PREDICCIÓN DEL DESEMPEÑO ESCOLAR EN NIVEL SUPERIOR. UN ANÁLISIS DESCRIPTIVO

PREDICTION OF ACADEMIC PERFORMANCE IN HIGHER LEVEL STUDENTS. A DESCRIPTIVE ANALYSIS

Abraham Santiago Martínez,[1] Vicente Josué Aguilera Rueda[2] y César Augusto Mejía Gracía[3]

RESUMEN

La predicción del desempeño escolar en estudiantes de nivel superior, entre otras ventajas, permite identificar tempranamente a los estudiantes que se encuentran en riesgo y reconocer las áreas de mejora del sistema educativo. Con algoritmos de inteligencia artificial es posible encontrar patrones en los datos que indiquen cuando un estudiante está en riesgo académico e intervenir a tiempo según el tomador de decisiones considere. Esta investigación presenta un análisis descriptivo e indagatorio de las publicaciones relacionadas con la predicción del desempeño escolar en Iberoamérica entre los años 2018 a 2022. Se analizaron las siguientes variables: año de publicación, algoritmos, tamaño de la muestra, resultados y factores relacionados con el desempeño escolar. Los resultados muestran que los algoritmos más utilizados

ABSTRACT

Among other benefits, predicting school performance in higher-level students allows for the early identification of at-risk students and the recognition of areas for improvement within the educational system. With artificial intelligence algorithms, it is possible to uncover patterns in the data that indicate when a student is at academic risk and intervene on time as decision-makers deem. This research presents a descriptive and exploratory analysis of publications on school performance prediction in Ibero-America between 2018 and 2022. The following variables were analyzed: year of publication, algorithms, sample size, results, and factors related to school performance. The results show that the most commonly used algorithms are classification trees and artificial neural networks, with varied sample sizes related

[1] Es egresado de la licenciatura en Sistemas Computaciones Administrativos en la Universidad Veracruzana, México. Su trabajo recepcional de dicho grado está relacionado con la inteligencia artificial en el ámbito educativo. https://orcid.org/0009-0006-1818-8410

[2] Es profesor en la Facultad de Contaduría y Administración en la Universidad Veracruzana, México, y profesor en la Escuela de Ingenierías de la Universidad de Xalapa. Ha publicado en revistas de alto impacto y participado como ponente en congresos internacionales. https://orcid.org/0000-0002-1952-7860

[3] Formado en Sistemas Computacionales Administrativos, con maestría y doctorado en Administración. Docente en la Universidad Veracruzana, México. Miembro de SNI y AMIAC. https://orcid.org/0000-0001-8874-0473.

DOI: https://doi.org/10.5281/zenodo.10357981

RECIBIDO: 21 de agosto, 2023
ACEPTADO: 13 de septiembre, 2023

son los árboles de clasificación y las redes neuronales artificiales, con tamaños de muestra variados relacionados con el número de variables analizadas, la cuales corresponden mayormente a factores socioeconómicos y cognitivos.

PALABRAS CLAVE: desempeño académico, inteligencia artificial, Instituciones de nivel superior.

to the number of analyzed variables, which mostly correspond to socioeconomic and cognitive factors.

KEYWORDS: academic performance, artificial intelligence, higher educational institution.

I. INTRODUCCIÓN

En los últimos años, las instituciones de educación superior (IES) han prestado más atención e invertido recursos para tareas que potencien el proceso de mejora de la calidad en la educación, puesto que la educación es uno de los componentes efectivos del desarrollo social y es una de las principales metas de los países.

En el marco de la educación y la formación, se encuentran innumerables intentos de comprender e interpretar el funcionamiento escolar desde diferentes ángulos, como la pedagogía, la sociología, la psicología y, más recientemente, la neurociencia, entre otras. Todos estos esfuerzos han proporcionado elementos importantes para la formulación de estrategias destinadas a mejorar la calidad de la educación y la formación de los estudiantes no solo en nivel superior, sino en todos los niveles educativos (Lemus et al., 2021).

El rendimiento académico de los estudiantes es un factor crucial, ya que los requisitos actuales demandan una educación integral con bajos niveles de deserción y altas tasas de retención y finalización de los estudios. Al graduarse, se espera que los estudiantes sean excelentes intérpretes, con habilidades y destrezas ideales para ejercer su disciplina (Ibañez Reyes et al., 2020).

Por lo que medir el aprendizaje de los estudiantes universitarios es importante para las IES porque ayuda a determinar el crecimiento académico. Por esta razón, conocer el desempeño escolar futuro de un estudiante permite tomar las decisiones correctas para que pueda mantener un buen progreso en su rendimiento escolar.

En cierta medida, el desempeño académico de los estudiantes no solo se ve influenciado por factores sociodemográficos y socioeducativos, sino también por otros que no son considerados, como los aspectos emocionales de los estudiantes y sus relaciones familiares e interpersonales con compañeros de clase y profesores (Oviedo, 2019).

Diversas investigaciones se enfocan en identificar los factores que influyen en el desempeño escolar de los estudiantes, de todos los niveles, como los trabajos de Meléndez-Armenta, 2022; Morales et al., 2022; Muñoz & Fortoul, 2022, y Pinzón et al., 2023, y otros trabajos se han dado a la tarea de utilizar algoritmos de inteligencia artificial que, dados esos factores, tienen como fin predecir el desempeño escolar que tendrán los estudiantes. Particularmente, estos algoritmos son parte de lo que se conoce como aprendizaje automático; según Samuel (1959), se define como una rama de la inteligencia artificial que proporciona a las computadoras la capacidad de aprender sin necesidad de ser programadas explícitamente.

El aprendizaje automático es una disciplina que desde sus orígenes ha apoyado a muchas otras en la mejora de sus procesos, es el caso de diversos aspectos relacionados con los procesos educativos; en este sentido, el aprendizaje automático se ha utilizado para predecir el desempeño escolar de los estudiantes en las IES. Esta predicción se da desde la identificación *a priori* de una variable objetivo (usualmente conocida como variable "target") que puede ser de tipo categórica, por ejemplo; calificaciones altas o bajas, a lo que se le conoce como una tarea de clasificación o se puede dar como una tarea de regresión cuando la variable objetivo es de tipo continúa; es decir, cuando se desea predecir el valor concreto de la calificación que tendrá el estudiante, por ejemplo 8.9 o 7.5.

