

RECIBIDO EL 17 DE AGOSTO DE 2021 - ACEPTADO EL 16 DE NOVIEMBRE DE 2021

# ANÁLISIS DEL RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO CON MÉTODOS DE ENSAMBLE

## ANALYSIS OF ACADEMIC PERFORMANCE USING MACHINE LEARNING TECHNIQUES WITH ASSEMBLY METHODS

Leonardo Emiro Contreras Bravo<sup>1</sup>

Héctor Javier Fuentes López<sup>2</sup>

Edwin Rivas Trujillo<sup>3</sup>

Universidad Distrital Francisco José de Caldas,  
Bogotá D.C., Colombia.

### RESUMEN

En los últimos años el campo educativo se ha visto permeado por modelos y algoritmos de análisis de datos que pretenden la búsqueda de conocimiento a partir de los datos para mejorar el rendimiento académico y otros indicadores. El

<sup>1</sup> Ingeniero, Estudiante de doctorado en ingeniería, Docente de planta. Facultad de ingeniería. Universidad Distrital Francisco José de Caldas (Colombia). [lecontrerasb@udistrital.edu.co](mailto:lecontrerasb@udistrital.edu.co). ORCID: <https://orcid.org/0000-0003-4625-8835>

<sup>2</sup> Economista, Estudiante de doctorado en Economía. Docente de planta. Facultad de ingeniería. Universidad Distrital Francisco José de Caldas (Colombia). [hj-fuentesl@udistrital.edu.co](mailto:hj-fuentesl@udistrital.edu.co). ORCID: <https://orcid.org/0000-0001-6899-4564>

<sup>3</sup> Ingeniero, Doctor en ingeniería. Docente de planta. Facultad de ingeniería. Universidad Distrital Francisco José de Caldas (Colombia) [erivas@udistrital.edu.co](mailto:erivas@udistrital.edu.co). ORCID: <https://orcid.org/0000-0003-2372-8056>

objetivo principal de este trabajo es predecir el rendimiento académico de estudiantes mediante técnicas de aprendizaje automático. Se analizan 324 variables mediante métodos de selección de características, con el fin de determinar las variables influyentes. El modelo de predicción del rendimiento académico universitario es evaluado por medio de algoritmos supervisados (KNN, SVC, Naive Bayes y árbol de decisión), los cuales son optimizados mediante lenguaje Python. Además, son implementados algoritmos de ensamble que permiten mejorar la exactitud de los clasificadores previos. Se implementan métodos Bagging (CART, Random Forest; ExtraTreesClassifier), métodos Boosting (AdaBoost, GBM, XGBoost, CatBoost, Light

Boost) y Voting (Blending, Stacking). Los resultados muestran que los algoritmos Stacking y Blending con valores de precisión en cada semestre que oscilan alrededor de 85% y 75% para entrenamiento y prueba respectivamente arrojan los mejores resultados.

**PALABRAS CLAVE:** Análisis de datos educativo; aprendizaje automático; educación superior; modelo; rendimiento académico

## SUMMARY

In recent years, the educational field has been permeated by data analysis models and algorithms that aim to search for knowledge from data to improve academic performance and other indicators. The main objective of this research is to predict the academic performance of students through machine learning techniques. Through feature selection methods are analyzed 324 variables, in order to determine the influential variables. The university academic performance prediction model is evaluated by means of supervised algorithms (KNN, SVC, Naive Bayes and decision tree), which are optimized using Python language. In addition, assembly algorithms are implemented that allow improving the accuracy of the previous classifiers. Bagging (CART, Random Forest; ExtraTreesClassifier), Boosting (AdaBoost, GBM, XGBoost, CatBoost, Light Boost) and Voting (Blending, Stacking) methods are implemented. The results show that the Stacking and Blending algorithms with accuracy values in each semester that oscillate around 85% and 75% for training and testing, respectively, yield the best results.

**KEYWORDS:** Boosting, Educational data analytics; Ensemble, machine learning, Student academic performance

## INTRODUCCIÓN

El sector educativo ha experimentado un aumento en la cantidad de información que generan los diferentes actores del proceso.

Debido a esto, se ha planteado la necesidad de búsqueda y desarrollo de modelos que permitan dar explicación a diversos fenómenos de este campo como lo son: el rendimiento académico, la tasa de deserción y la tasa de graduación entre otros. La búsqueda de modelos ha sido abordada desde diferentes campos tales como la psicología, medicina, educación e ingeniería, donde se pretende predecir con buena exactitud un indicador académico que facilite u oriente el transcurrir académico de los estudiantes

La eficiencia del proceso educativo puede ser medido por indicadores como: la tasa de deserción anual, la tasa de deserción por cohorte, la tasa de graduación y la tasa de ausencia intersemestral, los cuales poseen relación con el rendimiento académico (De La Hoz et al., 2019), el cual es una variable multidimensional afectada por variables sociales, económicas, demográficas ,etc.). De acuerdo con el Ministerio de Educación de Colombia, estas métricas no son las mejores debido a que en el año 2017 la tasa de deserción por cohorte era del 45,1% y la tasa de graduación era del 21.17% para el año 2019. Ahora bien, estos indicadores no solo preocupan la situación colombiana, ya que en una situación similar o quizás peor se encuentran muchos de los países latinoamericanos. (Ferreya et al., 2017)

El rendimiento académico puede ser definido como la expresión de las capacidades, conocimientos y características psicológicas que se desarrollan o se actualizan a través del proceso formativo de enseñanza-aprendizaje con el fin de que el estudiante alcance logros académicos, y que se evidencien al final de un semestre a través de una evaluación objetiva, que muchas veces es una nota final. Así mismo, puede predecirse según diversos niveles tales como lo expresa (Alyahyan & Düşteğör, 2020): a nivel de titulación: predecir el éxito de los estudiantes en el momento de la obtención de la titulación; a nivel del año: predecir el éxito de los

estudiantes al final del año o semestre; a nivel del curso: predecir el éxito de los estudiantes en un curso específico; y a nivel de examen: predecir el éxito de los estudiantes en un examen de un curso específico. Esta tarea de predicción se vuelve cada vez más desafiante debido al gran volumen de datos presentes en las instituciones, que integrados al desarrollo de la ciencia de datos con sus algoritmos de Machine Learning pueden facilitar la toma de decisiones en el campo educativo siendo útil tanto para estudiantes como para docentes (Murnion & Helfert, 2013).

