Análisis comparativo de algoritmos de aprendizaje automático supervisado para construir un modelo predictivo para evaluar el rendimiento de los estudiantes

Artículo *en* International Journal of Online and Biomedical Engineering (iJOE) - Febrero 2021

DOI: 10.3991/ijoe.v17i02.20025

CITACIONES
4

LECTURAS
572

4 autores:

Inssaf El Guabassi
Universidad Abdelmalek Essaâdi
15 PUBLICACIONES **84** CITAS
VER PERFIL

Zakaria Bousalem
Facultad de Ciencias y Tecnología, Universidad Hassan I
11 PUBLICACIONES **64** CITAS
VER PERFIL

Marah Rim
ENSI Tánger
15 PUBLICACIONES **45** CITAS
VER PERFIL

Aimad Qazdar
Universidad Cadi Ayyad
18 PUBLICACIONES **62** CITAS
VER PERFIL

Análisis comparativo de algoritmos de aprendizaje automático supervisado para construir un modelo predictivo para evaluar el rendimiento de los estudiantes

https://doi.org/10.3991/ijoe.v17i02.20025

Inssaf El Guabassi (✉)
Universidad Abdelmalek Essaadi, Tetuán, Marruecos
elguabassi@gmail.com

Zakaria Bousalem
Universidad Hassan 1, Settat, Marruecos

Rim Marah
Universidad Abdelmalek Essaadi, Tetuán, Marruecos

Aimad Qazdar
Universidad Cadi Ayyad, Marrakech, Marruecos

Resumen-En los últimos años, la población mundial exige cada vez más predecir el futuro con certeza, predecir la información correcta en cualquier ámbito se está convirtiendo en una necesidad. Una de las formas de predecir el futuro con certeza es determinar el futuro posible. En este sentido, el aprendizaje automático es una forma de analizar enormes conjuntos de datos para hacer predicciones o decisiones sólidas. El objetivo principal de este trabajo de investigación es construir un modelo predictivo para evaluar el rendimiento de los estudiantes. Por lo tanto, las contribuciones son tres. La primera es aplicar varios algoritmos de aprendizaje automático supervisado (es decir, ANCOVA, Regresión Logística, Regresión de Vecinos de Apoyo, Regresión Log-lineal, Regresión de Árbol de Decisión, Regresión de Formas Aleatorias y Regresión de Mínimos Cuadrados Parciales) en nuestro conjunto de datos de educación. El segundo objetivo es comparar y evaluar los algoritmos utilizados para crear un modelo predictivo basado en varias métricas de evaluación. El último propósito es determinar los factores más importantes que influyen en el éxito o el fracaso de los alumnos. Los resultados experimentales mostraron que la Regresión Log-lineal proporciona una mejor predicción, así como los factores de comportamiento que influyen en el rendimiento de los estudiantes.

Palabras clave-Rendimiento de **los estudiantes**, predicción, aprendizaje automático, regresión, modelos predictivos, minería de datos educativos

1 Introducción

En el mundo real, con un notable crecimiento dentro del universo de tamaños de almacenes de datos medidos, el análisis de los datos y la extracción de la información útil se está convirtiendo en

una necesidad y un tema rico para varios investigadores [1]. Muchas áreas de aplicación adoptan técnicas de aprendizaje automático en sus sistemas, como las finanzas, las plataformas de compra, los restaurantes, la economía, la medicina, los objetivos turísticos y el marketing. En las dos últimas décadas, el aprendizaje automático ha entrado también en el espacio del aprendizaje electrónico [2] [3] [4] [5]. Así, varios algoritmos de aprendizaje automático han sido explotados por los investigadores para predecir patrones ocultos de los entornos educativos [6] [7] [8].

La predicción de los alumnos con riesgo de fracaso escolar es de suma importancia y debe identificarse lo antes posible durante el curso académico. La previsión temprana del rendimiento de los estudiantes es necesaria en la educación superior para proporcionar una educación de alta calidad, reducir las tasas de abandono, aumentar las tasas de finalización de los estudios y mejorar los resultados educativos.

Sin embargo, los verdaderos y mayores problemas son:

- ¿Cómo identificar a los estudiantes "débiles" que necesitarán ayuda adicional para mejorar su rendimiento?
- ¿Cuál es el mejor algoritmo de aprendizaje automático (es decir, modelo) para predecir el rendimiento ac- démico de los estudiantes?
- ¿Qué factores pueden afectar al rendimiento académico de los estudiantes?

Este trabajo de investigación evalúa y compara la eficacia de diferentes algoritmos de aprendizaje automático. Aunque hay muchos algoritmos para crear modelos predictivos, este trabajo se concentra en siete de ellos, que son ANCOVA, Regresión Logística, Regresión de Vectores de Apoyo, Regresión Log-lineal, Regresión de Árboles de Decisión, Regresión de Bosques Aleatorios y Regresión de Mínimos Cuadrados Parciales. El presente trabajo también determina los factores que afectan al rendimiento académico de los estudiantes.

El esquema del presente trabajo es el siguiente: En la sección 2 se presentan los estudios recientes relativos al área especificada. En la sección 3 se describen brevemente los antecedentes del aprendizaje automático. La sección 4 se centra en el enfoque propuesto. En la sección 5 se presenta una descripción de los materiales y de los métodos. En la sección 6 se presenta nuestra implementación y los resultados. La sección 7 se centra en la evaluación experimental. La sección 8 contiene la discusión. Por último, la sección 9 presenta las principales conclusiones, considerando algunas direcciones de investigación futuras.

2 Trabajos relacionados

En las últimas décadas, muchos estudios realizados por varios equipos de investigación se han centrado en predecir el rendimiento de los estudiantes en función de los factores de los buceadores utilizando diversos algoritmos de aprendizaje automático.