En relación con los esfuerzos de las investigaciones que se han publicado respecto a la predicción del desempeño escolar en IES, hasta donde se conoce, pocos han actualizado el estado del arte de estas investigaciones en países iberoamericanos (Cruz et al., 2022). Este trabajo tiene como fin principal recopilar, describir, identificar y listar los trabajos que se han publicado en los últimos años en relación con la predicción del desempeño escolar en IES en Iberoamérica.

Este documento está organizado de la siguiente manera: en la sección del marco teórico se encuentran los conceptos más importantes de aprendizaje automático, así como los algoritmos más utilizados en la tarea de predicción del desempeño escolar. La sección de materiales y métodos pone de manifiesto la metodología empleada para la búsqueda y descripción de los trabajos que se han publicado sobre el tema de interés. En la sección de desarrollo se documentan los trabajos que se han publicado entre los años 2018 a 2022 en Iberoamérica donde se describen los objetivos, metodología y resultados a los que llegaron.

II. MARCO TEÓRICO

El aprendizaje automático se divide en cuatro tipos: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje por refuerzo. Este trabajo se sitúa en el campo del aprendizaje supervisado; si el lector está interesado en otro tipo de aprendizaje automático, se recomienda que revise a Mahesh (2020).

En el aprendizaje supervisado, según Sorias Olivas et al. (2023), interviene un agente externo que se encarga de asignar el valor o nombre de las etiquetas previamente

establecidas y el algoritmo emparejará los datos a la etiqueta que pertenezca.

Desde una perspectiva formal, dado un conjunto de datos, el aprendizaje supervisado encuentra patrones que relacionen los datos con las etiquetas, y a su vez encuentra nuevos datos que no se hayan observado antes.

El tipo de aprendizaje no supervisado se basa en encontrar patrones de similitud en un conjunto de datos, los algoritmos que utilizan este tipo de aprendizaje realizan la tarea a ciegas, por lo tanto, no hay un agente eterno que supervise el proceso de aprendizaje, por lo tanto, el algoritmo intenta encontrar similitudes entre el conjunto de datos y así realizar la clasificación de cada uno de ellos, este tipo de aprendizaje si bien es muy efectivo, puede presentar problemas de rendimiento, ya que los datos no están clasificados por etiquetas (Sorias Olivo et al., 2023).

Desde el aprendizaje automático se utilizan distintos algoritmos que se encargan de automatizar el proceso de predicción del desempeño escolar que tendrán los estudiantes durante o al final del curso, uno de los más utilizados son los árboles, que son los modelos que tiene mayor facilidad de interpretar los resultados, dentro de estos se encuentran los árboles de decisión y los Random Forest. Estos modelos son más precisos y complejos, y con ellos se obtiene una mejor representación visual de las variables de estudio, aunque el rendimiento en términos del tiempo no es el mejor de todos (Sandoval, 2018).

También se encuentran las redes neuronales, que son el modelo que mejor rendimiento ha presentado en los últimos años, en comparación con otros algoritmos. Estos modelos tratan de imitar al cerebro humano, donde hay millones de neuronas interconectadas, estos tipos de modelo suelen ser utilizados en la clasificación de imágenes, las redes neuronales son las que presentan un mayor rendimiento que otros algoritmos de aprendizaje automático. El propósito de estos patrones de conexión es permitirles comportarse como neuronas biológicas durante el aprendizaje imitando perfectamente el comportamiento lógicamente racional (Ráyon, 2017).

El algoritmo llamado K vecinos más cercanos o K-NN, es un modelo no paramétrico que utiliza las distancias cortas para clasificar o predecir a qué grupo pertenece un conjunto de datos individuales en función de etiquetas específicas. Para realizar la selección y clasificación de los datos, se realiza una comparación de distancias que hay entre un dato y otro (Raschka, 2018).

Las máquinas de soporte vectorial (SVM, por su nombre en inglés Support Vector Machines) reconocen dos tipos diferentes de superficies de decisión como punto de entrada. Como clasificador único, las descripciones proporcionadas por los datos del vector pueden formar límites de decisión en el dominio de los datos de entrenamiento, con pocos o ningún dato extraño en los límites (Betancourt, 2005).

Naïve Bayes es un algoritmo de aprendizaje simple que usa la regla de Bayes y asume que los atributos son condicionalmente

independientes de las categorías. Aunque esta suposición de independencia a menudo se viola en la práctica, Naïve Bayes generalmente proporciona una precisión de clasificación competitiva. El algoritmo proporciona un mecanismo para usar información en los datos de muestra para predecir la siguiente probabilidad P(y|x) para cada objeto x de clase y, una vez que se tienen estas predicciones, es posible usarlas para la clasificación u otras aplicaciones de soporte de decisiones (Webb, 2016).

En general, no son los únicos algoritmos que se han empleado en el campo de la predicción y la clasificación, sin embargo, los que se mencionan en este documento corresponden a los que se han utilizado para la predicción del desempeño escolar en el contexto de este estudio.

III. MATERIALES Y MÉTODOS

El presente documento se trata de un estudio descriptivo, indagatorio y considera una revisión sistematizada de las investigaciones que se han publicado en relación con la predicción del desempeño escolar de estudiantes de nivel universitario en Iberoamérica que abarcan el periodo de 2018 a 2022.

La búsqueda de artículos se realizó en las siguientes bases de datos internacionales:

- Google Académico
- IEEE Xplore
- ACM Digital Library
- SpringerLink

Las búsquedas se hicieron utilizando las siguientes palabras clave: predicción, desempeño escolar, aprendizaje automático + desempeño escolar.

Los artículos encontrados en la revisión bibliográfica corresponden al aprendizaje supervisado. Para cada uno de los artículos revisados se obtuvo la siguiente información:

- Año de publicación.
- Algoritmo o algoritmos empleados.
- Número total de registros utilizados para el entrenamiento y prueba del algoritmo.
- Porcentaje de predicción de cada algoritmo.
- Factores que se relacionan con el desempeño escolar.

La ilustración de resultados se hace a través de gráficos de frecuencia y se discuten las variables de estudio y sus coincidencias.

