de los registros perdidos por eliminación naive de filas, mejorando el $R^2$ en un 9.82% respecto al baseline. Los métodos híbridos RF+GAN reportados en la literatura (Ou et al., 2024) alcanzan mejoras de hasta el 23% en contextos industriales más complejos.

La eliminación de outliers en la variable absences mediante la regla IQR (15 registros eliminados con más de 20 ausencias) produjo una mejora del $R^2$ de +11.28%. Este resultado es consistente con Habib y Okayli (2024), quienes reportaron reducciones del MAE de hasta un 31% combinando esta técnica con normalización en un contexto distinto.

La codificación adecuada de variables categóricas a través del Target Encoding capturó la relación real entre la ocupación de la madre y el rendimiento académico, superando al label encoding simple con un incremento del $R^2$ de +11.58%. Este hallazgo confirma lo documentado por Pargent et al. (2022), quienes demostraron que el target encoding regularizado supera a los métodos tradicionales en aprendizaje supervisado.

Estos hallazgos son aplicables al desarrollo de modelos predictivos en el área de Tecnologías de la Información, incluyendo predicción de rendimiento académico, estimación de demanda tecnológica y modelos de calidad de software. A pesar de los aportes, existen limitaciones: el análisis se centra en regresión lineal; los resultados dependen del dataset utilizado; y la revisión se limita a publicaciones 2021–2026.

## FINANCIACIÓN

La presente investigación fue desarrollada en el marco de las actividades académicas de la asignatura Machine Learning de la Carrera de Tecnologías de la Información y Comunicación de la Escuela Superior Politécnica de Chimborazo (ESPOCH). No se contó con financiamiento externo; todos los recursos utilizados (acceso a bases de datos, software y tiempo de investigación) fueron provistos por los propios autores.

## CONFLICTO DE INTERESES

Los autores declaran que no existe ningún conflicto de intereses, directo ni indirecto, con respecto a la presente investigación, sus métodos, fuentes de financiamiento, resultados o conclusiones. Ninguno de los autores tiene vínculos comerciales, financieros o personales con organizaciones que pudieran influir en el desarrollo o interpretación de los hallazgos presentados en este artículo.

## CONTRIBUCIÓN DE AUTORÍA

### Tabla de contribución

### Contribución de autoría según la taxonomía CRediT

| Participar activamente en: | Camacho | Álvarez | Daqui | Mazacon | Olivo | Vele |
|---|---|---|---|---|---|---|
| Conceptualización | X | X | X | – | – | – |
| Análisis formal | X | – | X | X | X | – |
| Adquisición de fondos | – | – | – | – | – | – |
| Investigación | – | X | X | X | X | X |
| Metodología | X | X | – | – | X | X |
| Administración del proyecto | X | X | – | – | – | X |
| Recursos | – | X | – | X | X | X |
| Redacción – borrador original | X | X | X | X | X | X |
| Redacción – revisión y edición | X | X | X | X | X | X |
| Discusión de los resultados | X | X | X | X | X | X |
| Revisión y aprobación de la versión final | X | X | X | X | X | X |

*Nota: Distribución declarada conforme a la taxonomía CRediT.*

## REFERENCIAS

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3), 52. https://doi.org/10.3390/technologies9030052

Alshdaifat, E., Alshdaifat, D., Alsarhan, A., Hussein, F., & El-Salhi, S. M. F. S. (2021). The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. Data, 6(2), 11. https://doi.org/10.3390/data6020011

Amini, M., Roozbeh, M., & Mohamed, N. A. (2024). Separation of the Linear and Nonlinear Covariates in the Sparse Semi-Parametric Regression Model in the Presence of Outliers. Mathematics, 12(2), 172. https://doi.org/10.3390/math12020172

Anelli, D., Morano, P., Tajani, F., & Guarini, M. R. (2025). The Interpretative Effects of Normalization Techniques on Complex Regression Modeling. Information, 16(6), 486. https://doi.org/10.3390/info16060486

Avelino, J. G., Cavalcanti, G. D. C., & Cruz, R. M. O. (2024). Resampling strategies for imbalanced regression: A survey and empirical analysis. Artificial Intelligence Review, 57(4), 82. https://doi.org/10.1007/s10462-024-10724-3

Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., & Cho, Y.-I. (2024). Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. Mathematics, 12(16), 2553. https://doi.org/10.3390/math12162553