Bravo-Agapito et al [13] explicaron su estudio basándose en la predicción del rendimiento académico de 802 estudiantes de pregrado en el aprendizaje completamente online. Utilizaron el análisis factorial ex- ploratorio, las regresiones lineales múltiples y el análisis de conglomerados. Concluyeron que la "edad" es un factor que afecta al rendimiento académico del estudiante. Gray y Perkins [14] realizaron un estudio sobre la predicción de los resultados de los estudiantes a partir de la cuarta semana del semestre de otoño utilizando técnicas de aprendizaje automático. Hamsa et al [15] aplicaron dos métodos de clasificación que son el árbol

de decisión y el algoritmo genético difuso para predecir

el rendimiento de los estudiantes de grado y máster en Informática y Electrónica y Comunicación. Hussain et al [16] describieron un estudio de rendimiento sobre la predicción de las dificultades de los estudiantes a partir de los datos de las sesiones de aprendizaje. Utilizaron redes neuronales artificiales, máquinas de vectores de apoyo, regresión logística, clasificadores Naïve Bayes y árboles de decisión. Sus resultados muestran que las redes neuronales artificiales y las máquinas de vectores de apoyo son los mejores algoritmos para predecir el rendimiento del estudiante. Karthikeyan et al [17] investigaron el rendimiento de los estudiantes desarrollando un modelo híbrido de minería de datos educativos denominado HEDM. Su modelo combina dos técnicas que son el clasificador J48 y la clasificación de Naive Baye. Sus resultados muestran que HEDM supera los resultados obtenidos en EDM.

En resumen, muchos investigadores han obtenido en sus últimos trabajos resultados significativos en la minería de datos educativos. Sin embargo, la mayoría de ellos utilizan métodos de clasificación para predecir el rendimiento académico de los estudiantes. Además, se ha prestado muy poca atención a las características de interacción y participación de los padres.

3 Aprendizaje automático

El aprendizaje automático reproduce el comportamiento mediante algoritmos de aprendizaje que se alimentan a su vez de inmensas fuentes de información. El ordenador se entrena y mejora, de ahí la palabra aprendizaje; "aprende" de los datos y extrae conocimientos de ellos.

Los algoritmos son los motores del aprendizaje automático. En general, se utilizan tres tipos principales de algoritmos de aprendizaje automático: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

- Aprendizaje supervisado: El sistema aprende una función a partir de ejemplos.
- Aprendizaje no supervisado: El sistema no se basa en elementos predefinidos.
- Aprendizaje por refuerzo: consiste en dejar que el algoritmo aprenda de sus propios errores. Ante una elección aleatoria al principio, utiliza recompensas y castigos como señales para una mala y una buena decisión.

Tras describir brevemente los antecedentes del aprendizaje automático, en la siguiente sección presentaremos nuestro enfoque propuesto.

4 Enfoque propuesto

Cada vez más, el e-learning se ha convertido en una importante herramienta de enseñanza y aprendizaje en todo el mundo. Además, los alumnos tienen la oportunidad de pasar al aprendizaje a distancia en diversos campos científicos en cualquier momento y en cualquier lugar [9]. Por lo tanto, es evidente que muchos investigadores trabajan en los diversos aspectos del e-learning [10] [11] [12]. La identificación de los estudiantes "débiles" y de los factores que afectan al rendimiento académico de los estudiantes es un paso crucial para el éxito del aprendizaje. Por lo tanto, en el presente trabajo, se pretende evaluar el rendimiento académico de los estudiantes e identificar los factores que influyen en el rendimiento académico utilizando algoritmos de aprendizaje automático supervisado.

Este trabajo de investigación se centra en los siguientes pasos:

- Aplicar varios algoritmos de aprendizaje automático, como ANCOVA, regresión logística, regresión de vectores de apoyo, regresión logarítmica y lineal, regresión de árboles de decisión, regresión de bosques aleatorios y regresión de mínimos cuadrados parciales.
- Comparación y evaluación de algoritmos de aprendizaje automático para identificar los más adecuados mediante el uso de varias métricas de evaluación que son el error cuadrático medio (MSE), el error cuadrático medio (RMSE) y la R cuadrada (R^2).
- Identificar qué factores influyen en la predicción final de los resultados de los alumnos.

La siguiente sección describe los materiales y métodos utilizados en nuestro trabajo de investigación, que son el conjunto de datos, los métodos aplicados y los métodos de evaluación.

5 Materiales y métodos

5.1 Conjunto de datos

Los datos utilizados para la experimentación de este trabajo (disponibles aquí) proceden de un conjunto de datos denominado "Students' Academic Performance Dataset (xAPI-Edu-Data)" [18] [19]. Se trata, por tanto, de un conjunto de datos de código abierto disponible públicamente en el repositorio de conjuntos de datos de Kaggle para fines académicos y de investigación. La fuente principal del conjunto de datos es Elaf Abu Amrieh, Thair Hamtini e Ibrahim Aljarah, Universidad de Jordania, Am- man, Jordania, http://www.Ibrahimaljarah.com, www.ju.edu.jo. Estos datos se obtienen del sistema de gestión del aprendizaje conocido como Kalboard 360 [20]. Kalboard 360 ha sido creado para ayudar a las escuelas a mejorar su aprendizaje mediante el uso de tecnología punta. Normalmente, un sistema de este tipo comparte y proporciona a los usuarios acceso sincrónico a los recursos educativos desde cualquier dispositivo que tenga acceso a Internet. La tabla 1 ofrece un resumen de las características del conjunto de datos, incluido el nombre, la abreviatura, la fuente, las características, el número de muestras, el área, las características de los atributos, el número de atributos, la fecha, las tareas asociadas, el valor que falta y los formatos de archivo.