IV. DESARROLLO

A continuación, se describen los estudios que utilizan algoritmos de clasificación y que fueron publicados entre los años 2018 y 2022.

En el trabajo de Rico (2022), los autores desarrollaron un modelo predictivo progresivo del rendimiento escolar, para el que se utilizaron los registros de 260 estudiantes. El modelo se desarrolló en varias etapas durante el curso, lo que permitió una elección flexible de pasos para realizar la predicción. Los factores que utilizaron para la predicción del rendimiento escolar fueron 14 actividades que los estudiantes realizaron a lo largo del curso.

En su investigación, los autores probaron el rendimiento de tres algoritmos para clasificación: Naïve Bayes, K-NN, y Árbol de decisión. El algoritmo que mostró mejor desempeño en la clasificación fue Naïve Bayes con una precisión del 70.5%; a diferencias de los otros algoritmos utilizados, que obtuvieron una precisión del 68.6% aproximadamente.

En otro trabajo de Rico y Gaytán (2022), se construyó un modelo de predicción del rendimiento de 228 estudiantes de ingeniería. Para este trabajo se utilizaron los mismos algoritmos que en el anterior: K-NN, Naïve Bayes y Árbol de decisión.

Los autores consideraron varios factores para predecir el desempeño académico, tales como la educación de los padres, el ingreso familiar, el promedio en estudios anteriores, las materias reprobadas, el promedio actual, la preferencia de estudio y actividades, la frecuencia de estudio y la calificación en el curso. Según los resultados obtenidos, el modelo con mejor precisión fue el Naïve Bayes, con una exactitud del 65%, lo que permitió identificar a los estudiantes que tienen un mayor riesgo de reprobar.

En la tesis de Carmona (2022) se realizó la predicción del desempeño escolar por medio de una red neuronal utilizando perceptrón multicapa, y se empleó un base de datos de 23,000 estudiantes.

Para la realización de la red neuronal se utilizó el lenguaje de programación de Python y la biblioteca de Google llamada Tensorflow. La red neuronal se desarrolló en 9 etapas del rendimiento de los estudiantes durante 9 semestres, para los cuales se entrenaron modelos independientes a los que se les fue agregando paulatinamente la calificación y el índice de reprobación del semestre anterior, además en las etapas se consideraron los factores socioeconómicos de los estudiantes. La precisión del entrenamiento del modelo con más información, es decir el de la etapa 9, obtuvo una precisión del 98.27%.

En la publicación de Rodríguez-Hernández et al. (2021), se desarrolló un modelo de procesamiento sistemático para la implementación de una red neuronal utilizando los registros de 162,030 estudiantes de universidades públicas y privadas. Entre los factores que afectan al desempeño escolar que se consideraron en este estudio se encuentran: información socioeconómica, información del nivel educativo anterior de los estudiantes, además de información de su estado laboral.

En este modelo se clasificó a los estudiantes por alto y bajo rendimientos. La implementación sistemática de la red neuronal demuestra una mejor precisión en la clasificación con un 71% y 82%. Los modelos de Naïve Bayes, K-NN, regresión logística y Random Forest llegaron a presentar una precisión por debajo del 60%.

En el trabajo de Incio et al. (2021) se implementó una red neuronal para predecir el rendimiento académico de 50 estudiantes de la carreara de Ingeniería Civil utilizado una encuesta de coeficiente de confiabilidad de Alpha de Cronbach de 0.854 (Cronbach, 1951), de 18 preguntas sobre factores de orden cognitivo, emocional y socio-cultural.

La red neuronal desarrollada no puede ser aplicada en otras universidades, debido a que los factores que se toman en cuenta en la encuesta varían dependiendo de cada población y de las condiciones espacio-tiempo. Esta red neuronal utilizó dos algoritmos: en el primer entrenamiento se empleó Levenber-Marquardt (LM) que tiene una predicción del 86% y en el algoritmo Scaled Conjugate Gradient (SCG) se obtuvo una precisión del 70%.

En el trabajo de Díaz-Landa et al. (2021), se realizó la predicción de factores que influyen en el rendimiento escolar de 237 estudiantes de una universidad pública de Perú, mediante el uso del algoritmo de Árbol de decisión J48 ejecutado por WEKA (Witten et al., 2011).

El algoritmo obtuvo una precisión del 62.45% y los factores que más influyeron en el desempeño de los estudiantes fueron la pedagogía, adecuados horarios de clase, buena relación interpersonal docente-estudiante y la calidad académica.

En el trabajo de Urteaga et al. (2020), se aplicaron técnicas de aprendizaje automático para predecir la deserción escolar en cursos en línea; en este caso en la plataforma de Moodle, para este trabajo se utilizaron registros de 654 estudiantes. La métrica que este trabajo utilizó toma en cuenta el impacto que tiene el precio de un curso y la deserción de los estudiantes.

Se utilizaron varios algoritmos de clasificación de los cuales destacan la red neuronal y K-NN. En este caso, la red neuronal presentó un impacto del 92% y una precisión del 28% mientras que K-NN presentó un impacto del 94% y una precisión del 31%. Se resalta que los modelos independientes para cada curso presentan mejores resultados de precisión que el predecir utilizando más de un algoritmo a la vez.

En el trabajo de Urbina-Nájera et al. (2020), se aplicó la minería de datos para predecir la deserción de los estudiantes mediante la búsqueda de patrones. Para el desarrollo del estudio se utilizaron los registros de 300 estudiantes de escuela pública y 200 de escuela privada.

Para la predicción de la deserción escolar se tomaron en cuenta algunos factores que pudieran afectar en el rendimiento de los estudiantes, por ejemplo: datos demográficos, antecedentes familiares, medio ambiente, apoyo económico, entre otros.

Al final, se obtuvo una precisión del 92.6%. Entre los factores que afectan el desempeño de los estudiantes considerados en este trabajo se encuentran 27: falta de orientación, entorno estudiantil, falta de asesoramiento académico, entre otros. El algoritmo de Árbol de decisión encontró otros patrones, como el apoyo financiero, situaciones incómodas y las opciones de carrera, por mencionar algunos.

