

MODELOS DE APRENDIZAJE AUTOMÁTICO: APLICACIÓN Y EFICIENCIA

Josselyn Nicolle Mendoza Álava
Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López
josselyn.mendoza@espam.edu.ec
Calceta, Ecuador

Luis Antonio Macías Bermeo
Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López
luismaciasantonio2000@gmail.com
Calceta, Ecuador

Jessica Morales-Carrillo
Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López
jmorales@espam.edu.ec
Calceta, Ecuador

Luis Cedeño-Valarezo
Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López
lcedeno@espam.edu.ec
Calceta, Ecuador

DOI: <https://doi.org/10.56124/encriptar.v7i14.005>

RESUMEN

Existe una gran cantidad de sistemas que se estudian y se desarrollan en el campo del aprendizaje automático y la inteligencia artificial, lo que sin duda ha logrado transformar una multitud de industrias y aplicaciones en todo el mundo. El propósito de este estudio radica en una revisión exhaustiva de 120 artículos, donde se destaca la diversidad de tipos de datos utilizados en el aprendizaje automático, desde datos estructurados como tablas y series temporales hasta datos no estructurados como imágenes y texto, identificando su aplicación óptima en los diversos contextos y aplicaciones. Para el desarrollo de esta investigación se llevó a cabo una revisión sistemática de la literatura, originada por la búsqueda de los modelos computacionales utilizados en distintos ámbitos del conocimiento humano; se procedió con ejecución de la búsqueda donde se realiza la extracción de datos, describiendo cada uno de los campos considerados para el análisis de la información; se procesó y analizaron los datos considerando los modelos con mayor frecuencia de uso y sus métricas de rendimiento. Los resultados de esta investigación destacan la preferencia del modelo de Máquinas de Soporte Vectorial como el más frecuentemente empleado en variedad de aplicaciones, así mismo la investigación también revela que, en términos de eficiencia y precisión los modelos Gradient Boosting y Redes Neuronales Artificiales, sobresalen significativamente. Estos resultados evidencian la importancia de

una colaboración interdisciplinaria y la necesidad de aplicación adecuada para garantizar que el aprendizaje automático y la inteligencia artificial sigan siendo motores para el avance tecnológico.

Palabras clave: aprendizaje automático, inteligencia artificial, tipos de modelos.

MACHINE LEARNING MODELS: APPLICATION AND EFFICIENCY

ABSTRACT

There are many systems that are studied and developed in the field of machine learning and artificial intelligence which are constantly evolving fields that have transformed a multitude of industries and applications around the world. The purpose of this study lies in a comprehensive review of 120 articles, highlighting the diversity of data types used in machine learning, from structured data such as tables and time series to unstructured data such as images and text. To develop this research, a systematic review of the literature was carried out, originating from the search for computational models used in different areas of human knowledge; The search proceeded with the execution of the data extraction, describing each of the fields considered for the analysis of the information; The data was processed and analyzed considering the models with the highest frequency of use and their performance metrics. The results of this research highlight the preference of the Support Vector Machines model as the most frequently used in a variety of applications. Likewise, the research also reveals that, in terms of efficiency and precision, the Gradient Boosting and Artificial Neural Networks models stand out significantly. These results evidence the importance of interdisciplinary collaboration and the need for proper application to ensure that machine learning and artificial intelligence continue to be drivers for technological advancement.

Keywords: machine learning, artificial intelligence, types of models.

1. Introducción

La inteligencia artificial (IA) ha revolucionado nuestra forma de interactuar con la tecnología y ha impactado en diversos sectores de la sociedad (García Peñalvo et al., 2023). Esta disciplina, concebida como parte de las Ciencias de la Computación, proporciona una diversidad de métodos,

técnicas y herramientas para modelizar y resolver problemas, simulando el proceder de los sujetos cognoscentes. Estos métodos y técnicas permiten desarrollar sistemas capaces de realizar tareas que requieren inteligencia humana, como el reconocimiento de voz, la toma de decisiones o el procesamiento de datos complejos (Ocaña Fernández et al., 2019). Con el rápido avance de la tecnología ha convertido a la IA en un campo de investigación y desarrollo de gran importancia, impulsando la innovación en diferentes áreas.

La IA abarca una amplia gama de áreas de investigación y aplicaciones que permiten a las máquinas simular la inteligencia humana (Solórzano Álava et al., 2022). Entre estas áreas se encuentran el procesamiento del lenguaje natural, los sistemas expertos, la robótica, la computación paralela, el aprendizaje automático y el aprendizaje profundo. Cada uno de estos artistas ofrece un conjunto único de herramientas y técnicas que han impulsado la innovación y ampliado los límites de la tecnología actual (LOZANO ORTEGA et al., 2003). Desde la capacidad de comprender y procesar grandes cantidades de datos hasta la capacidad de aprender y adaptarse, la IA ha transformado la forma en que se interactúa con la tecnología.

El Aprendizaje Automático (Machine Learning) es un área fundamental de la Inteligencia Artificial que ha transformado nuestra capacidad para procesar y comprender datos de manera automatizada. Mediante el uso de algoritmos y modelos matemáticos, el Aprendizaje Automático permite a las máquinas aprender de los datos sin ser programadas explícitamente. (Tandon et al., 2019). Para hacer un modelo de aprendizaje automático, se necesita los datos sobre los que se puede hacer el modelo.

Existen diferentes modelos de aprendizajes, entre los más destacados se encuentran: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y el aprendizaje de refuerzo (Tamayo & Pérez Marín, 2017). La supervisión del aprendizaje se basa en la disponibilidad de datos etiquetados, donde el modelo se entrena para predecir o clasificar nuevas muestras. Por el contrario, el aprendizaje no supervisado busca encontrar patrones y estructuras en datos no etiquetados, lo que permite agrupar y detectar anomalías (Sandoval, 2018, #). En este tipo de aprendizaje hay dos algoritmos (entrenamientos): el de clasificación y el de regresión. El aprendizaje semisupervisado combina elementos de ambos enfoques, empleando datos etiquetados y no etiquetados para entrenar el modelo. Finalmente, el aprendizaje por refuerzo se basa en un proceso de ensayo y error en el que el modelo interactúa con un entorno y recibe recompensas o caso contrario una penalización. Esto en esencia es la aplicación de la retroalimentación en el proceso de aprendizaje (Pineda Pertuz, 2022).