Tabla 1. Resumen del conjunto de datos

Nombre	Conjunto de datos sobre el rendimiento académico de los estudiantes
Abreviatura	xAPI-Edu-Data
Fuente	Elaf Abu Amrieh, Thair Hamtini e Ibrahim Aljarah, Universidad de Jordania, Ammán, Jordania.
Características	Multivariante
Número de muestras	480
Área	E-learning, Educación, Modelos predictivos, Minería de datos educativos
Características de los atributos	Entero/Categórico
Número de atributos	16
Fecha	2016-11-8
Tareas asociadas	Clasificación
¿Valores perdidos?	No

Formatos de archivo	xAPI-Edu-Data.csv

Como se muestra en la Tabla 1, el conjunto de datos considerado consta de 480 registros de estudiantes de varios países y 17 características. Por otra parte, las características se clasifican en tres categorías principales, denominadas "Características demográficas", "Características de los antecedentes académicos" y "Características del comportamiento":

- **Características demográficas**: Incluye cualidades como el género, la nacionalidad y el lugar de nacimiento.
- **Las características de los antecedentes académicos**: Representa las características de los antecedentes de los estudiantes como la etapa educativa, el nivel de grado, la sección y el semestre.
- **Características del comportamiento**: Ilustran el comportamiento como una clase de mano levantada, la apertura de recursos, la respuesta a las encuestas por parte de los padres y la satisfacción de la escuela.

La tabla 2 contiene un resumen de las características del conjunto de datos utilizado para el entrenamiento y las pruebas. Contiene tres campos: característica, descripción y tipo. Cabe señalar que hay dos tipos principales de características, denominadas "Nominal" y "Numérica".

- **Nominal**: Etiqueta las variables proporcionando un valor no numérico.
  - Ejemplos: Sexo {Hombre o Mujer}, nivel {bajo, medio, alto}, color de ojos {azul, verde, marrón, avellana, ámbar, rojo y gris}
- **Numérico**: Etiqueta las variables proporcionando un valor cuantitativo.
  - Ejemplos: Rangos, tamaño, humedad, temperatura y tiempo.

Tabla 2. Características del conjunto de datos

Característica	Descripción	Tipo
Género	Sexo del estudiante (es decir, hombre o mujer)	Nominal
Nacionalidad	Nacionalidad del estudiante (por ejemplo, Marruecos, Kuwait, Líbano, Jordania, Egipto, Arabia Saudí, Estados Unidos, etc.)	Nominal
Lugar de nacimiento	País de nacimiento del estudiante (por ejemplo, Marruecos, Kuwait, Líbano, Jordania, Egipto, Arabia Saudí, Estados Unidos, etc.)	Nominal
Etapas educativas	El nivel educativo del alumno (es decir, nivel inferior, nivel medio o nivel superior)	Nominal
Niveles de grado	Nivel de grado del estudiante (es decir, G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11 o G-12)	Nominal
Sección ID	Aula del alumno (es decir, A, B o C)	Nominal
Tema	Tema del curso (es decir, inglés, español, francés, árabe, informática, matemáticas, química, biología, ciencias, historia, corán o geología)	Nominal
Semestre	Semestre del año (es decir, primero o segundo)	Nominal
Padre responsable	El progenitor responsable del alumno (es decir, la madre o el padre)	Nominal
Mano alzada	Número de veces que el alumno ha levantado la mano en el aula (es decir, de 0 a 100)	Numérico
Recursos visitados	Número de veces que el estudiante visitó el contenido de un curso (es decir, de 0 a 100)	Numérico
Ver anuncios	Número de veces que el alumno comprobó los nuevos anuncios (es decir, de 0 a 100)	Numérico
Grupos de discusión	Número de veces que el estudiante participó en grupos de discusión (de 0 a 100)	Numérico

Respuesta de los padres	Los padres han respondido o no a las encuestas que se facilitan desde la escuela (es decir, sí o no)	Nominal
Satisfacción de los padres con la escuela	El grado de satisfacción de los padres con respecto a la escuela (es decir, Sí o No)	Nominal
Días de ausencia de los estudiantes	El número de días de ausencia de cada estudiante (es decir, por encima de 7 o por debajo de 7)	Nominal
Clase	Grado del estudiante para el curso (es decir, nivel bajo, nivel medio o nivel alto)	Nominal

Después de ver el conjunto de datos utilizado en nuestra experimentación, en la siguiente sección pre sentamos los métodos seleccionados para predecir el rendimiento académico de los estudiantes.

5.2 Métodos seleccionados

Es imposible predecir el futuro con certeza, pero puede determinar un resultado altamente exitoso observando las fuentes de datos existentes. En la actualidad, existen muchos al- goritmos para el aprendizaje automático de modelos predictivos. En el presente trabajo, nos centramos especialmente en los algoritmos de aprendizaje automático supervisado porque son los más adecuados (véase la sección III para más detalles).

En las siguientes secciones, presentaremos los algoritmos utilizados para construir modelos predictivos, que son ANCOVA, regresión de vectores de apoyo, regresión de árboles de decisión, regresión de bosques aleatorios, regresión de mínimos cuadrados parciales, regresión logarítmica-lineal y regresión logarítmica.

ANCOVA (ANalysis of VAriance) [21] es una prueba estadística que permite comparar globalmente la expectativa matemática de varias muestras. El nombre de esta prueba se explica por su forma de proceder: descomponemos la varianza total de las muestras en dos varianzas parciales, la varianza entre clases y la varianza residual, y comparamos estas dos varianzas. El modelo ANCOVA se escribe como sigue (1):

$$yij = \mu + \tau j + \beta(xij - \bar{x}) + \varepsilon ij \quad (1)$$

Dónde:

y_ij: es la jª observación del i-ésimo grupo. μ: es una constante común a todos los individuos. τ_j: es el efecto del tratamiento del j-ésimo grupo.