El rendimiento académico es un tema que ha sido ampliamente estudiado y controvertido en el campo de la investigación educativa (Lamas, 2015; Montero et al., 2007) Estas investigaciones varían desde su propio concepto hasta su medición y estimación. Algunos autores consideran que puede definirse como una medida de las capacidades adquiridas en el proceso formativo, como el cumplimiento de logros en un programa cursado, como el nivel de conocimiento de un área respecto a una norma, entre otros (Lamas, 2015). Otros manifiestan que lo importante yace en el hecho de cómo medirlo. Tal es el caso de los planteamientos sobre la medición a través de las calificaciones (Lamas, 2015), o tal vez a través de la inclusión de otras variables que pueden incidir en su medición (Hernández, 2016; Rodríguez & Ruíz, 2009).

Este ha sido abordado desde el campo de la psicología (Guizado et al., 2020) (Zárate et al., 2020), (Céspedes et al., 2018), (Rivera et al., 2020), (Minichil et al., 2020) quienes por medio de encuestas y cuestionarios han permitido recolectar información la cual según los investigadores resulta muy útil para la investigación sobre sentimientos, pensamientos y percepciones de cada uno de los individuos a través de puntuaciones determinadas en instrumentos que son validados por medio

del coeficiente de Cronbach. Desde el campo estadístico, el análisis descriptivo se constituye en la estrategia más importante de análisis, por ser una herramienta estadística que engloba muchos indicadores, al igual que el Coeficiente de correlación de Pearson y ANOVA (Durán & Rosado, 2019), (Lenskiy et al., 2020), (Lloret-Segura et al., 2014; López-Aguado & Gutiérrez-Provecho, 2019; Mavrou, 2015) validación y adaptación de instrumentos de medida psicológicos. Su uso se extendió durante los años 60 y ha ido creciendo de forma exponencial al ritmo que el avance de la informática ha permitido. Los criterios empleados en su uso, como es natural, también han evolucionado. Pero los investigadores interesados en asuntos sustantivos que utilizan rutinariamente esta técnica permanecen en muchos casos ignorantes de todo ello. En las últimas décadas numerosos trabajos han denunciado esta situación. La necesidad de actualizar los criterios clásicos para incorporar aquellos más adecuados es una necesidad urgente para hacer investigación de calidad. En este trabajo se revisan los criterios clásicos y, según el caso, se sustituyen o se complementan con otros más actuales. El objetivo es ofrecer al investigador aplicado interesado una guía actualizada acerca de cómo realizar un Análisis Factorial Exploratorio consonante con la psicometría post-Little Jiffy. Esta revisión y la guía con las recomendaciones correspondientes se han articulado en cuatro grandes bloques: 1 que permiten establecer relaciones y efectos entre las variables de estudio.

El aprendizaje automático o Machine Learning puede definirse como el proceso de aprender un conjunto de reglas a partir de instancias o de manera más general por medio de la creación de un clasificador que se puede utilizar para generalizar a partir de nuevas instancias (Acharya & Sinha, 2014). Suele dividirse en varias áreas según la naturaleza del etiquetado de datos como se muestra en la Figura 1.