Buczak, P., Chen, J.-J., & Pauly, M. (2023). Analyzing the Effect of Imputation on Classification Performance under MCAR and MAR Missing Mechanisms. Entropy, 25(3), 521. https://doi.org/10.3390/e25030521

Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., & J. Tallón-Ballesteros, A. (2023). The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis. En 18th International Conference on Soft Computing Models (SOCO 2023) (pp. 344-353). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42536-3_33

Chen, J., & Yang, Z. (2024). Revolutionizing Time Series Data Preprocessing with a Novel Cycling Layer in Self-Attention Mechanisms. Applied Sciences, 14(19), 8922. https://doi.org/10.3390/app14198922

Chowdhury, S., Lin, Y., Liaw, B., & Kerby, L. (2022). Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance. 2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), 17-25. https://doi.org/10.1109/IDSTA55301.2022.9923169

de Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. Applied Soft Computing, 133, 109924. https://doi.org/10.1016/j.asoc.2022.109924

Demir, S., & Sahin, E. K. (2024). The effectiveness of data pre-processing methods on the performance of machine learning techniques using RF, SVR, Cubist and SGB. Stochastic Environmental Research and Risk Assessment, 38(8), 3273-3290. https://doi.org/10.1007/s00477-024-02745-9

Gan, Q., Gong, L., Hu, D., Jiang, Y., & Ding, X. (2023). A Hybrid Missing Data Imputation Method for Batch Process Monitoring Dataset. Sensors, 23(21), 8678. https://doi.org/10.3390/s23218678

Habib, M., & Okayli, M. (2024). Evaluating the Sensitivity of Machine Learning Models to Data Preprocessing Technique in Concrete Compressive Strength Estimation. Arabian Journal for Science and Engineering, 49(10), 13709-13727. https://doi.org/10.1007/s13369-024-08776-2

Hassanat, A. B., Alqaralleh, M. K., Tarawneh, A. S., Almohammadi, K., Alamri, M., Alzahrani, A., Altarawneh, G. A., & Alhalaseh, R. (2024). A Novel Outlier-Robust Accuracy Measure for Machine Learning Regression Using a Non-Convex Distance Metric. Mathematics, 12(22), 3623. https://doi.org/10.3390/math12223623

Hippel, P. T. von, & Bartlett, J. W. (2021). Maximum Likelihood Multiple Imputation: Faster Imputations and Consistent Standard Errors Without Posterior Draws. Statistical Science, 36(3), 400-420. https://doi.org/10.1214/20-STS793

Kappatou, C. D., Odgers, J., García-Muñoz, S., & Misener, R. (2023). An Optimization Approach Coupling Preprocessing with Model Regression for Enhanced Chemometrics. Industrial & Engineering Chemistry Research. https://doi.org/10.1021/acs.iecr.2c04583

Koukaras, P., & Tjortjis, C. (2025). Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. AI, 6(10), 257. https://doi.org/10.3390/ai6100257

Li, C., Ren, X., & Zhao, G. (2023). Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data. Algorithms, 16(9), 422. https://doi.org/10.3390/a16090422

Li, F., Sun, H., Gu, Y., & Yu, G. (2023). A Noise-Aware Multiple Imputation Algorithm for Missing Data. Mathematics, 11(1), 73. https://doi.org/10.3390/math11010073

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91-99. https://doi.org/10.1016/j.gltp.2022.04.020

Mahmud Sujon, K., Binti Hassan, R., Tusnia Towshi, Z., Othman, M. A., Abdus Samad, M., & Choi, K. (2024). When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI. IEEE Access, 12, 135300-135314. https://doi.org/10.1109/ACCESS.2024.3462434

Mallikharjuna Rao, K., Saikrishna, G., & Supriya, K. (2023). Data preprocessing techniques: Emergence and selection towards machine learning models – a practical review using HPA dataset. Multimedia Tools and Applications, 82(24), 37177-37196. https://doi.org/10.1007/s11042-023-15087-5

Mu, W., Cardelli, R., & Ferrari, S. (2026). Data Preprocessing Techniques for Machine Learning Towards Improving Building Energy Performance: A Systematic Review. Energies, 19(6), 1561. https://doi.org/10.3390/en19061561