Para el trabajo de Leonor y Cazarez (2020), se realizó la predicción de estudiantes que realizan un curso en línea, utilizando red neuronal probabilística (RNP) y un análisis discriminatorio (AD). Para el desarrollo del trabajo se utilizó una muestra de 1,181 estudiantes que se dividieron en dos experimentos de los cuales se tomaron las

calificaciones que obtuvieron en 4 cursos diferentes. En el primer experimento se utilizaron 693 registros y en el segundo experimento se eliminaron los registros de los estudiantes que ya no continuaron con el curso, siendo una base de datos de 488. Los dos experimentos se probaron en las 4 materias con diferentes números de estudiantes, donde la RNP demostró tener mejores resultados de precisión en cinco de los ocho conjuntos de datos, en contraste con el AD.

En el trabajo de Contreras et al. (2020) se realizó la predicción de éxito o fracaso de estudiantes de ingeniería utilizando SVM, árboles de decisión, K-NN, RNA y el lenguaje de programación de Python. En el estudio también se tomaron en cuenta los factores que pueden afectar el rendimiento escolar de los estudiantes. Se usaron los registros de 1620 estudiantes, donde se encontró que los factores que más repercusión tienen en el rendimiento escolar son la edad, género, habilidades matemáticas, condición matemática. El modelo que presentó mejores resultados de precisión fue SVM y la RNA con una precisión por arriba del 66.4%.

En el trabajo de Castrillón et al. (2020), se realizó la predicción de rendimiento académico de 121 estudiantes de una universidad pública. Se utilizó el *software* WEKA que tomó en cuenta algunas variables que pueden afectar el rendimiento de los estudiantes. Se alcanzó una precisión del 91.7% y se encontró que las variables que más afectan en el rendimiento académico son los métodos de enseñanza de los profesores, horarios de clase, relación docente-alumno, calidad de aprendizaje de los docentes y el exceso de tareas.

En la tesis de Candia (2019) se realizó la predicción de 12,698 estudiantes utilizando el registro de sus datos de ingreso, con la ayuda del modelo CRISP-DM y la plataforma WEKA, y los algoritmos de Árbol de decisión J48, Random Forest, K-NN, entre otros. El algoritmo que presentó una mayor precisión fue el Random Forest con un 69.3%, donde se encontró que los factores que afectan el rendimiento académico son: la clase de ingreso, preparatoria a la que asistió, semestre, género y método de inscripción.

En el trabajo de Rico et al. (2019) se construyó e implementó un modelo para predecir el rendimiento académico mediante los registros de aprobación y reprobación de 71 estudiantes de Ingeniería. Se utilizó un cuestionario donde se recopilaron algunos factores, como la escolaridad de los padres, ingresos en la familia, promedio final de bachillerato, promedio actual y hábitos de estudio. Para la predicción del rendimiento de los estudiantes se utilizó el algoritmo de Naïve Bayes que arrojó una precisión del 70.42%, el estudio también resalta que algunos de los factores como los hábitos de estudio o el trabajo en equipo influyen en el rendimiento de los estudiantes.

En el trabajo de Rico y Sánchez (2018) se desarrolló un modelo que automatiza la predicción del rendimiento escolar de 86 estudiantes inscritos a una carrera de ingeniería del Instituto Politécnico Nacional (IPN). Mediante la minería de datos y el algoritmo de clasificación Naïve Bayes,

se obtuvo una precisión del 73%, para ello se tomaron en cuenta cinco atributos correspondientes a las calificaciones de las actividades que se hicieron en el curso.

En la tesis de Cabana (2018) se realizó la predicción del desempeño escolar de 69 estudiantes de Ingeniería con una red neuronal. Se analizaron las variables de razonamiento verbal, aritmética y álgebra, física, química, razonamiento matemático, lógica, geometría y trigonometría, y lengua. El modelo tuvo una precisión del 95% y durante el entrenamiento se obtuvo un error del 6.2%; se encontró que el razonamiento matemático es el factor principal que afecta el rendimiento de los estudiantes.

A continuación, se presentan los trabajos que abordan la predicción del desempeño escolar utilizando algoritmos de regresión.

El trabajo de Layza (2020) se realizó con un modelo matemático, utilizando factores que puedan afectar el rendimiento de 66 estudiantes de un curso de matemáticas, para la predicción se tomaron en cuenta los registros de exámenes y la calificación final del curso. Se utilizaron modelos lineales y de regresión, donde se tomaron en cuenta los factores cognitivos, motivacional, socio-ambiental, institucional e instruccional, calificaciones de finales del curso y la del examen.

Los modelos lineales demostraron que sí existe una relación entre los factores con el rendimiento de los estudiantes, lo que significa que si se presta atención a los factores se obtendrá un mayor desempeño escolar.

En el trabajo de Lemus et al. (2021), su objetivo fue definir un indicador académico estudiantil basado en los créditos académicos de 288 estudiantes, esto con el fin de tener una alternativa al indicador por promedio de calificaciones. Estos dos indicadores se compararon tomando en cuenta variables sociodemográficas, académicas y motivacionales.

Se compararon modelos de regresión lineal y logística y modelos de ecuaciones estructurales, los modelos lineales demostraron una pérdida de potencia al categorizar los indicadores; por otro lado, los modelos de ecuaciones estructurales que utilizan agrupación de ítems son una alternativa a los modelos de rutas. Dentro de las variables revisadas se llegó a la conclusión de que los aspectos demográficos y socioeconómicos y el estado de salud son lo que presentan una mayor repercusión en el rendimiento de los estudiantes.

En el trabajo de Martínez et al. (2021), Se llevó a cabo un análisis predictivo para la deserción escolar de 87 estudiantes de la carrera de medicina, mediante la aplicación de dos métodos. El primer método consistió en un análisis bivariado, que permitió identificar las variables que presentaban asociación con la deserción escolar. El segundo método fue un análisis multivariado, el cual se utilizó para analizar la capacidad de predicción de estas variables.

Luego de aplicar modelos de correlación y regresión logística, se identificaron las variables que presentaban una relación significativa con la deserción escolar. De las

9 variables analizadas, 4 mantuvieron una relación constante, por lo que se utilizaron como variables predictoras. Estas variables fueron: dedicarle menos de 15 horas por semana al estudio, ser de sexo femenino, haber repetido cursos previamente y tener un bajo rendimiento académico en la materia de Morfofisiología.