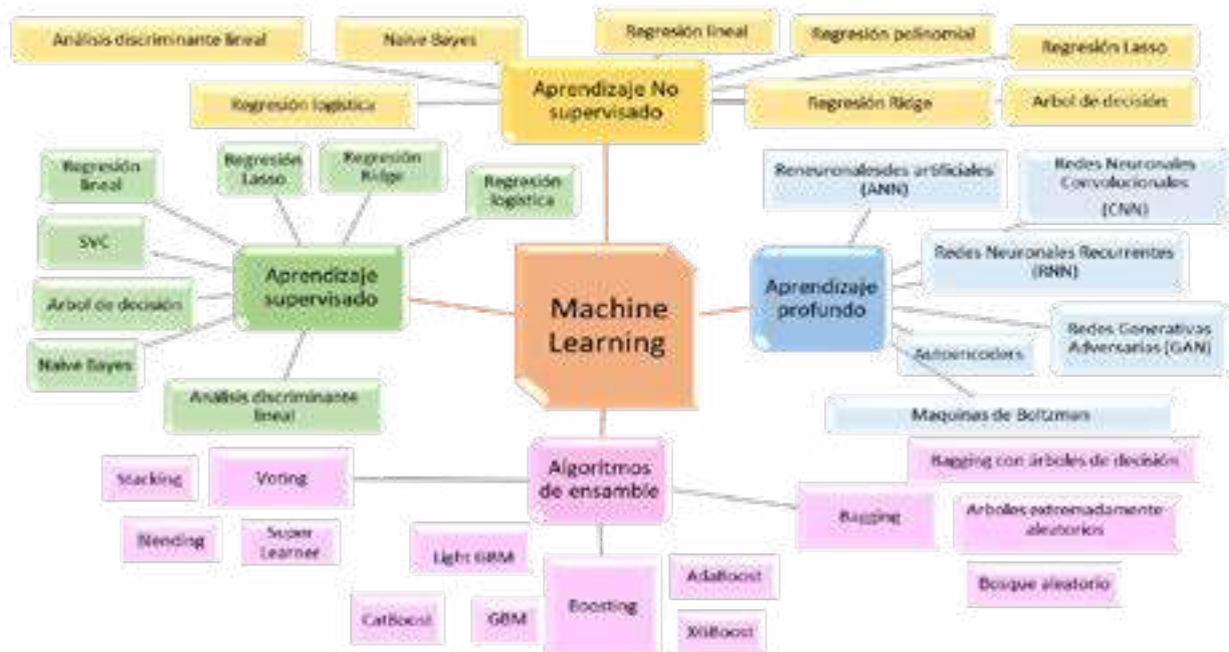


Figura 1. Líneas o campos del Machine Learning. Fuente: Elaboración propia.

Con el avance de la ciencia de datos y los algoritmos de aprendizaje automático, también se han planteado una serie de trabajos relacionados con la influencia de diversas variables sobre el rendimiento académico, por ser este quien establece el éxito o fracaso en el proceso estudiantil (Garbanzo & María, 2007).

El aprendizaje supervisado clásico se origina cuando cada una de las observaciones o muestras del conjunto de datos tiene relacionado una variable o un dato que indica lo que sucedió, lo que pasó, es decir las entradas están etiquetadas. Este tipo de aprendizaje se subdivide en clasificación y regresión. En clasificación las salidas del sistema son finitas y discretas y son interpretadas como la clase a la que pertenece, (G. García, 2014)

Dentro de esta línea del Machine Learning o aprendizaje automático supervisado se ha planteado diversas investigaciones utilizando algoritmos clásicos como KNN, SVC, regresión logística, Naive Bayes y árbol de decisión (Mengash, 2020a), (Castrillón et al., 2020), (Viloria et al., 2020) many studies have emerged about regarding the topic of school failure, showing a growing interest in determining the multiple factors that may influence it [1]. Most of the researches that attempt to solve this issue [2] are focused on determining the factors that most affect the performance of students (dropout and failure con resultados prometedores de su aplicación. El aprendizaje no supervisado es considerado una subdivisión del aprendizaje automático, que consiste en que solo se proporcionan muestras de entrada al sistema de aprendizaje a partir de las cuales

el algoritmo plantea agrupaciones o jerarquías. Desde el aprendizaje no supervisado clásico existen investigaciones para identificar patrones de comportamiento en las variables, en donde se agrupan a los estudiantes en clústeres de acuerdo a diversas características (Guleria & Sood, 2018), (Sajjadi, S. Shapiro, B. Mckinlay, C. Sarkisyan, A. Shubin & Osoba, 2018), (Martinez-Rodriguez et al., 2019), (Joshika, P. and Rajeshwari, 2019), (Santoso & Yulia, 2019), (Moubayed et al., 2020), (Kumar, V. Krishna, A. Neelakanteswara, P. Basha, 2020). Por otro lado, se han usado algoritmos supervisados clásicos cuyo fin es realizar predicciones sobre diversos valores de salida o clases, los cuales son clasificados según el tipo de variable de salida (Mengash, 2020b), (Viloria et al., 2020)(Bhutto et al., 2020) (Anderson et al., 2019), (Jalota & Agrawal, 2019). En estas investigaciones suelen utilizarse variables de tipo académico y demográficas procesadas mediante algoritmos de árboles de decisión y KNN, ya que permite a los investigadores tener un buen panorama del rendimiento estudiantil.

El aprendizaje automático contemporáneo ha sido dividido según la literatura en dos líneas: los métodos de conjunto o ensamble y las redes neuronales o aprendizaje profundo. Los primeros consisten en métodos que a partir de un mismo conjunto de datos, construye varios predictores y los combina de alguna forma para poder obtener un predictor más estable y con mayor rendimiento que los predictores débiles (algoritmos supervisados) si trabajasen en solitario (Bourel, 2012; Orihuela Maita, 2019; Patacsil, 2020). Las principales técnicas que se desarrollan dentro de los métodos de ensamble en Machine Learning son Bagging, Boosting y Stacking cuyas investigaciones se han enfocado al rendimiento del estudiante dentro de una asignatura o un grupo de asignaturas de un semestre específico o un semestre (Kostopoulos et al., 2018), (Campo-Ávila et al., 2018); al finalizar su primer año de universidad (Yamao

et al., 2018), (Candia Oviedo, 2019), (Sweeney et al., 2016), (Ochoa et al., 2017), (Vega García, 2019), (Adekitan & Salau, 2019).

Esta investigación tiene como objetivos identificar los atributos que influyen de mejor manera en la determinación del rendimiento académico de los estudiantes partiendo de un gran volumen de variables establecidas en la investigación referencial. Así mismo, busca definir las herramientas de análisis de datos de ensamble pertenecientes al campo de machine Learning que mejor permiten determinar el rendimiento académico de estudiantes de Ingeniería de la Universidad Distrital Francisco José de Caldas. Por lo anterior, se evidencia la necesidad de investigar y de pretender un modelo (con futuro desarrollo de software) que prediga de la mejor manera la variable multidimensional denominada rendimiento académico, implementando diversos mecánicos y técnicas que permitan predecirlo y de esta manera faciliten la toma de decisiones en el campo educativo tanto para estudiantes como para docentes. Este trabajo es un aporte en ese camino de búsqueda.