Los modelos son representaciones simplificadas de sistemas o fenómenos complejos que permiten comprender, predecir y tomar decisiones (García Falckenheiner et al., 2023). En el contexto del aprendizaje automático, los modelos son herramientas fundamentales para procesar datos y extraer patrones que ayuda a tomar decisiones informadas. Al explorar las diversas técnicas en este campo, se destacan en particular los árboles de decisión, un algoritmo ampliamente utilizado para tareas de clasificación y regresión. Además, se han desarrollado extensiones notables de este enfoque, como los Bosques Aleatorios (Random Forest), que buscan aumentar la precisión de los modelos de clasificación. Por otro lado, Naive Bayes es otro algoritmo probabilístico bastante interesante y usado con frecuencia en tareas de aprendizaje automático, el support vector machine (SVM) es un conjunto de algoritmos de aprendizaje supervisado, estos métodos están propiamente

relacionados con problemas de clasificación y regresión entre otras (Herrera Zurita, 2016).

Las métricas que se pueden usar para validar un modelo de regresión cumplen un papel fundamental en la evaluación del desempeño y la precisión del modelo en relación con los datos de prueba o validación (Orozco Iguasnia et al., 2021) . Estas métricas proporcionan una medida cuantitativa de qué tan bien se ajusta el modelo a los datos y permiten comparar diferentes modelos o ajustes. Cada una de estas métricas posee una función o método correspondiente al cual proporcionar como parámetros los valores reales de la variable objetivo (y) y los valores predichos por el modelo. Algunas de las métricas utilizadas para validar modelos son las siguientes como el error absoluto medio (MAE) y el Error Cuadrático Medio (MSE) para medir la precisión de las predicciones y en problemas principalmente de clasificación, se utilizan métricas como Recall, F1-Score, Precisión, Exactitud (Accuracy), G-Mean y Área bajo la Curva (AUC) para evaluar el desempeño en diferentes aspectos. Estas métricas ofrecen una visión completa del rendimiento del modelo, ayudando a tomar decisiones informadas sobre ajustes y mejoras. (Beunza Nuin et al., 2023).

El propósito fundamental de esta investigación radica en llevar a cabo un análisis bibliográfico profundo y abarcador de los modelos de aprendizaje automático que son ampliamente adoptados y utilizados con frecuencia. Este análisis tiene como finalidad establecer una referencia sólida y esclarecedora para futuras investigaciones que aborden el procesamiento y análisis de conjuntos de datos.

2. Metodología

Para el desarrollo de esta investigación se llevó a cabo una revisión sistemática de la literatura, originada por la búsqueda de los modelos computacionales utilizados en distintos ámbitos del conocimiento humano. La revisión de la literatura es una fase imprescindible en cualquier trabajo de investigación, puesto que ayuda a situar la investigación y a sustentar teórica y conceptualmente a partir de lo que otros investigadores e investigadoras han escrito previamente sobre la temática (Arnau Sabatés & Sala Roca, 2020).

Se aplicó las tres etapas de la Revisión Sistemática (Carrizo & Moller, 2018), mismas que se detallan a continuación:

Definición para la búsqueda

Se inició con la investigación de los modelos de aprendizaje automático más utilizados y su eficiencia para establecer una línea base en investigaciones relacionadas, las comprensión de las técnicas y enfoques que han demostrado ser efectivos en una amplia gama de problemas, es recomendable adaptar el análisis a la tarea o problema específico en cuestión y considerar factores como la disponibilidad de datos, se pueden utilizar técnicas como el análisis bibliométrico, que implican analizar la frecuencia de aparición de ciertos modelos en publicaciones científicas y la cantidad de citas que reciben. Además, es posible realizar estudios comparativos en los que se evalúan diferentes modelos utilizando conjuntos de datos.

Es importante destacar que cada plataforma de búsqueda académica o base de datos puede tener su propia sintaxis y características para realizar búsquedas más avanzadas, como la combinación de palabras claves cuando se recopilan y preparan los datos relevantes para las búsquedas de artículos

en el campos, es común utilizar palabras clave específicas para dirigir la búsqueda hacia los temas de interés, se emplearon las siguientes palabras claves: Aprendizaje automático, Inteligencia artificial y Tipos de modelos.

Las mismas que se usaron en la base de datos, es el que delimita las fuentes de información o bases de datos más pertinentes en función del ámbito disciplinario es aquí donde se realizó la revisión del tema de estudio a través de buscadores genéricos como: Google Scholar, Scielo, Redalyc, Dialnet, IEEE, Google Académica, estas bases de datos son herramientas valiosas para los investigadores y académicos, ya que permiten acceder a una amplia gama de contenido académico y científico para llevar a cabo investigaciones, realizar revisiones bibliográficas y estar al tanto de los avances en sus respectivas áreas de estudio. Cada una de estas bases de datos tiene sus propias características y alcance, por lo que es recomendable explorarlas y utilizarlas según sus necesidades específicas.

Un desafío clave en la construcción de sistemas de aprendizaje automático es la necesidad de incluir datos diversos y representativos, entre más acceso a información de alta calidad que se adapte a las intenciones del modelo, mayor será la exactitud, de acuerdo con lo expuesto, se logró identificar más de 400 artículos pero se excluyeron algunos artículos, ya que habían tema relacionado con la inteligencia artificial pero no tenía relación con las métricas, los tipos de modelos y otros datos relevante para la investigación; solo 120 artículos especificaba en la información requerida de lo que se buscaba .