β: es la pendiente de la regresión correspondiente a la covariable xij. x_ij: es la covariable del i-ésimo sujeto del j-ésimo grupo.

x es la media global de x.

ε_ij: es un término de error gaussiano.

Como se muestra en la figura 2, el ANCOVA ayuda a comparar dos o más líneas de regresión entre sí.

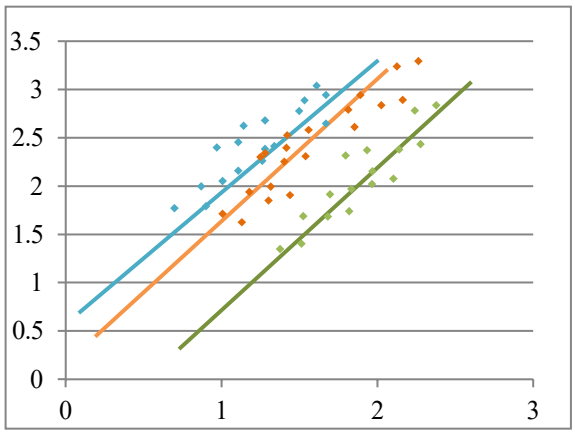

Fig. 1. ANCOVA

La Regresión Logística o Regresión Logit (Logit-R) [22] es un método estadístico para realizar clasificaciones binarias como sano/enfermo, ganar/perder, pasar/fallar o vivo/muerto.

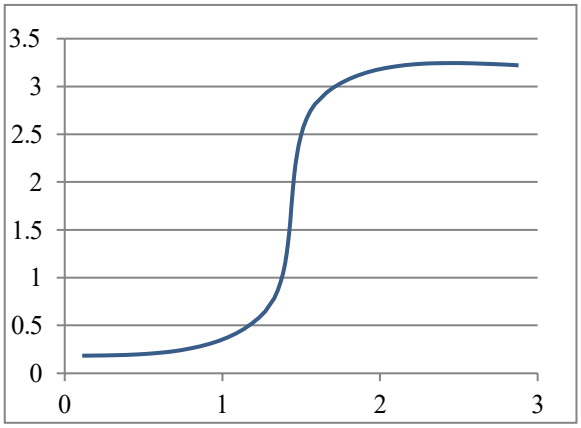

Fig. 2. Regresión logística

Toma como entrada variables predictoras cualitativas y/o ordinales y mide la probabilidad del valor de salida utilizando la función sigmoidea mostrada en la figura 2 y definida por la fórmula (1):

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

La regresión de vectores de apoyo (SVR) [23] es un algoritmo de clasificación binaria. Al igual que la Regresión Logística. Si tomamos la imagen de arriba, tenemos dos clases (por ejemplo, supongamos que se trata de correos electrónicos, y que los correos spam están en rojo y los no spam en azul). La regresión logística puede separar estas dos clases definiendo la línea en rojo. El SVR optará por separar las dos clases mediante la línea verde (véase la figura 3).

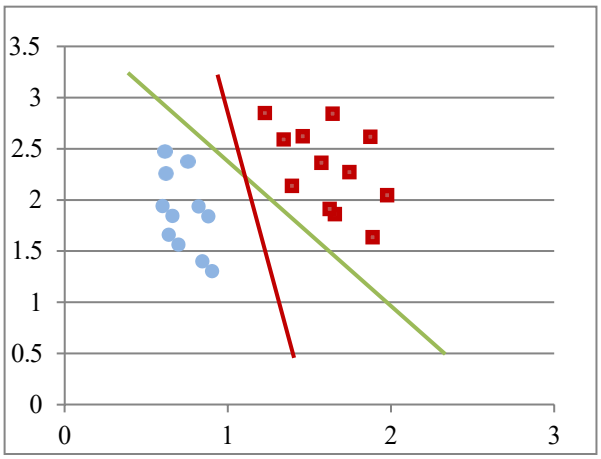

Fig. 3. Regresión de vectores de apoyo

La regresión de árboles de decisión (DTR) [25] es un algoritmo que utiliza un modelo de grafos (árboles) para definir la decisión final. Cada nodo tiene una condición, y las ramas se basan en esta condición (Verdadero o Falso). Cuanto más se desciende en el árbol, más condiciones se acumulan. La figura 4 ilustra este funcionamiento.

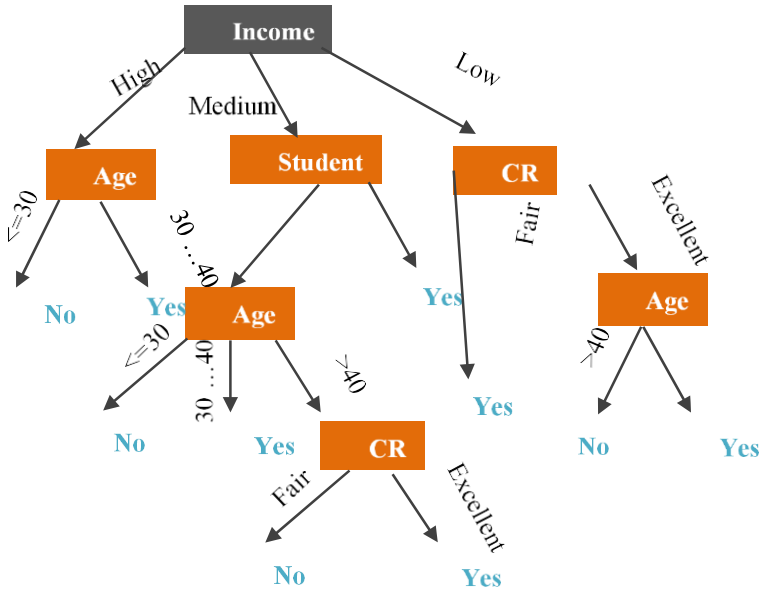

Fig. 4. Árbol de decisión

La Regresión Log-Lineal (Log-LR) [24] forma parte de la familia de modelos lineales generalizados para datos con distribución exponencial, gamma o de Poisson. Este método es un modelo lineal

para modelar la relación entre una variable de respuesta y una o más variables explicativas. Suponemos que la variable de respuesta se escribe como el logaritmo de una función afín de las variables explicativas