Este artículo presenta el orden secuencial de las secciones de la siguiente manera: La sección 2 describe la metodología y herramientas utilizadas en la investigación. Esto incluye recopilación, procesamiento y algoritmos de modelado de datos. La sección 3 detalla las experiencias realizadas y sus resultados. La sección 4 proporciona la discusión de los resultados. Finalmente, la sección 5 explica las conclusiones de tipo general, así como las sugerencias de trabajo futuro.

## MATERIALES Y METODOS

La metodología empleada se desarrolla en tres etapas complementarias (Figura 2) y se resume en los siguientes pasos: (1) Participantes y variables; (2) tratamiento y transformación, (3) selección de características; (4) algoritmos de ensamble y evaluación.

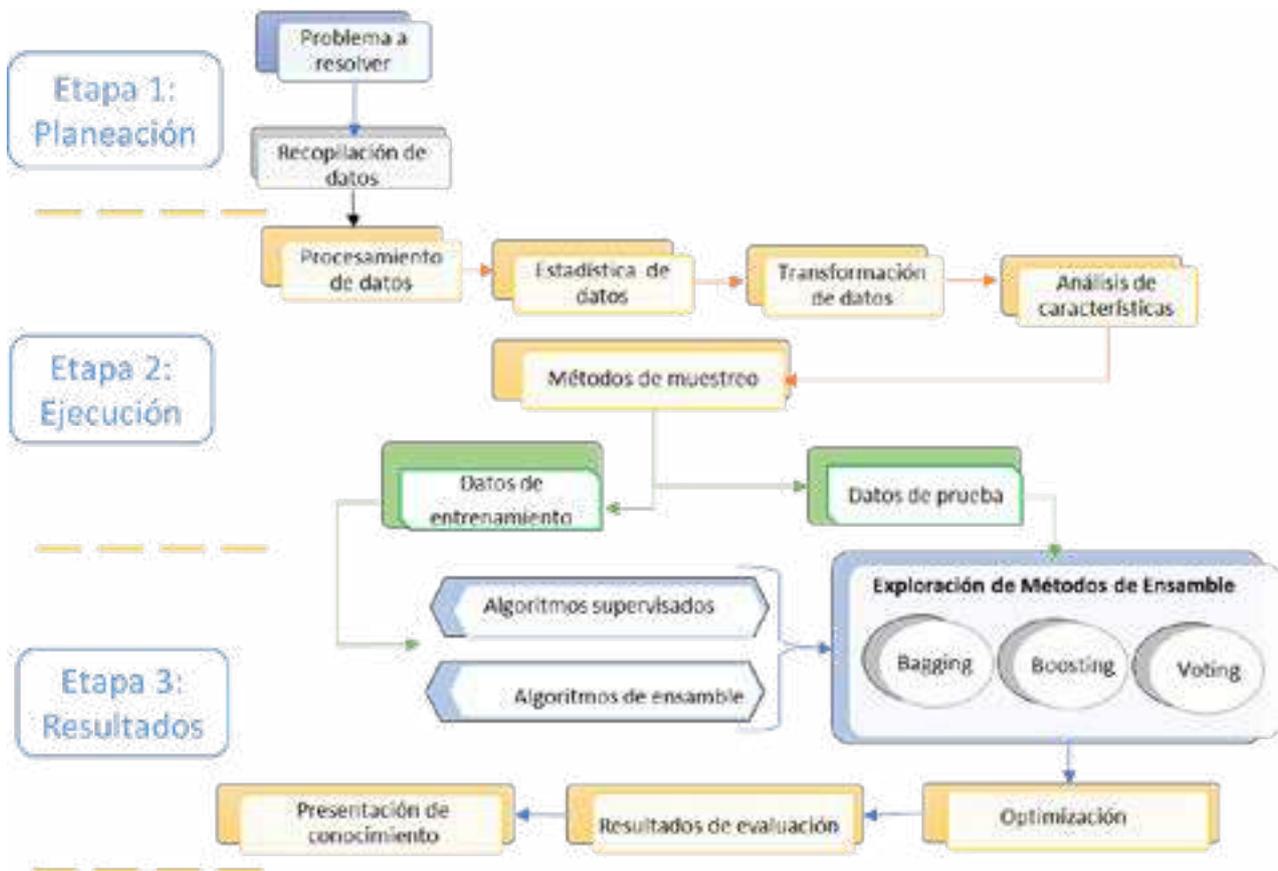


Figura 2: Pasos del modelo de predicción del rendimiento académico con método de ensamble.

Fuente: Elaboración propia.

## PARTICIPANTES Y VARIABLES

La investigación parte de un conjunto de datos suministrados por la Universidad Distrital (Colombia), que corresponden a 2476 estudiantes de ingeniería Industrial del periodo 2008 al 2018 (total de 843.696 datos). Las 324 variables fueron analizadas mediante Jupyter Notebook mediante lenguaje Python.

## TRATAMIENTO Y TRANSFORMACIÓN

La base de datos fue sometida a un proceso de limpieza (por ejemplo: datos perdidos, incompletos, erróneos) que arroja 2300 registros con los cuales se continúa el proceso de estudio. Este paso es importante ya que permite mejorar la calidad de los datos y posiblemente obtener un mejor proceso de modelado y rendimiento de

los algoritmos (Costa et al., 2017). Así mismo con el fin de tener un mejor conocimiento básico de los datos fueron realizadas algunas estadísticas descriptivas de variables del conjunto de datos con el fin de identificar outliers y distribución de los datos que conforman a cada una de las variables (Santosh, 2020) which was identified in late 2019, requires special attention because of its future epidemics and possible global threats. Beside clinical procedures and treatments, since Artificial Intelligence (AI).