Ejecución de la búsqueda

A continuación, se muestran los campos que se tomaron en cuenta para la extracción de información de los 120 artículos. En la tabla 1 se describen cada uno de los campos considerados para el análisis de la información. Los atributos más relevantes fueron: Año, tema, autores, categoría, sigas modelos, nombre de modelo, modelo base, tipo de dataset, precisión, accuracy, f1, recall, error MAE, error MSE, G-mean, AUC, url y observaciones de la investigación.

Para llevar a cabo un análisis exhaustivo de los trabajos primarios seleccionados, se han establecido criterios de inclusión precisos. En cuanto al idioma, se han considerado tanto trabajos en inglés como en español, con el propósito de abarcar una amplia diversidad de fuentes y perspectivas. En lo que respecta a la diversidad de foco, se ha priorizado la inclusión de trabajos que traten temáticas relacionadas con el aprendizaje automático (machine learning). Esto garantiza que los trabajos seleccionados se alinean con el objetivo de la investigación y permitan profundizar en los desarrollos más recientes y relevantes en este campo.

En términos del tipo de publicación, se han tomado en consideración fuentes respetadas y confiables como Scielo, Redalyc, Dialnet, IEEE y Google Académico. Esta elección de fuentes de renombre asegura la calidad y la rigurosidad de los trabajos incluidos en el análisis. Además, se ha limitado el rango de Año de Publicación a partir de 2018 en adelante. Esto permite que se examinen investigaciones recientes, lo que es crucial para capturar las tendencias actuales y las novedades en el ámbito del aprendizaje automático. Con estos criterios específicos, se busca establecer una base sólida para un análisis detallado y una visión integral de los avances en el aprendizaje automático en los últimos años.

Tabla 1 – Campos que se consideran en la recopilación de información

Campos	Descripción
Año	Se refiere al año en que se registró o se realizó una determinada observación o medición, es decir en qué año se publicó el artículo.
Tema	Se refiere al área de interés, título o materia de la que trata la información registrada en el conjunto de datos.
Autor(es)	Se refiere a la persona o personas que han creado o recopilado la información contenida en el conjunto de datos.
Categoría	Se refiere al tipo de modelos que se utilizó regresión o clasificador.
Siglas modelos	Son las abreviaciones utilizadas para identificar diferentes modelos
Nombre de modelo	Se refiere a los nombres de los diferentes modelos o sistemas que se han utilizado para generar o analizar la información contenida en el conjunto de datos.
Modelo Base	Se refiere al modelo inicial utilizado como punto de partida en el desarrollo de modelos más complejos y precisos en el campo del aprendizaje automático.
Tipo de Dataset	Se refiere al conjunto de datos en función de sus características y propósitos principales, del cual estaban los datos numéricos, gráficos e imágenes.
Precision	<p>Es la medida de evaluación de la calidad de un modelo de clasificación en machine learning, y se utiliza para evaluar qué tan bien el modelo identifica correctamente los verdaderos positivos entre todos los positivos predichos. La fórmula para calcular la precisión es:</p> $Precision = \frac{TP}{TP + FP}$ <p>Donde: Verdaderos positivos (True Positives, <i>TP</i>) son los casos en que el modelo predijo correctamente una clase positiva. Falsos positivos (False Positives, <i>FP</i>) son los casos en que el modelo predijo incorrectamente una clase positiva cuando en realidad era negativa.</p>
Accuracy	<p>Esta medida se utiliza para evaluar qué tan bien el modelo clasifica correctamente todas las clases, tanto las positivas como las negativas. La fórmula para calcular el accuracy es:</p> $Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$ <p>Donde: Verdaderos positivos (True Positives, <i>TP</i>) son los casos en que el modelo predijo correctamente una clase positiva. Falsos positivos (False Positives, <i>FP</i>) son los casos en que el modelo predijo incorrectamente una clase positiva cuando en realidad era negativa. Verdaderos negativos (True Negatives, <i>TN</i>) son los casos en que el modelo predijo correctamente una clase negativa. Falsos negativos (False Negatives, <i>FN</i>) son los casos en que el modelo predijo incorrectamente una clase negativa cuando en realidad era positiva</p>
F1	Es la métrica de evaluación de modelos de clasificación que combina la precisión y el recall. La precisión mide la proporción de predicciones positivas que son correctas, mientras que el recall mide la proporción de casos positivos que son identificados correctamente por el modelo.

	<p>La fórmula de la F1-score es:</p> $F1 = \frac{2 * (precisión * recall)}{(precision + recall)}$ <p>Dónde la precisión es la proporción de verdaderos positivos sobre la suma de verdaderos positivos y falsos positivos, y recall es la proporción de verdaderos positivos sobre la suma de verdaderos positivos y falsos negativos.</p>
Recall	<p>Es la métrica de evaluación de modelos de clasificación que mide la proporción de casos positivos que son identificados correctamente por el modelo. En otras palabras, el <i>recall</i> mide la capacidad del modelo para encontrar todos los casos positivos.</p> <p>La fórmula de la recall es:</p> $Recall = \frac{TP}{TP + FN}$ <p>Donde los verdaderos positivos son los casos positivos que el modelo identifica correctamente y los falsos negativos son los casos positivos que el modelo clasifica incorrectamente como negativos.</p>
Error MAE	<p>Es la métrica que se utiliza para evaluar la precisión de un modelo. Esta métrica mide la diferencia absoluta promedio entre las predicciones del modelo y los valores reales en el conjunto de datos.</p> <p>La fórmula del <i>MAE</i> es:</p> $MAE = (1/n) * \sum y_i - \hat{y}_i $ <p>Donde n es el número de muestras en el conjunto de datos, y_i son los valores reales de las muestras y \hat{y}_i son las predicciones del modelo para las muestras correspondientes. La barra vertical $$ indica el valor absoluto.</p>
Error MSE	<p>Es utilizada para evaluar la precisión de un modelo. Esta métrica mide el promedio de las diferencias cuadráticas entre las predicciones del modelo y los valores reales en el conjunto de datos elevado al cuadrado.</p> <p>La fórmula del <i>MSE</i> es:</p> $MSE = (1/n) * \sum (y_i - \hat{y}_i)^2$ <p>Donde n es el número de muestras en el conjunto de datos, y_i son los valores reales de las muestras y \hat{y}_i son las predicciones del modelo para las muestras correspondientes.</p> <p>Para obtener una medida más fácilmente interpretable, se puede calcular la Raíz del Error Cuadrático Medio (RMSE), que es la raíz cuadrada del MSE:</p> $RMSE = \sqrt{MSE}$ <p>El <i>RMSE</i> se expresa en las mismas unidades que las variables de la variable objetivo y, por lo tanto, es más fácil de interpretar en términos de la magnitud del error en las predicciones del modelo.</p>