En el trabajo de Martínez et al. (2019), se evaluó la correlación lineal de los hábitos de estudio y la autoestima con el desempeño escolar de 153 estudiantes de nuevo ingreso. Para la recopilación y diagnóstico de los hábitos de estudio se utilizó el instrumento del Inventario de Técnicas y Actitudes de Estudio (ITAE), donde se consideraron las siguientes categorías: concentración, relaciones interpersonales, memoria, motivación para estudiar, administración del tiempo, y presentación de evaluaciones. Se utilizaron modelos de regresión lineal múltiple y el modelo de agrupación K-medias. Los resultados de este trabajo demostraron que existe una baja a moderada relación entre las variables de hábitos de estudio y autoestima, ante el desempeño escolar. Por otra parte, el algoritmo K-medias sí demostró una relación entre las variables y el desempeño escolar.

En el trabajo de Martínez-González et al. (2018), se realizó un análisis estadístico sobre el grado de conocimiento de 27,624 estudiantes de primer ingreso mediante una evaluación diagnóstica de 120 reactivos utilizando un algoritmo lineal logístico.

Los estudiantes se clasificaron en dos etapas, la primera etapa se dividió por el total de aciertos en la evaluación diagnóstica: bajo, menor a 35.9% de aciertos; medio bajo, entre 36 a 41.9%; medio, de 42 a 46.9%; medio alto, de 47 a 53.9%, y alto, mayor a 54%.

La segunda etapa se dividió por desempeño escolar durante el curso tomando en cuenta los créditos obtenidos durante este: abandono 0% de créditos, rezago extremo de 1 a 25%, rezago alto de 26 a 50%, rezago intermedio de 51 a 75%, rezago recuperable de 76 a 99% y egreso o regularidad 100%. En el estudio se llegó a la conclusión de que los conocimientos previos que se obtienen en el bachillerato son el factor principal que afecta el desempeño de los estudiantes; también se demostró que obtener un 44.2% de aciertos durante las evaluaciones se consideran insuficientes para enfrentar los retos académicos de las licenciaturas de la universidad. El trabajo en cuestión hace uso de técnicas clásicas para la predicción del desempeño escolar.

En el trabajo de González-Beltrán et al. (2018), se realizó un análisis de impacto sobre cursar el "Taller de matemáticas" en el rendimiento escolar de los estudiantes durante el curso de Matemáticas Suplementarias e Introducción al Análisis. Este análisis se realizó para saber si es indispensable tomar el curso, aprobarlo por examen o eliminarlo del plan de estudios. Del análisis de los datos de 5,121 estudiantes analizados se presentaron las siguientes premisas: los estudiantes que no asistan al "Taller de matemáticas" por haber aprobado la prueba diagnóstica, tendrán un buen desempeño en los cursos de Matemáticas Complementarias e Introducción al Análisis. A los estudiantes que les vaya bien en el Taller de

Matemáticas también les irá bien en los cursos de Matemáticas Suplementarias e Introducción al Análisis. Los resultados demostraron que el contenido del "Taller de matemáticas" se debe de analizar para saber por qué no es útil para futuros cursos. En la siguiente sección se presentan lo resultados como consecuencia de un análisis descriptivo de los artículos que se describieron en esta sección.

V. PRESENTACIÓN Y DISCUSIÓN DE RESULTADOS

Como se mencionó en la introducción, el aprendizaje automático está conformado por diversas tareas; típicamente la regresión, cuando la variable de interés es de tipo continua, la tarea de clasificación, cuando la variable de interés es de tipo categórica, adicionalmente en esta revisión se encontró al menos un trabajo que utilizó técnicas clásicas de regresión estadística. En la Figura 1, es posible observar cómo los algoritmos de clasificación son los más empleados, en este contexto, entre los años 2018-2022. En algunas ocasiones es deseable conocer cuál será el desempeño de los estudiantes usando una variable clase como: rendimiento-ato, rendimiento-medio y rendimiento-bajo; lo cual permite tomar decisiones, como la implementación de cursos remediales, actualización de contenidos en los programas de estudio, actualización de planes de estudio o incluso para determinar la correlación entre los factores que están afectando el desempeño escolar y el rendimiento obtenido, como en los trabajos de Candia Oviedo (2019), Carmona Jáquez (2022) y Castrillón et al. (2020).

De las publicaciones revisadas, dentro de los países iberoamericanos, como se puede ver en la Figura 2, México se destaca en la utilización del aprendizaje automático para la predicción del desempeño escolar, sobre todo en 2018 y 2022.

Figura 1. Número de publicaciones por tarea y año (elaboración propia)

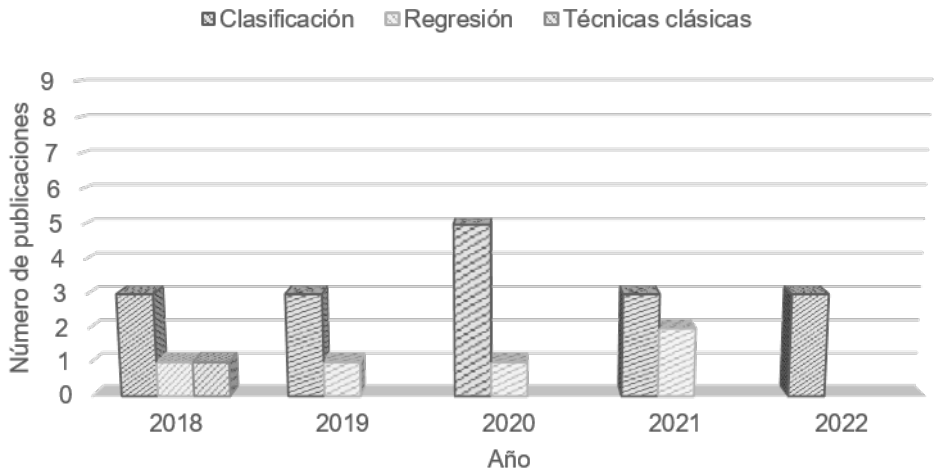
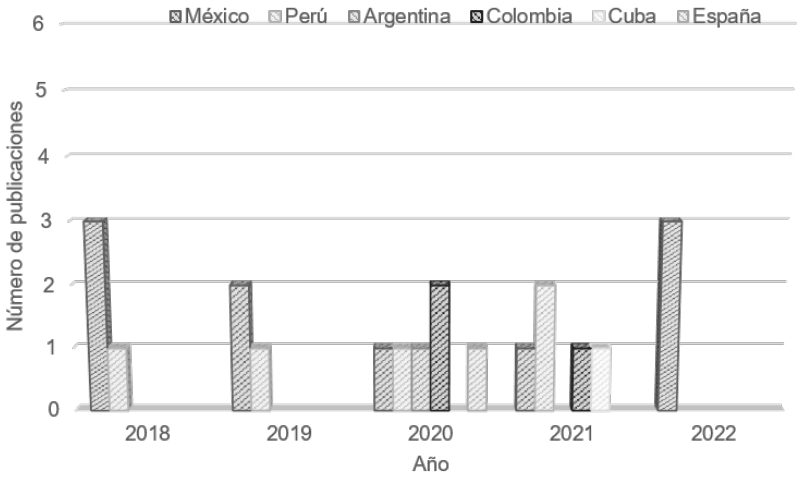