La Regresión de Bosque Aleatorio (RFR) [26] es un algoritmo de aprendizaje supervisado que combina múltiples predicciones para hacer una predicción más precisa que un solo modelo (ver Figura 5)

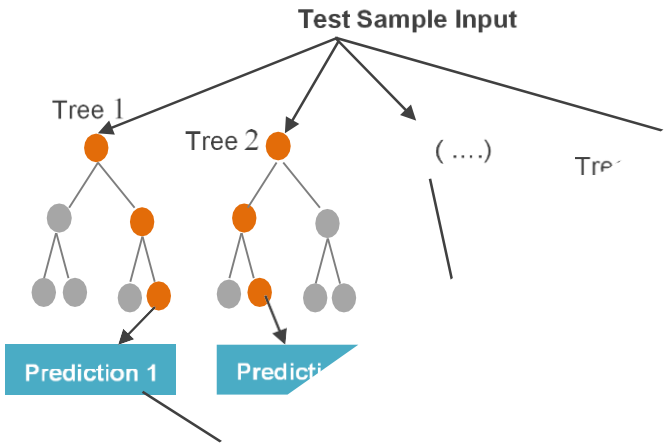

Fig. 5. Bosque aleatorio

La regresión por mínimos cuadrados parciales (PLS-R) [27] es una técnica estadística flexible aplicable a cualquier tipo de datos. Permite modelar las relaciones entre las entradas y las salidas, incluso cuando las entradas están correlacionadas y son ruidosas, las salidas son múltiples y las entradas son más numerosas que las observaciones. En la siguiente sección, nos centraremos en las métricas de evaluación utilizadas en nuestro estudio experimental para identificar el mejor algoritmo de aprendizaje de máquinas.

5.3 Métodos de evaluación

La evaluación de un modelo es una parte fundamental de la construcción de un modelo de aprendizaje automático eficaz. Hay muchos métodos de evaluación que pueden utilizarse. Sin embargo, la pregunta es: ¿qué métrica debemos utilizar para evaluar las técnicas de regresión en el aprendizaje automático? La figura 6 se representa para responder a esta pregunta.

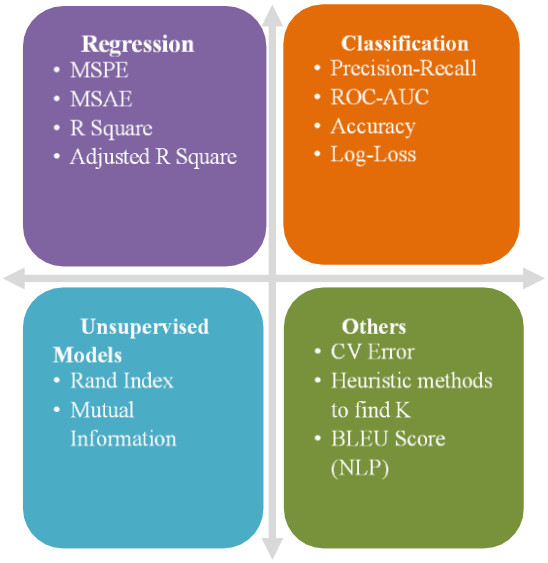

Fig. 6. Métrica correcta para evaluar los modelos de aprendizaje automático [28]

A continuación, hablaremos de las tres principales métricas que utilizaremos en nuestra evaluación.

El R-cuadrado (R2 o el coeficiente de determinación) [29] es un indicador que permite juzgar la calidad de la regresión lineal simple. Mide el ajuste entre el modelo y los datos observados o lo bien que la ecuación de regresión describe la distribución de los puntos.

- Si el R^2 es cero, significa que la ecuación de la recta de regresión determina el 0% de la distribución de los puntos. Esto significa que el modelo matemático utilizado no explica la distribución de los puntos.
- Si la R^2 es 1, significa que la ecuación de la recta de regresión puede determinar el 100% de la distribución de los puntos. Esto significa entonces que el modelo matemático utilizado, así como los parámetros a y b calculados, son los que determinan la distribución de los puntos.

En resumen, cuanto más se acerque el coeficiente de determinación a 0, más se dispersa el gráfico de dispersión alrededor de la línea de regresión. Por el contrario, cuanto más tiende el R^2 a 1, más se estrecha la nube de puntos alrededor de la línea de regresión. Cuando los puntos están exactamente alineados en la línea de regresión, entonces $R^2 = 1$.

El error cuadrático medio (MSE) [30] es la media aritmética de los cuadrados de las pre dicciones entre el modelo y las observaciones. Es el valor que debe minimizarse en el contexto de una regresión única o múltiple. El método se basa en la nulidad de la media de los residuos. Pero la media de sus cuadrados no suele ser nula.

El error cuadrático medio (RMSE) es una forma estándar de medir el error en los estudios de evaluación de modelos. Es la raíz cuadrada de la media del cuadrado de todos los errores.

6 Aplicación y resultados

El presente trabajo representa una comparación y evaluación de algoritmos de aprendizaje automático supervisado para predecir el rendimiento académico de los estudiantes. Se realizaron muchos experimentos en siete pasos principales dependiendo de los métodos de regresión, a saber, ANCOVA, Regresión Logística (Logit-R), Regresión de Vectores de Apoyo (SVR), Regresión Log-lin (Log-LR), Regresión de Árboles de Decisión (DTR), Regresión de Bosques Aleatorios (RFR) y Regresión de Mínimos Cuadrados Parciales (PLS-R). Estos métodos de regresión se aplicaron utilizando el entorno XLSTAT [31]. A continuación se presenta el resultado experimental de cada algoritmo.