Una etapa no menos importante, es la de realizar un proceso de transformación. Esto es, poder llevar la distribución de las variables que conforman el conjunto de datos a una distribución cuasi – gaussiana (Jahangiri & Rakha, 2015). Este paso puede hacer que datos medidos en escalas diferentes sean más comparables

entre sí eliminando los efectos de influencia que pueden darse cuando por ejemplo una variable independiente ejerza mayor influencia sobre la variable dependiente por el hecho de que su escala numérica es mayor que para las otras variables (J. García et al., 2019). Algunos de los métodos de transformación utilizados fueron: Reescalar, Normalizar, estandarización robusta, Box-cox y Yeo-Johnson.

### SELECCIÓN DE CARACTERÍSTICAS

Este tópico es importante en el proceso ya que ayuda a eliminar variables del conjunto de datos educativos que no están altamente relacionadas con la variable de salida, aumentando la exactitud de los clasificadores. Los métodos de selección de características implementados se pueden agrupar en cuatro grupos: de filtro, wrapper, embebidos y de ensamble.

### ALGORITMOS DE ENSAMBLE

Los algoritmos usados en este estudio son descritos brevemente a continuación:

Los métodos Bagging (Bootstrap Aggregation) intentan construir un clasificador fuerte a partir del número de clasificadores débiles (un mismo algoritmo) al que se le pasa diferentes subgrupos aleatorios de datos. Cada una de las submuestras que se le pasan al algoritmo contiene elementos con reemplazo (lo que significa que se puede seleccionar el mismo valor varias veces). Al final se promedian las respuestas (Zhi-Hua Zhou, 2012). Dentro de este grupo se encuentran los algoritmos Bagging Decisión Tree, Randomforest y ExtraTreeClassifier

Los métodos Boosting son un grupo de algoritmos de ensamble que intentan construir un clasificador fuerte (que logra una precisión arbitrariamente buena) a partir del número de clasificadores. Su forma de trabajar consiste en que cada uno de los algoritmos elegidos

se entrenan secuencialmente permitiendo crear un clasificador fuerte a partir de varios clasificadores débiles ya que se enfoca cada vez más en instancias mal clasificadas por clasificadores generados previamente (Zhang & Ma, 2012). Esto se hace construyendo un modelo a partir de los datos de entrenamiento, luego creando un segundo modelo que intente corregir los errores del primer modelo (Zhi-Hua Zhou, 2012). Los modelos se agregan hasta que el conjunto de entrenamiento se predice perfectamente o se agrega un número máximo de modelos. Estos métodos buscan que cada nuevo algoritmo (*siempre es el mismo*) aprenda a corregir los errores del algoritmo anterior (Gareth, 2013). Dentro de este grupo se encuentran los algoritmos AdaBoost (Adaptive Boosting), Gradient Boosting Machine (GBM), XGBoost (Extreme Gradient Boosting), CatBoost (Categorical Boosting) y LightGBM (Light Gradient Boosted Machine).

Los métodos Voting (votación por mayoría), son un grupo de técnicas que permite combinar las predicciones de múltiples algoritmos de aprendizaje automático que tal vez han hecho bien su tarea de predicción cuando se usan por separado. Funciona creando primero dos o más modelos independientes a partir de un conjunto de datos de entrenamiento (aprendices de primer nivel). Los resultados que arrojan cada uno de los modelos pasan a un clasificador de votación (aprendiz de segundo nivel o meta-aprendizaje) el cual será el encargado de predecir la clase de la variable de salida (Zhi-Hua Zhou, 2012). Básicamente el clasificador voting trabaja en función de los resultados arrojados por cada uno de los clasificadores débiles previos (modelos de nivel 0) creándose un modelo único (modelo de nivel 1) que será entrenado con estos valores. La predicción de una nueva clase es realizada por el clasificador voting y no por los clasificadores débiles (modelos de nivel 0) previamente usados. Dentro de este grupo se encuentran las técnicas: Votación dura y un solo

tipo de algoritmo base, Votación dura y varios tipos de algoritmo base, Votación suave o dura con varios algoritmos base y diferentes pesos, Blending y Stacking.

## MEDIDAS DE EVALUACIÓN Y OPTIMIZACIÓN

Independiente del tipo de algoritmo de aprendizaje supervisado o de ensamble utilizado en los modelos, a estos se les debe evaluar la calidad o rendimiento de la clasificación. Las métricas (exactitud, precisión, especificidad, sensibilidad) para tareas de clasificación por lo general se extraen por medio de una matriz de confusión (Kaunang & Rotikan, 2018).

## RESULTADOS

Posterior a la aplicación de la secuencia de

pasos descritos en la metodología, se obtuvieron los siguientes resultados.

## TRATAMIENTO Y TRANSFORMACIÓN

A partir de la base de datos que consta de 434 variables, es posible obtener información básica como notas, materias, año de ingreso, pruebas de estado – ICFES; pero esta información no fue suficiente para los análisis que se deseaba hacer. Por tanto, se crearon nuevas variables que permiten tener una base estructural más sólida de acuerdo con las variables influyentes en el rendimiento académico según la investigación referencial (Contreras et al., 2020),(Contreras & López, 2020),(Contreras et al., 2021). La mejora en la exactitud de los modelos de ensamble por efecto del uso de la transformación Yeo – Jhonson es alrededor de un 5% (Figura 3)