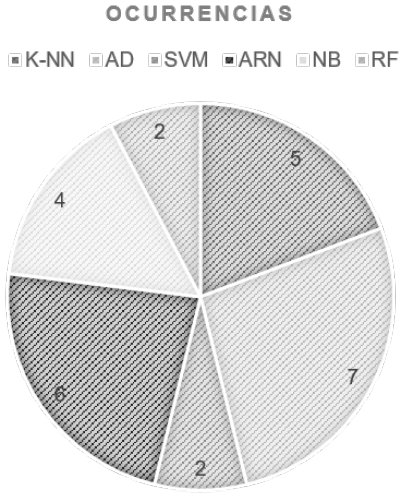
Figura 2. Número de publicaciones por país (elaboración propia)



En el transcurso de los cinco años estudiados, se utilizaron varios algoritmos de aprendizaje automático supervisado. Como se ve en la Figura 3, los algoritmos más utilizados son los Árboles de decisión (AD) con 7 ocurrencias, seguido de las redes neuronales con 6, K-vecinos más cercanos (K-NN) con 4; al igual que Naïve Bayes (NB) y Random Forest (RF) con 2; al igual que las máquinas de soporte vectorial (SVM). Algo a destacar es que los algoritmos de clasificación tienen más presencia que los algoritmos lineales y de regresión.

En el transcurso de los cinco años estudiados, se utilizaron varios algoritmos de aprendizaje automático supervisado, como se ve en la Figura 3. Los algoritmos más utilizados son los Árboles de decisión (AD) con 7 ocurrencias, seguido de las redes neuronales con 6, K-vecinos más cercanos (K-NN) con 4, al igual que Naïve Bayes (NB);

Figura 3. Cantidad de ocurrencias de algoritmos de aprendizaje en los trabajos revisados



y Random Forest (RF) con 2 al igual que las máquinas de soporte vectorial (SVM). Una de las cualidades más importantes de los árboles de clasificación es que son un modelo explicativo del fenómeno que se estudia (Lizares Castillo, 2017), lo cual implica que el investigador puede visualmente determinar relaciones entre las variables o, en este contexto, de los factores que afectan el desempeño escolar.

Como se puede ver en la Tabla 1, las publicaciones utilizan más los algoritmos AD teniendo, en promedio, una precisión del 72.8%, a pesar de que este tipo de algoritmos tienen una mayor facilidad para interpretar los datos; si se compara con las ARN, que son consideradas cajas negras, es posible notar que su porcentaje de precisión no es el más alto, con 78%.

En relación con el total de registros que se utilizan para el entrenamiento, como se puede ver en la Tabla 3, este es variado, pues depende del grado de confidencialidad, el tamaño de la población y el margen

Tabla 1. Porcentaje de precisión de los modelos de clasificación (elaboración propia)

Publicaciones	Precisión por modelos					
	AD	ARN	RF	NB	K-NN	SVM
		2022				
Carmona, 2022		98.27%				
Rico y Gaytán, 2022	46%			65%	32%	
Rico, 2022	68%			70%	68%	
		2021				
Rodríguez-Hernández et al., 2021	82%	80%	82%			
Incio et al., (2021)		86% y 70%				
Díaz-Landa et al., (2021)	62.45%					
		2020				
Contreras et al., 2020						66%
Urbina-Nájera et al., 2020	92%					
Urteaga et al., 2020		28%			31%	
Castrillón et al., 2020	92%					
Leonor y Cazarez, 2020		94%				89%
		2019				
Candia, 2019	67%		69%		63%	
Rico et al., (2019)				70.42 %		
		2018				
Cabana, 2018		90%				
Rico y Sánchez, (2018)				70%		
Promedio	72.8	78.0	75.5	68.9	48.5	77.5

de error de cada uno de los estudios. Se puede observar, en la misma tabla, cómo predominan los factores cognitivos y socioeconómicos; sin embargo, otros factores se han estudiado y algunos autores se han referido a la importancia del papel que juegan las universidades en este aspecto y cómo deben proveer de las herramientas necesarias para la mejora del rendimiento escolar (Guadalupe y González, 2017; Huecas, 2020; Urbina-Nájera et al., 2020).