Tabla 3. Resultados experimentales

	MSE	RMSE	R^2
ANCOVA	0.157256464	0.396555752	0.71890384
Logit-R	0.156250000	0.395284708	0.73799242
SVR	0.212447120	0.460919863	0.6271547
Log-LR	0.158611894	0.398261088	0.71667276
DTR	0.195293449	0.441920184	0.65025193
RFR	0.171994444	0.414722128	0.69480482
PLS-R	0.205659323	0.453496773	0.63238366

La tabla anterior representa, por tanto, los resultados resumidos de los siete algoritmos utilizados en este trabajo de investigación. Las métricas de evaluación utilizadas en este experimento son el error cuadrático medio (MSE), el error cuadrático medio (RMSE) y el R-cuadrado (R^2). Cabe señalar que el RMSE es simplemente la raíz cuadrada del MSE.

7 Evaluación

Después de evaluar rigurosamente los siete algoritmos en los 480 estudiantes de nuestro conjunto de datos, comparamos los resultados para determinar qué modelo predice mejor. De acuerdo con los resultados experimentales, está claro que la regresión logarítmica lineal (Log-LR) proporciona un mejor rendimiento porque tiene un MSE bajo, un RMSE bajo y una puntuación R alta^2 , seguido de cerca por ANCOVA. Por otro lado, observamos que la regresión de vectores de apoyo (SVR) no es adecuada para predecir el rendimiento académico de los estudiantes porque tiene un MSE alto, un RMSE alto y una puntuación R baja.^{2}

8 Discusión

Dado el R^2= 73% de la variabilidad de la variable dependiente, la clase es explicada por las 16 variables explicativas. El resto de la variabilidad se debe a otras variables explicativas que no han sido consideradas durante la presente investigación experimental. La tabla 4 muestra el análisis de la suma de cuadrados de tipo III. Esta tabla es muy importante para determinar si las variables explicativas proporcionan o no información significativa.

Tabla 4. Análisis de la suma de cuadrados de tipo III

Caracte rística	DF	Suma de cuadrados	Cuadrados medios	F	Pr > F
Mano alzada	1.000	2.470	2.470	14.380	0.000
Recursos visitados	1.000	4.327	4.327	25.197	0.000
Ver anuncios	1.000	0.625	0.625	3.638	0.057
Grupos de debate	1.000	0.353	0.353	2.056	0.152
Género	1.000	1.432	1.432	8.336	0.004
Nacionalidad	8.000	1.981	0.248	1.442	0.177
Lugar de nacimiento	8.000	2.018	0.252	1.469	0.166
Etapas educativas	1.000	0.120	0.120	0.699	0.403
Niveles de grado	9.000	1.723	0.191	1.115	0.350
Sección ID	2.000	0.013	0.007	0.038	0.963
Tema	11.000	2.347	0.213	1.243	0.256
Semestre	1.000	0.050	0.050	0.294	0.588
Padre responsable	1.000	2.497	2.497	14.541	0.000
Padres que responden a la encuesta	1.000	2.319	2.319	13.506	0.000
Satisfacción de los padres con la escuela	1.000	0.276	0.276	1.606	0.206
Días de ausencia de los estudiantes	1.000	23.444	23.444	136.517	0.000

Según la prueba F de Fisher, cuanto menor sea la probabilidad F correspondiente a una determinada variable, mayor será el impacto de la variable en el modelo. En la tabla anterior, podemos ver que el valor p de "Ver anuncios". "Grupos de discusión". "Género". "Nacionalidad". "Lugar de nacimiento". "Etapas educativas". "Niveles de grado". "ID de la sección". "Tema". "Semestre" y "Satisfacción escolar de los padres" son 0,057. 0.152. 0.004. 0.177.
0.166. 0.403. 0.350. 0.963. 0.256. 0,588. y 0,206 respectivamente. Esto confirma el escaso impacto de estos parámetros en el modelo. Por otra parte, es evidente que el valor p de "Mano alzada". "Recursos visitados". "Padres responsables". "Encuesta de respuesta de los padres" y "Días de ausencia de los estudiantes" es 0. Por lo tanto, estos parámetros aportan información significativa a nuestro modelo. Además, basándose en los errores de tipo III, se puede inferir que la variable explicativa más influyente es "Días de ausencia del estudiante". El siguiente gráfico indica los valores predichos frente a los valores observados. Además. Los intervalos de confianza para la media permiten detectar posibles valores atípicos.

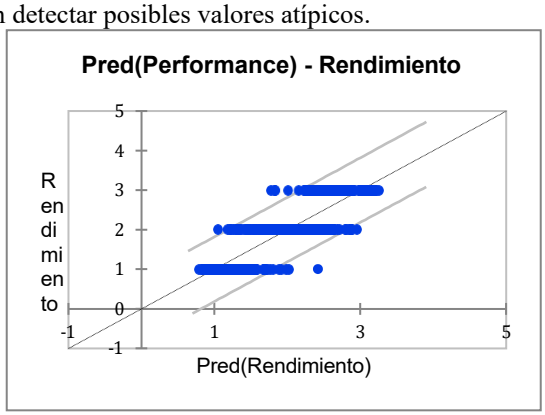

Fig. 7. Valores predichos frente a los valores observados

El siguiente histograma representa los residuos estandarizados frente al rendimiento. Indica que los residuos crecen con el rendimiento. Como podemos ver en la Figura 8, el gráfico de barras de los residuos permite mostrar rápidamente los residuos que están fuera del rango [-2. 2].