G-mean	<p>También conocido como Geometric Mean Score, es una métrica utilizada para evaluar la eficacia de modelos de clasificación en conjuntos de datos desequilibrados. El G-mean es la raíz cuadrada del producto de la sensibilidad (recall) y la especificidad del modelo.</p> <p>La fórmula del G-mean es:</p> $G - mean = \sqrt{(sensibilidad * especificidad)}$ <p>Donde la sensibilidad mide la proporción de casos positivos que son identificados correctamente por el modelo, y la especificidad mide la proporción de casos negativos que son identificados correctamente por el modelo.</p> <p>La sensibilidad se calcula como:</p> $Sensibilidad = \frac{TP}{TP + FN}$ <p>Mientras que la especificidad se calcula como:</p> $Especificidad = \frac{TN}{TN + FN}$
AUC	<p>(Area Under the Curve) es una métrica utilizada para evaluar la eficacia de modelos de clasificación en conjuntos de datos desequilibrados. El AUC mide la capacidad del modelo para distinguir entre clases positivas y negativas en el conjunto de datos, independientemente del umbral de decisión utilizado para hacer las predicciones.</p> <p>La fórmula del AUC es:</p> $AUC = integral (sensibilidad (tasa positiva verdadera) d(especificidad (tasa negativa verdadera)))$
URL	<p>Se refiere a la dirección web que es una cadena de caracteres que identifica una ubicación en internet de un recurso específico, como una página web.</p>
Observación	<p>Son los datos que se agregan o se recolectan en un estudio y son fundamentales para la toma de decisiones basada en los datos de análisis.</p>

En esta investigación, se llevó a cabo la identificación y descripción de los modelos definidos en los artículos seleccionados, así como la determinación de un "modelo base" para el análisis comparativo.

Para identificar el modelo base en esta investigación, se siguió un proceso riguroso de análisis. Se comenzó identificando todos los modelos definidos en los artículos seleccionados, lo que resultó en un conjunto inicial de 65 modelos diferentes. Sin embargo, para facilitar la comparación y el análisis, estos modelos se categorizaron en función de sus características y propiedades comunes.

Este proceso de categorización permitió reducir el conjunto inicial de 65 modelos a un grupo más manejable de 13 modelos base. Estos modelos base incluían una variedad de enfoques, como Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree (DT), Artificial Neural Network (ANN), Naive Bayes (NB), Logistic Regression (LR), Multilayer Perceptron (MLP), Regularized Linear (RL), Convolutional Neural Network (CNN), Gradient Boosting (GB), y Recurrent Neural Network (RNN).

La selección de estos 13 modelos base se fundamentó en la representatividad de los diferentes enfoques y en la capacidad de estos para abordar una amplia variedad de problemas en el campo del aprendizaje automático. Estos modelos base se convirtieron en los puntos de referencia para comparar y evaluar el rendimiento de otros modelos y enfoques específicos en el análisis comparativo.

Discusión de los resultados

Después de realizar la revisión sistemática y recopilar los datos relevantes, se procedió a realizar un análisis de los datos cuantitativos obtenidos. Este análisis se enfoca en los modelos que tienen mayor frecuencia de uso y también se evalúan las métricas de rendimiento, se identificaron varios modelos de aprendizaje automático utilizados en diferentes contextos, los modelos identificados incluyen, entre otros, redes neuronales, árboles de decisión, máquinas de vectores de soporte (SVM) y algoritmos de agrupamiento como k-means etc.

En la tabla 2 se muestra la cantidad de modelo base identificados y sus frecuencias encontradas, los modelos que se identificaron durante la investigación, así como la frecuencia con la que aparecen en los datos

recopilados, también se realizó un análisis y en la tabla 3 se analiza los tipos de datos, en este caso se tomaron datos de porcentajes, ya que había artículos que utilizaron múltiples tipos de datos en una sola investigación, como se indica en la tabla 4 del documento se presenta un análisis detallado sobre las diversas aplicaciones a las que se destinan los modelos que se han estudiado, este análisis puede proporcionar una visión general de las áreas o contextos en los que estos modelos son más comunes o relevantes, en la tabla 5, se presenta información relacionada con los tipos de datos utilizados en relación con clasificadores y regresión en un contexto de aprendizaje automático, se detalla cómo se han empleado diferentes tipos de datos en la construcción o evaluación de clasificadores y modelos de regresión en un estudio o investigación particular. Con esta información analizada, se obtuvieron los modelos de aprendizaje automático más utilizados y eficientes de acuerdo con las diferentes métricas consideradas.

3. Resultados

Los resultados obtenidos a partir de la aplicación del método seleccionado se describen a continuación. Para llevar a cabo esta investigación sobre inteligencia artificial y los modelos de aprendizaje automático, se realizó un proceso exhaustivo de búsqueda y recopilación de artículos relacionados con el tema. Inicialmente, se identificaron aproximadamente 400 artículos relevantes en diversas fuentes académicas y bases de datos especializadas.