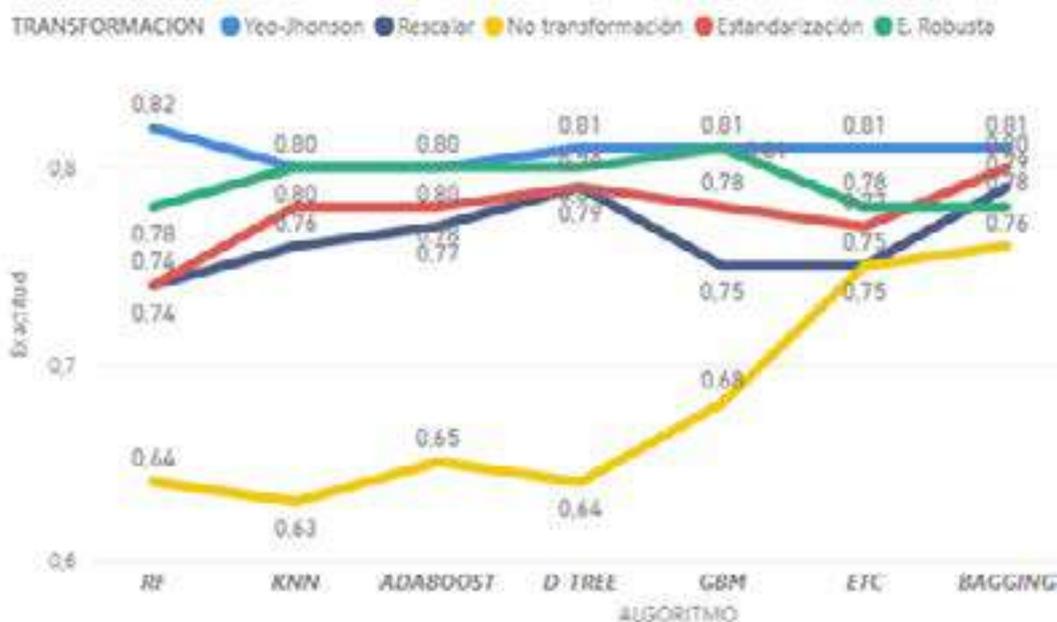


Figura 3. Rendimiento de algoritmos supervisados y de ensamble con diversas transformaciones

## SELECCIÓN DE CARACTERÍSTICAS

La Figura 4a muestra para cada uno de los intervalos de año de vida académica, la cantidad de variables que se analizan y la cantidad de variables que según los métodos son las que más influyen en cada año para la determinación

del rendimiento académico. Los resultados evidencian que no solo es conveniente aplicar métodos de transformación a los datos, sino proporcionar a los algoritmos las variables más influyentes, hecho que se evidencia en mejora de las métricas de rendimiento de los algoritmos. En promedio la mejora

resulta ser de aproximadamente un 6% de la exactitud. Se utilizaron métodos de selección de características tales como: Chi Cuadrado, anova, regularización Lasso, regularización Ridge, Backward Selection, Forward Selection, eliminación de características recursivas (RFE), GBM, XGBoost, Random Forest, ExtraTreesClassifier, LightGBM y CatBoost. En la Figura 4b muestra las características

relevantes para estimar el rendimiento académico del año 4 (semestre 9). Las variables influyentes que arrojaban los métodos Boosting (XGBoost, LightGBM y CatBoost) y las que arrojaban los métodos de ensamble Bagging (Random Forest, ExtraTreesClassifier), son las que proporcionaban mejores métricas de evaluación de los algoritmos implementados

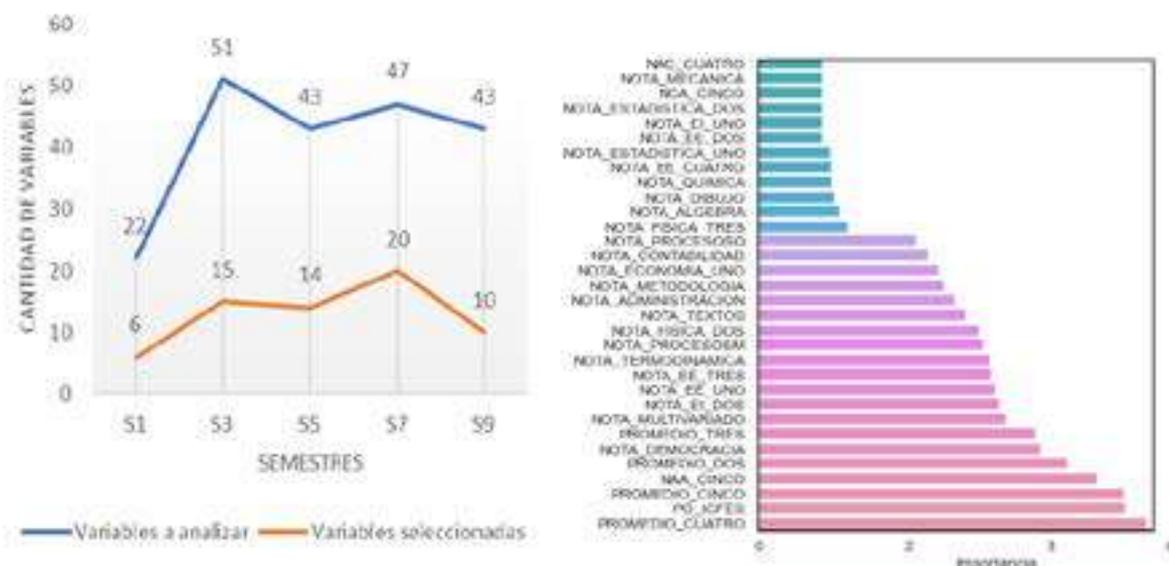


Figura 4. a) Cantidad de características óptimas para predecir el rendimiento académico para cada año b) Mejores características según método LightGBM

Las variables tenidas en cuenta para predecir el rendimiento académico en cada uno de los años académicos son mostradas en la Tabla 1. Las variables cuando se pretende predecir el rendimiento semestre tras semestre de la vida académica de un estudiante son diferentes a las mostradas aquí.