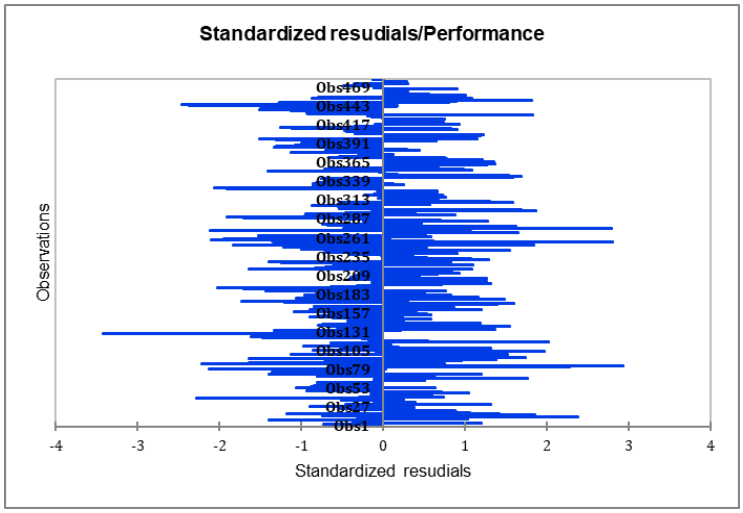

Fig. 8. Residuos normalizados frente al rendimiento

Como conclusión. "Mano alzada". "Recursos visitados". "Padre responsable". "Padres que responden a la encuesta" y "Días de ausencia del alumno" nos permiten explicar el 73% de la variabilidad del rendimiento. Sería necesario un análisis más profundo porque una cantidad de información no es explicada por nuestro modelo.

9 Conclusión y trabajo futuro

En los últimos años, predecir el rendimiento académico de un estudiante es el principal objetivo de todas las instituciones educativas. Los numerosos estudios demuestran que el aprendizaje automático puede ser una tecnología eficiente para cumplir este objetivo. En este trabajo de investigación, nuestro primer objetivo fue comparar varios algoritmos de aprendizaje automático para predecir el rendimiento académico de los estudiantes. Por lo tanto, aplicamos y evaluamos varios algoritmos que son ANCOVA. Logit-R. SVR. Log-LR. DTR. RFR y PLS-R. Nuestro segundo objetivo era determinar las relaciones entre las características y el rendimiento académico de los estudiantes. Como resultado de nuestro estudio experimental, podemos concluir que "Mano levantada". "Recursos visitados". "Padre responsable". "Padres que responden a la encuesta" y "Días de ausencia del alumno" proporcionan una cantidad significativa de información para predecir el rendimiento académico del alumno. Ciertamente, este trabajo de investigación tiene algunas limitaciones. Por ello, las principales orientaciones para futuros trabajos podrían centrarse en lo siguiente: En primer lugar, aplicar técnicas como la agrupación y las redes neuronales artificiales para mejorar la predicción. En segundo lugar, utilizar un conjunto de datos de gran tamaño y con diversas características para abordar el problema de

escalabilidad. El último aspecto que puede mejorarse es la explotación de algunos algoritmos híbridos de selección de características.

10 Referencias

[1] Kalaivani. S. Priyadharshini. B. &Nalini. B. S. (2017). Analizar el per- fecto académico del estudiante basado en el enfoque de minería de datos. International Journal of Innovative Research in Computer Science and Technology. 5(1). 194-197. https://doi.org/10.21276/ijircst.2017. 5.1.4

[2] Moubayed. A. Injadat. M. Shami. A. &Lutfiyya. H. (2020). Student engagement level in e-learning environment: Clustering using k-means. American Journal of Distance Education. 1-20. https://doi.org/10.1080/08923647.2020.1696140

[3] Alenezi. H. S. &Faisal. M. H. (2020). Utilizing crowdsourcing and machine learning in education: Literature review. Education and Information Technologies. 1-16. https://doi.org/10.1007/s10639-020-10102-w

[4] El Guabassi. I. Al Achhab. M. Jellouli. I. & El Mohajir. B. E. (2016. Octubre). Recommender system for ubiquitous learning based on decision tree. En 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt) (pp. 535-540). IEEE. https://doi.org/10.1109/cist.2016.7805107

[5] Hew. K. F. Hu. X. Qiao. C. & Tang. Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. Computers & Education. 145. 103724. https://doi.org/10.1016/j.compedu.2019.103724

[6] Qazdar. A. Er-Raha. B.. Cherkaoui. C. &Mammass. D. (2019). A machine learning algorithm framework for predicting students' performance: Un estudio de caso de estudiantes de bachillerato en Marruecos. Educación y tecnologías de la información. 24(6). 3577-3589. https://doi.org/10.1007/s10639-019-09946-8

[7] Huang. A. Y. Lu. O. H. Huang. J. C. Yin. C. J. & Yang. S. J. (2020). Predicción del rendimiento académico de los estudiantes mediante el uso de big data educativo y learning analytics: evaluación de los métodos de clasificación y los registros de aprendizaje. Interactive Learning Environments. 28(2). 206-230. https://doi.org/10.1080/10494820.2019.1636086

[8] Waheed. H. Hassan. S. U. Aljohani. N. R. Hardman. J. Alelyani. S. & Nawaz. R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior. 104. 106189. https://doi.org/10.1016/j.chb.2019.106189

[9] El Guabassi. I. Al Achhab. M. Jellouli. I. &Mohajir. B. E. E. (2018). Aprendizaje ubicuo personalizado a través de un motor adaptativo. Revista internacional de tecnologías emergentes en el aprendizaje (iJET). 13(12). 177-190. https://doi.org/10.3991/ijet.v13i12.7918