Es importante destacar que el proceso de selección y clasificación de estos artículos se llevó a cabo con un criterio riguroso para asegurar la inclusión de los trabajos más relevantes y pertinentes al tema de investigación.

Tras una revisión exhaustiva, se clasificaron un total de 120 artículos que cumplían con los criterios de selección. En el enlace a continuación se encuentra la tabla con los artículos revisados: https://docs.google.com/spreadsheets/d/1L11_OWo_XdHS4jUirj7cvfJKTE6v9-Lv6PXXtcaM/edit?usp=sharing

Una vez identificados los artículos relevantes, se procedió a la elección de aquellos específicos que serían sometidos a un análisis detallado y a la clasificación de los modelos de inteligencia artificial y aprendizaje automático que se alineaban de manera más precisa con los objetivos de nuestra investigación. En esta fase, Datalor desempeñó un papel crucial al proporcionar herramientas y recursos para obtener resultados precisos y confiables. La plataforma Datalor permitió realizar un análisis detallado de cada modelo, evaluando su idoneidad para nuestro estudio en función de diversos criterios, como su capacidad de adaptación a conjuntos de datos específicos, su eficacia en tareas particulares y su relevancia en el contexto de nuestra investigación.

Datalore fue esencial en la demostración gráfica de nuestros resultados. La plataforma brindó la capacidad de visualizar de manera efectiva y atractiva los datos y conclusiones clave de nuestro estudio. Esto no solo facilitó la comprensión de los resultados por parte de nuestro público objetivo, sino que también enriqueció la presentación de nuestra investigación con representaciones visuales. Gracias a las capacidades de visualización de datos de Datalore, se pudo destacar patrones, tendencias y hallazgos de manera convincente, lo que fortaleció aún más la credibilidad de la investigación.

A continuación, se presentan los resultados de manera estructurada y visualmente atractiva, utilizando tablas y gráficas para transmitir de manera efectiva la información recopilada.

En la tabla 2 muestra que en estas investigaciones se identificaron 13 modelos en el aprendizaje automático en las búsquedas de los artículos de los cuales, es importante notar que algunos modelos aparecen más frecuentemente que otros en el conjunto de datos analizado.

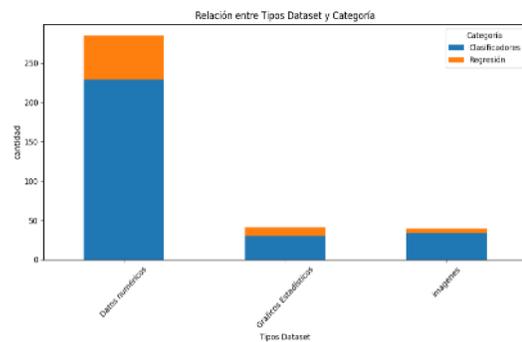
Entre los modelos más destacados, se encuentra el de Máquinas de Vectores de Soporte (SVM), que aparece con mayor frecuencia, siendo mencionado en 73 ocasiones. Le sigue Random Forest (RF) con 57 menciones, y la Red Neuronal Artificial (ANN) con 43 menciones, destacándose como modelos ampliamente utilizados en las investigaciones revisadas. Por otro lado, K-Means es el modelo menos frecuente, mencionado en tan solo 2 ocasiones en el conjunto de datos analizado, lo que indica su menor presencia en los estudios revisados. Este análisis proporciona una visión clara de la popularidad y relevancia de los diferentes modelos de aprendizaje automático en la literatura revisada.

Tabla 2 – Cantidad de modelos identificados y sus frecuencias encontradas.

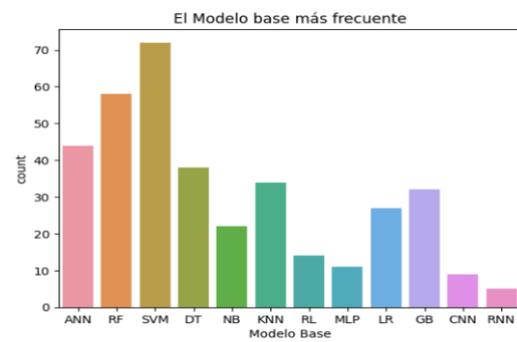
Modelo Base	Frecuencias
Máquinas de vectores de soporte (SVM)	73
Random Forest (RF)	57
Red Neuronal Artificial (ANN)	43
Árbol de decisión (DT)	36
K-vecinos más cercanos (KNN)	34
Gradient Boosting (GB)	27

Regresión logística (LR)	27
Naive Bayes (NB)	22
Aprendizaje por refuerzo (RL)	13
Perceptrón multicapa (MLP)	11
Red neuronal convolucional (CNN)	8
Red neuronal Recurrente (RNN)	5
K-Means	2

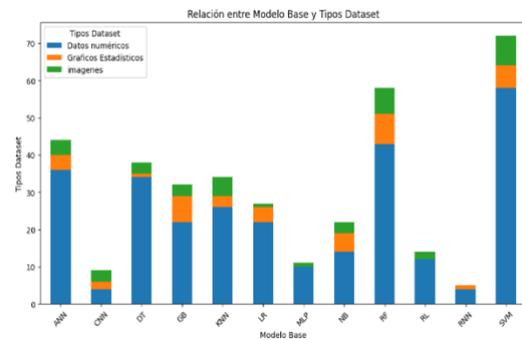
Figura 1. Análisis de modelos de aprendizaje automático (AA) frente a diversas variables.