Ou, H., Yao, Y., & He, Y. (2024). Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network. Sensors, 24(4), 1112. https://doi.org/10.3390/s24041112

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021a). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … McKenzie, J. E. (2021b). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ, 372, n160. https://doi.org/10.1136/bmj.n160

Pandit, P., Dey, P., & Krishnamurthy, K. N. (2021). Comparative Assessment of Multiple Linear Regression and Fuzzy Linear Regression Models. SN Computer Science, 2(2), 76. https://doi.org/10.1007/s42979-021-00473-3

Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics, 37(5), 2671-2692. https://doi.org/10.1007/s00180-022-01207-6

Park, H.-J., Koo, Y.-S., Yang, H.-Y., Han, Y.-S., & Nam, C.-S. (2024). Study on Data Preprocessing for Machine Learning Based on Semiconductor Manufacturing Processes. Sensors, 24(17), 5461. https://doi.org/10.3390/s24175461

Pinheiro, J. M. H., Oliveira, S. V. B. de, Silva, T. H. S., Saraiva, P. A. R., Souza, E. F. de, Godoy, R. V., Ambrosio, L. A., & Becker, M. (2025). The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks. IEEE Access, 13, 199903-199931. https://doi.org/10.1109/ACCESS.2025.3635541

Rasyidah, Efendi, R., Nawi, N. Mohd., Deris, M. M., & Burney, S. M. A. (2023). Cleansing of inconsistent sample in linear regression model based on rough sets theory. Systems and Soft Computing, 5, 200046. https://doi.org/10.1016/j.sasc.2022.200046

Rauf, R. I., Alrasheedi, M. A., Sadiq, R., & Aldawsari, A. M. A. (2024). Evaluating Predictive Accuracy of Regression Models with First-Order Autoregressive Disturbances: A Comparative Approach Using Artificial Neural Networks and Classical Estimators. Mathematics, 12(24), 3966. https://doi.org/10.3390/math12243966

Shin, W. G., Lee, J. S., Ju, Y. C., Hwang, H. M., & Ko, S. W. (2025). Data preprocessing and machine learning method based on ameliorated mathematical models for inferring the power generation of photovoltaic system. Energy Conversion and Management, 333, 119793. https://doi.org/10.1016/j.enconman.2025.119793

Tawakuli, A., Havers, B., Gulisano, V., Kaiser, D., & Engel, T. (2025). Survey: Time-series data preprocessing: A survey and an empirical analysis. Journal of Engineering Research, 13(2), 674-711. https://doi.org/10.1016/j.jer.2024.02.018

Veltri, G. A. (2025). The Effects of Data Preprocessing Choices on Behavioral RCT Outcomes: A Multiverse Analysis. Multivariate Behavioral Research, 0(0), 1-16. https://doi.org/10.1080/00273171.2025.2575399

Vescan, A., Găceanu, R., & Şerban, C. (2024). Exploring the impact of data preprocessing techniques on composite classifier algorithms in cross-project defect prediction. Automated Software Engineering, 31(2), 47. https://doi.org/10.1007/s10515-024-00454-9

Wanyonyi, E. N., & Masinde, N. W. (2025). The Impact of Data Preprocessing on Machine Learning Model Performance: A Comprehensive Examination. IJSRCSEIT, 11(2), 3814-3827. https://doi.org/10.32628/CSEIT25112854

Zeng, G., & Tao, S. (2023). A Generalized Linear Transformation and Its Effects on Logistic Regression. Mathematics, 11(2), 467. https://doi.org/10.3390/math11020467

Zheng, D., Hao, X., Khan, M., Wang, L., Li, F., Xiang, N., Kang, F., Hamalainen, T., Cong, F., Song, K., & Qiao, C. (2022). Comparison of machine learning and logistic regression as predictive models for adverse maternal and neonatal outcomes of preeclampsia: A retrospective study. Frontiers in Cardiovascular Medicine, 9. https://doi.org/10.3389/fcvm.2022.959649

Zhong, J., Ma, C., Hou, L., Yin, Y., Zhao, F., Hu, Y., Song, A., Wang, D., Li, L., Cheng, X., & Qiu, L. (2023). Utilization of five data mining algorithms combined with simplified preprocessing to establish reference intervals of thyroid-related hormones for non-elderly adults. BMC Medical Research Methodology, 23(1), 108. https://doi.org/10.1186/s12874-023-01898-5