[10] Syed. A. M. Ahmad. S. Alaraifi. A. & Rafi. W. (2020). Identificación de los riesgos operativos que impiden la implementación del eLearning en el sistema de educación superior. Education and In- formation Technologies. 1-17. https://doi.org/10.1007/s10639-020-10281-6

[11] Bousalem. Z. El Guabassi. I. &Cherti. I. (2018. Julio). Hacia un contenido de aprendizaje adaptable y reutilizable utilizando esquemas de etiquetado dinámico XML y bases de datos relacionales. En Conferencia Internacional sobre Sistemas Inteligentes Avanzados para el Desarrollo Sostenible (pp. 787-799). Springer. Cham. https://doi.org/10.1007/978-3-030-11928-7_71

[12] El Guabassi. I. Bousalem. Z. Al Achhab. M.. y EL Mohajir. B. E. (2019). Identificación del estilo de aprendizaje a través de la tecnología de seguimiento ocular en los sistemas de aprendizaje adaptativo. International Journal

de Ingeniería Eléctrica e Informática (2088-8708). 9. https://doi.org/10.11591/ijece.v9i5.pp4408-4416

[13] Bravo-Agapito. J. Romero. S. J. y Pamplona. S. (2020). Predicción Temprana del Rendimiento Académico del Estudiante de Grado en el Aprendizaje Completo en Línea: A Five-Year Study. Com- puters in Human Behavior. 106595. https://doi.org/10.1016/j.chb.2020.106595

[14] Gray. C. C. y Perkins. D. (2019). Utilizando el compromiso temprano y el aprendizaje automático para pre dictar los resultados de los estudiantes. Computers & Education. 131. 22-32. https://doi.org/10.1016/j. compedu.2018.12.006

[15] Hamsa. H. Indiradevi. S. &Kizhakkethottam. J. J. (2016). Modelo de predicción del rendimiento académico de los estudiantes mediante árbol de decisión y algoritmo genético difuso. Procedia Technology. 25. 326-332. https://doi.org/10.1016/j.protcy.2016.08.114

[16] Hussain. M. Zhu. W. Zhang. W. Abidi. S. M. R. & Ali. S. (2019). Uso del aprendizaje automático para predecir las dificultades de los estudiantes a partir de los datos de las sesiones de aprendizaje. Revista de inteligencia artificial. 52(1). 381-407. https://doi.org/10.1007/s10462-018-9620-8

[17] Karthikeyan. V. G. Thangaraj. P. y Karthik. S. (2020). Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. Soft Computing. 1-11. https://doi.org/10.1007/s00500-020-05075-4

[18] Amrieh. E. A. Hamtini. T. &Aljarah. I. (2016). Minería de datos educativos para predecir el rendimiento académico de los estudiantes utilizando métodos de conjunto. International Journal of Database Theory and Application. 9(8). 119-136. https://doi.org/10.14257/ijdta.2016.9.8.13

[19] Amrieh. E. A. Hamtini. T. &Aljarah. I. (2015. Noviembre). Preprocesamiento y análisis del conjunto de datos educativos utilizando X-API para mejorar el rendimiento de los estudiantes. En 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) (pp. 1-5). IEEE. https://doi.org/10.1109/aeect.2015.7360581

[20] "Sistema de aprendizaje Kalboard360-E". http://kalboard360.com/

[21] Rutherford. A. (2001). Introducing ANOVA and ANCOVA: a GLM approach. Sage.

[22] Kleinbaum. D. G.. Dietz. K. Gail. M.. Klein. M. & Klein. M. (2002). Logistic regression. New York: Springer-Verlag.

[23] Smola. A. J. &Schölkopf. B. (2004). A tutorial on support vector regression. Statistics and computing. 14(3). 199-222. https://doi.org/10.1023/b:stco.0000035301.49549.88

[24] Heien. D. M. (1968). A note on log-linear regression. Journal of the American Statistical Association. 63(323). 1034-1038. https://doi.org/10.2307/2283895

[25] Safavian. S. R. &Landgrebe. D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems. man. and cybernetics. 21(3). 660-674. https://doi.org/10.1109/21.97458

[26] Liaw. A. y Wiener. M. (2002). Clasificación y regresión por randomForest. R news. 2(3). 18-22.

[27] Geladi. P. y Kowalski. B. R. (1986). Partial least-squares regression: a tutorial. Analytica chimica acta. 185. 1-17. https://doi.org/10.1016/0003-2670(86)80028-9

[28] Swalin. A. (2018). Elección de la métrica correcta para evaluar los modelos de aprendizaje automático.

[29] Miles. J. (2014). R al cuadrado. R al cuadrado ajustado. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat06627

[30] Willmott. C. J. y Matsuura. K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research. 30(1). 79-82. https://doi.org/10.3354/cr030079

[31] Addinsoft. X. (2015). Análisis de datos y estadística con MS Excel. Addinsoft. NY. USA. xlstat disponible en http://www.xlstat.com/en/home

11 Autores

Inssaf El Guabassi se doctoró en febrero de 2019 en la Universidad Abdelmalek Es- saadi, Facultad de Ciencias, Marruecos, Tetuán.

Zakaria Bousalem es estudiante de doctorado en la Facultad de Ciencias y Tecnología de Settat, Marruecos.

Rim Marah se doctoró en julio de 2018 en la Universidad Abdelmalek Essaadi, Facultad de Ciencias, Marruecos, Tetuán.

Aimad Qazdar es profesor adjunto en la Facultad de Ciencias Semlalia, Laboratorio ISI - Universidad Cadi Ayyad en Marrakech, Marruecos.

Artículo presentado el 24 de noviembre de 2020. Reenviado 2020-12-08. Aceptación final 2020-12-11. Versión final p u b l i c a d a tal y como fue presentada por los autores.