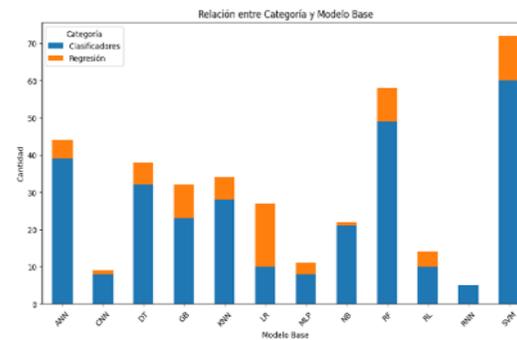
a. Datasets y categorías



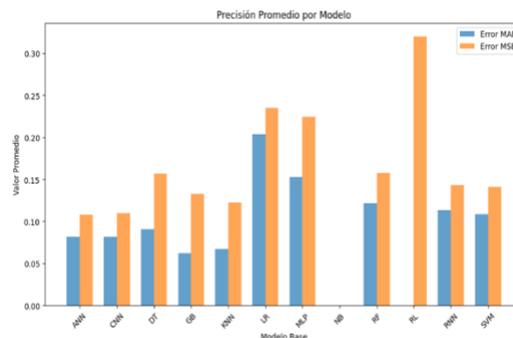
b. Modelos de aprendizaje automático (AA)



c. Modelos de AA y tipos de dataset



d. Modelos de AA y categoría



e. Precisión de los modelos de AA

En la Figura 1.b se presenta una representación gráfica de las frecuencias de los modelos identificados en los artículos analizados. Se observa que la frecuencia de mención de estos modelos varía considerablemente, lo que proporciona una visión interesante de su prevalencia en la literatura revisada, además como los datos con respecto a las métricas como Precision, Accuracy, F1, Recall, Error MAE, Error MSE, G-mean estas métricas permiten medir diferentes aspectos del desempeño de un modelo, como su capacidad para clasificar correctamente, su nivel de precisión y su capacidad para minimizar errores.

En conjunto, estos resultados reflejan la diversidad de modelos utilizados en la investigación de inteligencia artificial y aprendizaje automático. Algunos modelos son claramente más prevalentes en la literatura científica que otros. La recopilación de frecuencias y métricas proporciona información valiosa sobre las tendencias en la selección de modelos en este campo, lo que podría influir en las decisiones futuras de investigación y servir como guía para investigaciones posteriores.

Tabla 3 – Análisis de los tipos de datos.

Tipos de datos	Frecuencias	Porcentajes	Análisis
Datos Numéricos	99	74%	Valores cuantitativos utilizados en cálculos matemáticos, y predicción y toma de decisiones.
Imágenes	18	14%	Datos visuales que representan una escena o un objeto.
Gráficos estadísticos	16	12%	Representación visual de datos numéricos utilizando gráficos.

La Tabla 3 presenta una visión detallada de los tipos de datos fundamentales que se han utilizado en el contexto del aprendizaje automático. Estos resultados proporcionan una valiosa perspectiva sobre la naturaleza de los datos empleados en nuestro análisis.

En primer lugar, los datos numéricos son el tipo de dato más prevalente, representando un impresionante 74% del conjunto de datos. Esta predominancia es coherente con la importancia de los datos cuantitativos en numerosos campos de estudio, y subraya su relevancia en el análisis y la toma de decisiones impulsados por el aprendizaje automático. Por otro lado, las imágenes representan el 14% del conjunto de datos, desempeñando un papel crucial en áreas que requieren análisis visual, como el procesamiento de imágenes y la visión por computadora. Estos datos visuales tienen aplicaciones significativas en el análisis exploratorio y en la resolución de problemas complejos. Finalmente, los gráficos estadísticos contribuyen con un 12% del conjunto de datos. Estos gráficos son esenciales para representar y comunicar eficazmente patrones y tendencias en los datos, lo que los convierte en una herramienta valiosa en el análisis y la presentación de resultados.

La presencia de esta diversidad de tipos de datos subraya la complejidad y la riqueza de la investigación en el campo del aprendizaje automático. Además, resalta la importancia de seleccionar las metodologías y algoritmos adecuados según el tipo de datos con el que se esté trabajando, lo que puede tener un impacto significativo en la precisión y la eficacia de los modelos de aprendizaje automático desarrollados.

En la tabla 4 se observa una clara tendencia en su uso. La mayoría de estos modelos se emplean principalmente en tareas de clasificación, siendo esta la aplicación más frecuente con un total de 93 menciones. Esto sugiere

que el proceso de categorizar y etiquetar datos es una aplicación esencial y ampliamente explorada en la literatura revisada.

Por otro lado, la aplicación de regresión también es relevante, aunque menos frecuente en comparación con la clasificación. Se registraron 23 menciones de modelos utilizados en tareas de regresión, lo que indica que la predicción de valores numéricos es una aplicación importante, pero menos común en los estudios revisados.

Tabla 4 – Análisis de los tipos de aplicación del modelo.

Tipo de Aplicaciones de modelos	Frecuencias	Porcentajes	Análisis
Clasificación	93	80%	Asigna una etiqueta o categoría a una instancia basada en características predefinidas.
Regresión	23	20%	predice un valor numérico o continuo basado en variables independientes.

En la Figura 3, muestra una representación gráfica en forma de pastel que ilustra los porcentajes relativos de las categorías en los modelos de inteligencia artificial y se observa que la clasificación es la aplicación más predominante, abarcando aproximadamente el 80% de los casos, mientras que la regresión representa el 20% restante.

Esta visualización permite apreciar de manera efectiva la distribución de estas aplicaciones en la muestra de estudios. A pesar de estas diferencias en frecuencia, es importante destacar que cada una de estas aplicaciones

desempeña un papel valioso en el análisis de datos. La clasificación y la regresión, aunque utilizadas en diferentes contextos, contribuyen de manera significativa a la comprensión y toma de decisiones basadas en datos en diversos campos de estudio.

En el análisis de los tipos de datos por clasificadores y regresión, muestra que la mayoría de las tareas realizadas fueron clasificaciones en todos los tipos de datos, mientras que las regresiones fueron menos frecuentes como se muestra en la tabla 5.

Tabla 5 – Tipos de datos por clasificadores y regresión.