Tabla 1: Variables influyentes en rendimiento académico por cada año de vida académica del estudiante

Semestre	Variables relevantes
1	Puntaje global ICFES, puntaje ICFES para área de condición matemática, puntaje ICFES para área de física, puntaje ICFES para área de biología, puntaje ICFES para área de idioma, localidad del colegio
3	Puntaje global ICFES, puntaje ICFES para área de condición matemática, puntaje ICFES para área de física, puntaje ICFES para área de biología, puntaje ICFES para área de idioma, localidad del colegio, promedio de estudiante (1 semestre), condición de estado al final del semestre, nota de cálculo diferencial, nota de dibujo de ingeniería, nota de catedra, nota de comprensión de textos, numero de asignaturas cursadas en semestre 1, numero de asignaturas aprobadas en 1 semestre
5	puntaje global ICFES, promedio de estudiante (1 semestre), nota de electiva extrínseca uno, promedio de estudiante (2 semestre), numero de asignaturas aprobadas en 2 semestre, nota de materiales, nota de electiva extrínseca dos, promedio de estudiante (3 semestre), numero de asignaturas cursadas en semestre 3, numero de asignaturas aprobadas en 3 semestre, nota de cálculo multivariado, nota de estadística uno, nota de termodinámica, nota de ética, nota de electiva extrínseca tres
7	Puntaje global ICFES, puntaje ICFES para área de condición matemática, promedio de estudiante (1 semestre), nota de comprensión de textos, nota de electiva extrínseca uno, promedio de estudiante (2 semestre), nota de materiales, promedio de estudiante (3 semestre), nota calculo multivariado, nota de ética, nota de prog. orientada, nota de electiva extrínseca tres, promedio de estudiante (4 semestre), nota de economía uno, promedio de estudiante (5 semestre), nota de metodología, nota de contabilidad, nota de procesos mecánicos, numero de asignaturas cursadas en semestre 5, numero de asignaturas aprobadas en 5 semestre
9	Puntaje global ICFES, puntaje ICFES para área de condición matemática, promedio de estudiante (1 semestre), nota de electiva extrínseca uno, promedio de estudiante (2 semestre), promedio de estudiante (3 semestre), promedio de estudiante (4 semestre), promedio de estudiante (5 semestre), nota de ética, nota de derecho, nota de ingeniería económica, promedio de estudiante (7 semestre)

## ALGORITMOS DE ENSAMBLE Y EVALUACIÓN

Después de aplicar las técnicas de transformación y selección de características, se procedió a la búsqueda de los mejores hiperparámetros de los algoritmos débiles que pudieran arrojar buena métrica de rendimiento. La Tabla 2 muestra estos resultados utilizando una validación cruzada de 10 k-fold (compensaría los efectos de la varianza y el sesgo en el conjunto de datos de entrenamiento) en cada semestre. El clasificador

de árbol de decisión genera mejores resultados en comparación con los otros clasificadores. Ofrece un 74.28% de precisión en promedio.

Tabla 2: Resultados de algoritmos supervisados utilizando Validación cruzada K-fold

Algoritmo	SEM_1	SEM_3	SEM_4	SEM_5	SEM_6	SEM_7	SEM_8	SEM_9	SEM_10
KNN	0,621	0,782	0,773	0,8	0,789	0,745	0,692	0,744	0,725
SVC	0,637	0,660	0,755	0,7	0,235	0,520	0,603	0,737	0,670
Arbol	0,602	0,792	0,796	0,8	0,806	0,7110	0,746	0,786	0,713
Naive Bayes	0,607	0,644	0,654	0,7	0,660	0,644	0,517	0,737	0,710

Tanto los algoritmos supervisados como los de ensamble, que se basan en establecer diferentes métodos que trabajan en conjunto con el fin de reducir los errores que proporcionarían cada uno de los algoritmos débiles supervisados, pueden ser optimizados. La búsqueda de mejores hiperparámetros para cada uno de los algoritmos de ensamble, se realizó por medio del método GridSearch, que es un enfoque para el ajuste que construirá y evaluará metódicamente un

modelo para cada combinación de parámetros del algoritmo especificados con antelación en una malla. La Figura 5 muestra un gráfico de optimización de algunos hiperparámetros tales como "n\_estimators" que representa la cantidad de árboles para el algoritmo Randomforest y "max\_depth" que corresponde a la profundidad máxima del árbol; es decir, es la expansión máxima de las hojas de los árboles de decisión que se empleen en el algoritmo Randomforest.