Tabla 2. Total de registros y factores utilizados en cada uno de los trabajos revisados (elaboración propia)

Publicaciones	Registros	Factores
Rico, (2022)	260	Cognitivos.
Rico y Gaytán, (2022)	228	Socioeconómicos y cognitivos.
Carmona, (2022)	23,000	Socioeconómicos y cognitivos.
Rodríguez-Hernández et al., (2021)	162,030	Socioeconómicos y cognitivos.
Incio et al., (2021)	50	Cognitivos, emocionales y socioculturales.
Díaz-Landa et al., (2021)	237	Instruccionales, institucionales, relaciones interpersonales.
Lemus et al., (2021)	288	Cognitivos, demográficos, socioeconómicos, salud
Martínez et al., (2021)	87	Cognitivo, sociodemográfico, socioeconómico
Urteaga et al., (2020)	654	Cognitivos, institucionales
Urbina-Nájera et al., (2020)	500	Demográficos, socioeconómicos, institucionales e instruccionales.
Leonor y Cazarez, (2020)	1,181	Cognitivos.
Contreras et al., (2020)	1620	Sociodemográficos y cognitivos.
Castrillón et al., (2020)	121	Instruccionales es institucionales.
Layza, (2020)	66	Cognitivos, motivacionales, socioambientales, institucionales e instruccionales.
Candia, (2019)	12,698	Socioeconómicos, cognitivos, sociodemográficos e institucionales.
Rico et al., (2019)	71	Socioeconómicos, cognitivos, motivacional, relaciones interpersonales.
Martínez et al., (2019)	153	Cognitivo, socioeconómico, sociodemográficos, relaciones interpersonales.
Cabana, (2018)	69	Cognitivos.
Rico y Sánchez, (2018)	86	Cognitivos.
Martínez-González et al., (2018)	27,624	Cognitivos y sociodemográficos, institucionales.
González-Beltrán et al., (2018)	5,121	Cognitivos.

Finalmente, Cruz et al. (2022) señala que, aunque las soluciones actuales no resuelven todas las necesidades de los estudiantes y las instituciones, los resultados obtenidos son muy prometedores. Sin embargo, el autor argumenta que, a pesar del avance en la tecnología de predicción de la deserción escolar y la mejora del rendimiento de los estudiantes, todavía queda trabajo por hacer.

VI. CONCLUSIONES

Luego de revisar y analizar diversas publicaciones sobre la predicción del rendimiento académico, incluyendo los algoritmos utilizados, el tamaño de las muestras y los factores relacionados con el rendimiento de los estudiantes, se concluye que la predicción del rendimiento escolar es un campo de investigación que busca identificar factores que puedan predecir el éxito académico de los estudiantes. Se han utilizado diferentes enfoques para predecir el rendimiento escolar, incluyendo la medición de habilidades cognitivas, aspectos sociodemográficos, la evaluación del entorno familiar y socioeconómico, y la observación de la motivación y el compromiso del estudiante. A pesar de que existen muchos factores que pueden influir en el rendimiento académico, la predicción del rendimiento escolar sigue siendo un desafío para los investigadores y educadores. Sin embargo, esta área de investigación es importante para ayudar a los estudiantes a alcanzar su máximo potencial académico y para mejorar la calidad de la educación en general.

Como se mencionó anteriormente los modelos predictivos tienen un campo de aplicación bastante amplio. Lo que facilita la aplicación de un modelo predictivo es adaptación para manejar distintas cantidades de registros.

En futuras investigaciones, se puede tomar como base alguno de los modelos y sus factores para aplicarlo de manera específica en una institución de educación superior, con el objetivo de obtener un diagnóstico que permita intervenir en los factores que puedan afectar a los estudiantes, con el objetivo de fomentar la eficiencia terminal y un mejor desempeño académico. Esto puede ayudar a las instituciones a desarrollar políticas y estrategias específicas que aborden las necesidades y desafíos únicos de sus estudiantes, lo que a su vez puede mejorar su retención y éxito académico.

REFERENCIAS

Betancourt, G. A., (2005). Las máquinas de soporte vectorial SVMs. *Scientia Et Technica*, *11*(27), 67–72.

Cabana Yupanqui, S. B. (2018). *Análisis predictivo del rendimiento académico en los alumnos de la escuela profesional de ingeniería en informática y sistemas de la unjbg, utilizando redes neuronales, semestre 2017-i*. http://repositorio.unjbg.edu.pe/handle/UNJBG/3200

Candia Oviedo, D. I. (2019). *Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático*. https://repositorio.unsaac.edu.pe/handle/20.500.12918/4120

Carmona Jáquez, J. R. (2022). *Predicción de desempeño académico por medio de redes neuronales* [Tesis, Universidad Autónoma de Chihuahua]. http://repositorio.uach.mx/id/eprint/503

Castrillón, O. D., Sarache, W., y Ruiz-Herrera, S. (2020). Prediction of academic performance using artificial intelligence techniques. *Formación Universitaria*, *13*(1), 93–102. https://doi.org/10.4067/S0718-50062020000100093

Contreras, L. E., Fuentes, H. J., y Rodriguez, J. I. (2020). Academic performance prediction by machine learning as a success/failure indicator for engineering students. *Formacion Universitaria*, *13*(5), 233–246. https://doi.org/10.4067/S0718-50062020000500233

Cruz, E., González, M., y Rangel, J. C. (2022). Técnicas de *machine learning* aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión. *Prisma Tecnológico*, *13*(1), 77–87. https://doi.org/10.33412/pri.v13.1.3039

Díaz-Landa, B., Meleán-Romero, R., y Marín-Rodriguez, W. (2021). Rendimiento académico de estudiantes en Educación Superior: predicciones de factores influyentes a partir de árboles de decisión. *Telos Revista de Estudios Interdisciplinarios en Ciencias Sociales*, *23*(3), 616–639. https://doi.org/10.36390/telos233.08

Guadalupe, E., y González, C. (2017). Factores que inciden en el rendimiento académico de los estudiantes de la Universidad Politécnica del Valle de Toluca. *Latinoamericana de Estudios Educativos*, *47*, 91–108. https://www.redalyc.org/articulo.oa?id=27050422005

Huecas Dueñas, J. (2020). *Análisis de la relación entre la situación socioeconómica de los padres y rendimiento académico de los hijos en España*. https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/37120/TFG%20-%20Duenas%20Huecas%2C%20Jesus.pdf?sequence=1

Ibañez Reyes, S., Lozano Jimenez, J., y Lastre Gómez, D. (2020). *Diseño de un modelo predictivo entre las características de personalidad y el desempeño académico de los estudiantes del Programa de Psicología la Facultad de Ciencias Humanas y Sociales de la Universidad de la Costa*.