Tipos de datos	Clasificadores	Regresión
Datos Numéricos	229	56
Imágenes	30	11
Gráficos estadísticos	34	6

En la Figura 1.a con base en el análisis bivariado del tipo de dataset y categoría muestra que la mayoría de las tareas realizadas fueron clasificaciones en todos los tipos de datos, mientras que las regresiones fueron menos frecuentes. Esto puede deberse a la naturaleza de los datos y las necesidades específicas de los proyectos, además, se podría profundizar en el rendimiento y la eficacia de cada clasificador en diferentes tipos de datos, realizando una evaluación más detallada para determinar cuál de los clasificadores se adapta mejor a cada tipo de problema.

En el análisis bivariado del modelo base y la categoría, se observa que la mayoría de las tareas realizadas corresponden a clasificaciones en todos los tipos de datos, por otro lado, las regresiones son menos frecuentes en

comparación, esta tendencia puede explicarse por la naturaleza de los datos y las necesidades específicas de los proyectos, para obtener una comprensión más profunda del rendimiento y la eficacia de cada clasificador en diferentes tipos de datos, como se muestra en la figura 1.d. Esto implicó llevar a cabo el análisis más específico para determinar cuál de los clasificadores se adapta mejor a cada tipo del modelo base.

En la figura 1.c, se presenta un análisis bivariado que relaciona el modelo base empleado en tareas de análisis de datos y aprendizaje automático con el tipo de dataset utilizado, uno de los hallazgos destacados es que los datos numéricos se utilizaron con mayor frecuencia en comparación con otros tipos de datos, esto sugiere una preferencia generalizada por trabajar con conjuntos de datos que contienen variables numéricas, posiblemente debido a su versatilidad y facilidad de procesamiento. Por otro lado, se observa que los gráficos estadísticos y las imágenes fueron menos utilizados en las tareas realizadas, esta diferencia en la frecuencia de uso puede indicar que los proyectos se han centrado en gran medida en el análisis cuantitativo de datos numéricos, mientras que las visualizaciones y el procesamiento de imágenes pueden no haber sido tan prominentes en las necesidades específicas de análisis.

La Tabla 6 proporciona una visión detallada de las métricas de error para cada uno de los modelos de nuestra investigación. Estas métricas son esenciales para comprender el rendimiento y la precisión de cada modelo en el contexto de nuestra aplicación específica.

La figura 1.e muestra de manera gráfica los datos, y con base a la tabla y a la gráfica determinar cuál de los modelos es el más eficiente. Al evaluar los

errores de los dos modelos en cuestión, el Error Absoluto Medio (MAE) y el Error Cuadrático Medio (MSE), llegamos a las siguientes conclusiones:

Los resultados del análisis indican que el modelo de Gradient Boosting (GB) sobresale de manera significativa al presentar el menor Error Absoluto Medio (MAE) en comparación con los otros modelos evaluados. Esta métrica de evaluación pone de manifiesto la destacada capacidad del modelo de Gradient Boosting para realizar predicciones altamente precisas y cercanas a los valores reales. Estos hallazgos establecen a Gradient Boosting como la opción más prometedora para aplicaciones futuras y la toma de decisiones fundamentadas, especialmente si se busca minimizar los errores en las predicciones en general y se está dispuesto a penalizar de manera más significativa los errores grandes.

Tabla 6 – Promedio de cada métrica

Modelo Base	Error MAE	Error MSE
Máquinas de vectores de soporte (SVM)	0.108307	0.141327
Random Forest (RF)	0.121481	0.157329
Red Neuronal Artificial (ANN)	0.081778	0.108028
Árbol de decisión (DT)	0.090929	0.157051
K-vecinos más cercanos (KNN)	0.067337	0.122570
Gradient Boosting (GB)	0.062300	0.132720
Regresión logística (LR)	0.204100	0.234764
Naive Bayes (NB)	nan	nan
Aprendizaje por refuerzo (RL)	nan	0.520000
Perceptrón multicapa (MLP)	0.153129	0.224917
Red neuronal convolucional (CNN)	0.081500	0.110000
Red neuronal Recurrente (RNN)	0.113333	0.143333

Además, en cuanto a las Redes Neuronales Artificiales (ANN), si bien también ofrecen un desempeño notable, su eficiencia se destaca principalmente cuando se considera el Error Cuadrático Medio (MSE), lo que las hace una elección sólida en escenarios donde se busca minimizar los errores de manera más generalizada, incluso a costa de penalizar los errores grandes.

4. Conclusiones

En este estudio, se realizó una exhaustiva revisión de modelos de aprendizaje automático y su aplicabilidad en una diversidad de categorías y tipos de conjuntos de datos. A través de un análisis meticuloso, se identificó patrones destacados en la elección de modelos por parte de la comunidad científica y profesional. Entre los hallazgos, se destaca la preeminencia del modelo de Máquinas de Soporte Vectorial (SVM) como el más frecuentemente empleado en una amplia variedad de aplicaciones. Sin embargo, nuestra investigación también ha revelado que, en términos de eficiencia y precisión, dos modelos, Gradient Boosting (GB) y Redes Neuronales Artificiales (ANN), sobresalen significativamente. Estos modelos se han destacado por su capacidad para ofrecer predicciones altamente precisas y cercanas a los valores reales, estableciéndose así como herramientas a considerar para futuras investigaciones y aplicaciones en el campo del aprendizaje automático. En un panorama en constante evolución de la inteligencia artificial, la elección estratégica de estos modelos no sólo puede mejorar la calidad de las predicciones, sino también impulsar la toma de decisiones fundamentadas en diversas disciplinas.

Los modelos de aprendizaje automático son altamente sensibles a la calidad y naturaleza de los datos con los que se entrenan, la preparación y limpieza adecuadas de los datos, junto con la selección de características relevantes, pueden tener un impacto significativo en el rendimiento del modelo, diferentes tipos de datos, como datos numéricos, categóricos, de texto o de imágenes, requieren enfoques de preprocesamiento específicos. Además, la recopilación de una cantidad suficiente y representativa de datos es esencial para evitar problemas de sobreajuste, este estudio muestra que el aprendizaje automático y la inteligencia artificial están en constante evolución y se aplican en una diversidad de áreas. Los modelos utilizados son diversos y se adaptan a los diferentes tipos de datos disponibles. Esta versatilidad demuestra el potencial de estas tecnologías para abordar una amplia gama de problemas y oportunidades en la actualidad y en el futuro. Además, se subraya la importancia de seguir investigando y desarrollando modelos y aplicaciones de aprendizaje automático para aprovechar al máximo el poder transformador de la inteligencia artificial en diversas industrias.

Este estudio proporciona una visión integral de la aplicación de modelos de inteligencia artificial y aprendizaje automático en la investigación actual. Los modelos SVM, RF y ANN han demostrado ser herramientas poderosas para clasificar y predecir datos numéricos en una variedad de contextos. Sin embargo, se reconoce la necesidad de abordar las limitaciones metodológicas identificadas y de explorar nuevas oportunidades para mejorar la precisión y la aplicabilidad de estos modelos. Se sugiere como trabajo futuro la integración de métodos adicionales de sistematización y la evaluación continua de métricas de rendimiento para avanzar en el campo del aprendizaje automático y su impacto en decisiones basadas en datos.

5. Referencias

- Arnau Sabatés, L., & Sala Roca, J. (2020, Abril 23). La revisión de la literatura científica: Pautas, procedimientos y criterios de calidad. https://ddd.uab.cat/pub/recdoc/2020/222109/revliltcie_a2020.pdf
- Beunza Nuin, J. J., Sanz, E. P., Vila, B. R., Moreno, E. C., & Sanz, J. B. (2023). Manual práctico de inteligencia artificial en entornos sanitarios (DRK edicion ed.). Elsevier España, S.L.U. https://books.google.com.ec/books?id=aQWtEAAAQBAJ&newbks=1&newbks_redir=0&printsec=frontcover&hl=es#v=onepage&q&f=false
- Carrizo, D., & Moller, C. (2018, Agosto 06). Estructuras metodológicas de revisiones sistemáticas de literatura en Ingeniería de Software: un estudio de mapeo sistemático. *Revista chilena de ingeniería*, 26, 45-54. <https://www.scielo.cl/pdf/ingeniare/v26s1/0718-3305-ingeniare-26-00045.pdf>
- Herrera Zurita, A. (2016, 06). Aprendizaje Automático para la detección de ataques informáticos. <https://core.ac.uk/download/pdf/78544263.pdf>
- LOZANO ORTEGA, M. A., COLOMINA PARDO, O., ESCOLANO RUIZ, F., CAZORLA QUEVEDO, M. A., & ALFONSO GALIPIENSO, M. I. (2003). *Inteligencia artificial*. Ediciones Paraninfo, S.A. https://books.google.com.ec/books?hl=es&lr=&id=_spC6S7UfZgC&oi=fnd&pg=PP1&dq=articulos+de+Inteligencia+artificial&ots=sQmsJFNuBW&sig=NKwzb6z0DGe4pfRQnD6Rdb63Z28#v=onepage&q&f=false
- Ocaña Fernández, Y., Valenzuela Fernández, L. A., & Garro Aburto, L. L. (2019, mayo/agosto). *Inteligencia artificial y sus implicaciones en la educación superior. Propósitos y Representaciones*, volumen 7. <http://dx.doi.org/10.20511/pyr2019.v7n2.274>
- Pineda Pertuz, C. M. (2022). *Aprendizaje automático y profundo en Python: una mirada hacia la inteligencia artificial*. Ra-Ma Editorial.

https://books.google.es/books?hl=es&lr=&id=NEi9EAAAQBAJ&oi=fnd&pg=PA7&dq=aprendizaje+automatico+y+profundo+en+python&ots=b2D58vflb&sig=Yun_SbwxLBwDGm6E2sBAXUuFYp0#v=onepage&q&f=false

Sandoval, L. J. (2018, Julio 19). MACHINE LEARNING ALGORITHMS FOR DATA ANALYSIS AND PREDICTION. REVISTA TECNOLÓGICA. http://redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Tandon, A., salimath, N., Bhatia, s., Sethi, K., & tripathy, B. k. (2019). Introduction to Machine Learning. Book Bazooka. https://www.google.com.ec/books/edition/Introduction_to_Machine_Learning/3YTADwAAQBAJ?hl=es-419&gbpv=1&dq=machine+learning&printsec=frontcover.

Garcia Falckenheiner, A. E., Villanes Rojas, S., Cerna Barco, R. A., Felipe Mori, M. A., Paliza Champi, L. E., & Pajuelo Aguirre, P. (2023). IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LA EDUCACION SUPERIOR. *AD MAJOREM PATRIAE GLORIAM*, 6(1). <https://doi.org/10.61556/ampg.v5i06.71>

García Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2023). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia*, 27(1). <https://doi.org/10.5944/ried.27.1.37716>

Orozco Iguasnia, W. A., Villao Balón, A. J., Orozco Iguasnia, J., & Villarreal Sánchez, M. (2021). Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela 'Lic. Angélica Villón L.' *Revista Científica y Tecnológica UPSE*, 8(2). <https://doi.org/10.26423/rctu.v8i2.637>

Solórzano Álava, W. L., Rodríguez Rodríguez, A., Anzules Ávila, X. L., & Cornelio, O. M. (2022). Impacto del uso de la tecnología en la formación integral de los estudiantes de la carrera tecnologías de la información.

Journal TechInnovation, 1(2), 71–77.

<https://doi.org/10.47230/journal.techinnovation.v1.n2.2022.71-77>

Tamayo, S., & Pérez Marín, D. (2017). Propuesta de Evaluación basada en Big Data para facilitar la integración de Agentes Conversacionales Pedagógicos en las aulas. *IE Comunicaciones: Revista Iberoamericana de Informática Educativa*, 26.