Corporación Universidad de la Costa. https://doi.org/https://hdl.handle.net/11323/7085

Incio Flores, F. A., Capuñay Sanchez, D. L., Estela Urbina, R. O., Delgado Soto, J. A., y Vergara Medrano, S. E. (2021). Diseño e implementación de una red neuronal artificial para predecir el rendimiento académico en estudiantes de Ingeniería Civil de la UNIFSLB. *Revista Veritas et Scientia-UPT*, *10*(1), 107–117. https://doi.org/10.47796/ves.v10i1.464

Layza Bermudez, F. H. (2020). *Modelo matemático de predicción del rendimiento académico del curso matemática II en la facultad de ingeniería química, unac 2019*. http://hdl.handle.net/20.500.12952/5484

Lemus Amezcua, A. A., Cantón Croda, R. M., y Morales Salgado, M. del R. (2021). Construcción de un modelo predictivo para determinar el rendimiento académico de los estudiantes del colegio de estudios científicos y tecnológicos del estado de Michoacán. *Ciencia Latina Revista Científica Multidisciplinar*, *5*(5), 7709–7749. https://doi.org/10.37811/cl_rcm.v5i5.872

Leonor, R., y Cazarez, U. (2020). Aplicación de una red neuronal probabilística para predecir el desempeño académico de estudiantes de educación superior en línea. In *Research in Computing Science*, *149*(8).

Lizares Castillo, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico* [Tesina, Universidad del Perú]. https://hdl.handle.net/20.500.12672/7122

Mahesh, B. (2020). Machine Learning Algorithms-A Review. *International Journal of Science and Research*, *9*(1), 381–386. https://doi.org/10.21275/ART20203995

Martínez Pérez, J. R., Pérez Leyva, E. H., Ferrás Fernández, Y., y Bermúdez Cordoví, B. C. (2021). Análisis predictivo de la deserción estudiantil en la carrera de Medicina. *EDUMECENTRO*, *13*(3), 217–236. https://orcid.org/0000-0002-0415-6200

Martínez R, Álvarez-Xochihua O, Mejía O, Jordan A, y Gonzáles-Fraga J. (2019). *Use of Machine Learning to Measure the Influence of Behavioral and Personality Factors on Academic Performance of Higher Education Students*. https://doi.org/10.1109/tla.2019.8891928

Martínez-González, A., Patricia Manzano-Patiño, A., García-Minjares, M., Johana Herrera-Penilla, C., Ricardo Buzo-Casanova, E., y Sánchez-Mendiola, M. (2018). *Grado de conocimientos de los estudiantes al ingreso a la licenciatura y su asociación con el desempeño escolar y la eficiencia terminal. Modelo multivariado.* https://www.scielo.org.mx/scielo.php?pid=S0185-27602018000400057&script=sci_abstract

Meléndez-Armenta, R. A. (2022). La salud mental y su influencia en el desempeño académico de estudiantes durante la pandemia COVID-19. *Revista Electrónica Educare*, *27*(1), 1–12. https://doi.org/10.15359/ree.27-1.14538

Morales Hernández, M. Á., González Camacho, J. M., Robles Vásquez, H., Del Valle Paniagua, D. H., y Durán Moreno, J. R. (2022). Algoritmos de aprendizaje automático para la predicción del logro académico. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *12*(24). https://doi.org/10.23913/ride.v12i24.1180

Muñoz Comonfort, A., y Fortoul van der Goes, T. I. (2022). Preparación académica previa y desempeño académico de estudiantes de primer año de una escuela de medicina. *Investigación En Educación Médica*, *11*(43), 90–98. https://doi.org/10.22201/fm.20075057e.2022.43.21423

Pinzón Pérez, D. F., Román González, M., y González Palacio, E. V. (2023). El pensamiento algorítmico como estrategia didáctica para el desarrollo de habilidades de resolución de problemas en el contexto de la educación básica secundaria. *Revista de Educación a Distancia (RED)*, *23*(73), 2–22. https://doi.org/10.6018/red.542111

Raschka, S. (2018). STAT 479: *Machine Learning Lecture Notes*. http://stat.wisc.edu/sraschka/teaching/stat479-fs2018/

Ráyon, A. (2017, abril). Guía para comenzar con algoritmos de Machine Learning. *Deusto Data*. https://blogs.deusto.es/bigdata/guia-para-comenzar-con-algoritmos-de-machine-learning/

Rico Páez, A. (2022). Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *12*(24). https://doi.org/10.23913/ride.v12i24.1196

Rico Páez, A., y Gaytán Ramírez, N. D. (2022). Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería. *IE Revista de Investigación Educativa de La REDIECH*, *13*, e1426. https://doi.org/10.33010/ie_rie_rediech.v13i0.1426

Rico Páez, A., Gaytán Ramírez, N. D., y Sánchez Guzmán, D. (2019). Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes. *Diálogos Sobre Educación*, *19*. https://doi.org/10.32870/dse.v0i19.509

Rico Páez, A., y Sánchez Guzmán, D. (2018). Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN / Design of a model to automate the prediction of academic performance in students of IPN. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *8*(16), 246–266. https://doi.org/10.23913/ride.v8i16.340

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., y Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100018

Samuel, A. L. (1959). *Some Studies in Machine Learning Using the Game of Checkers.*

Sandoval Serrano, L. J. (2018). *Algoritmos de aprendizaje automático para análisis y predicción de datos. 11.* http://www.redicces.org.sv/jspui/handle/10972/3626

Soria Olivas, E., Sánchez-Montañés, M. A., Gamero Cruz, I. R., Borja Castillo, C., y Cano Michelena, P., (2023). *Sistemas de aprendizaje automático. Ra-Ma* .

Urbina-Nájera, A. B., Camino-Hampshire, J. C., y Cruz Barbosa, R. (2020). University dropout: Prevention patterns through the application of educational data mining. *RELIEVE - Revista Electronica de Investigacion y Evaluacion Educativa*, *26*(1), 1–19. https://doi.org/10.7203/relieve.26.1.16061

Urteaga, I., Siri, L., y Garófalo, G. (2020). Predicción temprana de deserción mediante aprendizaje automático en cursos profesionales en línea. *RIED. Revista Iberoamericana de Educación a Distancia*, *23*(2), 147. https://doi.org/10.5944/ried.23.2.26356

Webb, G. I. (2016). *Naïve Bayes. In Encyclopedia of Machine Learning and Data Mining*. Springer US. https://doi.org/10.1007/978-1-4899-7502-7_581-1

Witten H. Ian, Eibe Frank, y Mark A. Hall. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. https://doi.org/https://doi.org/10.1016/C2009-0-19715-5