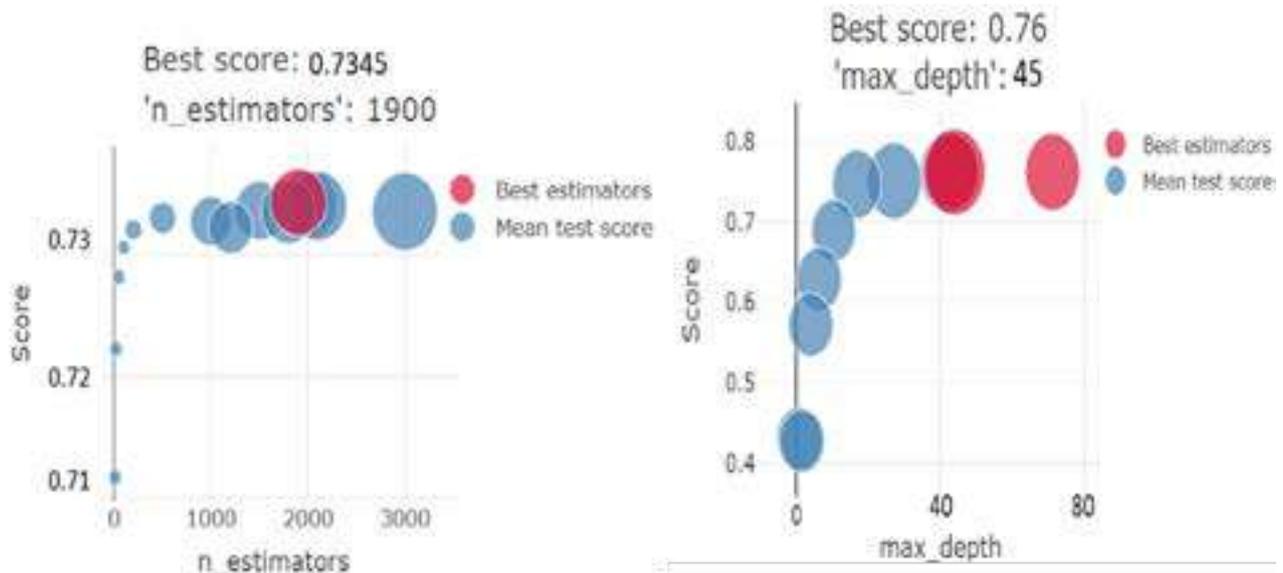


Figura 5. Optimización de algunos hiperparámetros

Los resultados de rendimiento de cada uno de ellos al predecir el rendimiento académico de los estudiantes, cuando se este se estima por año de estudio se muestran en la Figura 6. Los métodos de ensamble XGBoost y CatBoost muestran valores de "exactitud" (entre otras

métricas) superiores a los otros métodos de ensamble, cuyos valores son aún superiores a los conseguidos cuando se plantearon como algoritmos para el modelo de rendimiento los algoritmos supervisados (SVC,KNN, árbol de decisión, Naive Bayes).

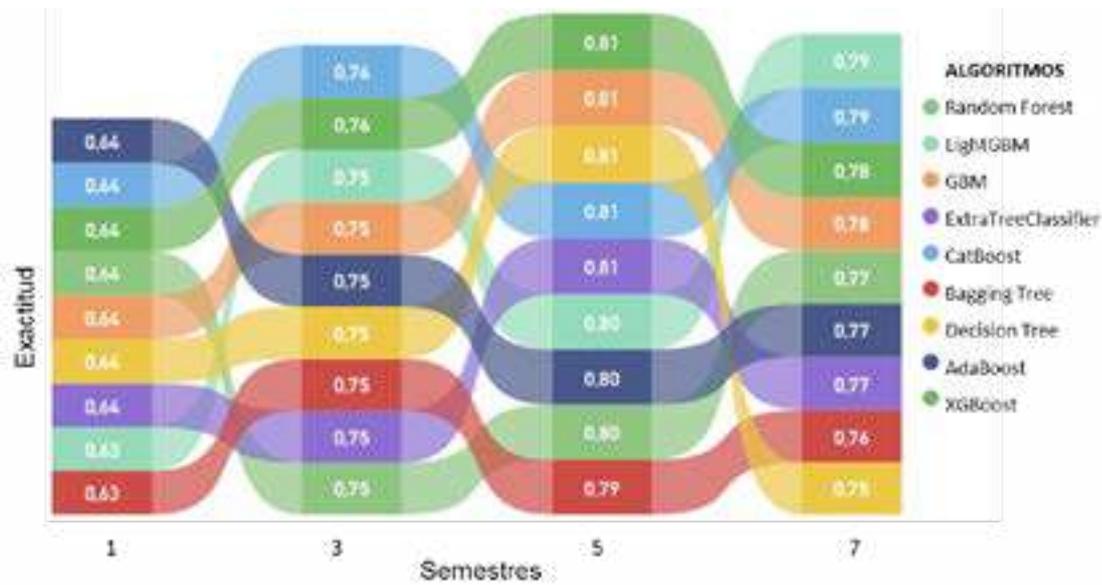


Figura 6. Resumen de métrica de evaluación de Modelos de ML con aprendizaje de conjunto

Aunque los algoritmos de ensamble Bagging y Boosting arrojaron mejores resultados que los supervisados, había que ensayar la metodología "Voting" con sus posibles variantes según la literatura acerca de Machine Learning. Las opciones por ensayar fueron: Voting - Votación Suave y varios tipos de algoritmo base (débil), Voting - Votación Suave y varios tipos de algoritmo base (débil), Blending (modelos débiles y metamodelo lineal), Stacking (modelos débiles y metamodelo no lineal). Cada uno de los algoritmos de ensamble luego de su optimización, fueron probados con sus mejores hiperparámetros con el fin de evidenciar cual arrojaría el mejor resultado de predicción del

rendimiento académico de los estudiantes cuando se este se estima por año de estudio del estudiante (cada dos semestres). Por tanto, el resumen de una de las métricas de evaluación por semestres es mostrado en la Figura 7. Un valor de métricas como la exactitud "accuracy" cercano a 1 indica que todas las predicciones son correctas. Un valor cercano a 0 sugiere un pésimo modelo de predicción pésimo. Cabe mencionar que el algoritmo Stacking proporciona mejor métrica de rendimiento. Los mejores resultados se consiguen mediante la opción Stacking. La diferencia con el Blending radica en el tipo de metamodelo, ya que no debe usa un modelo lineal como en el caso del Blending sino un metamodelo no lineal.

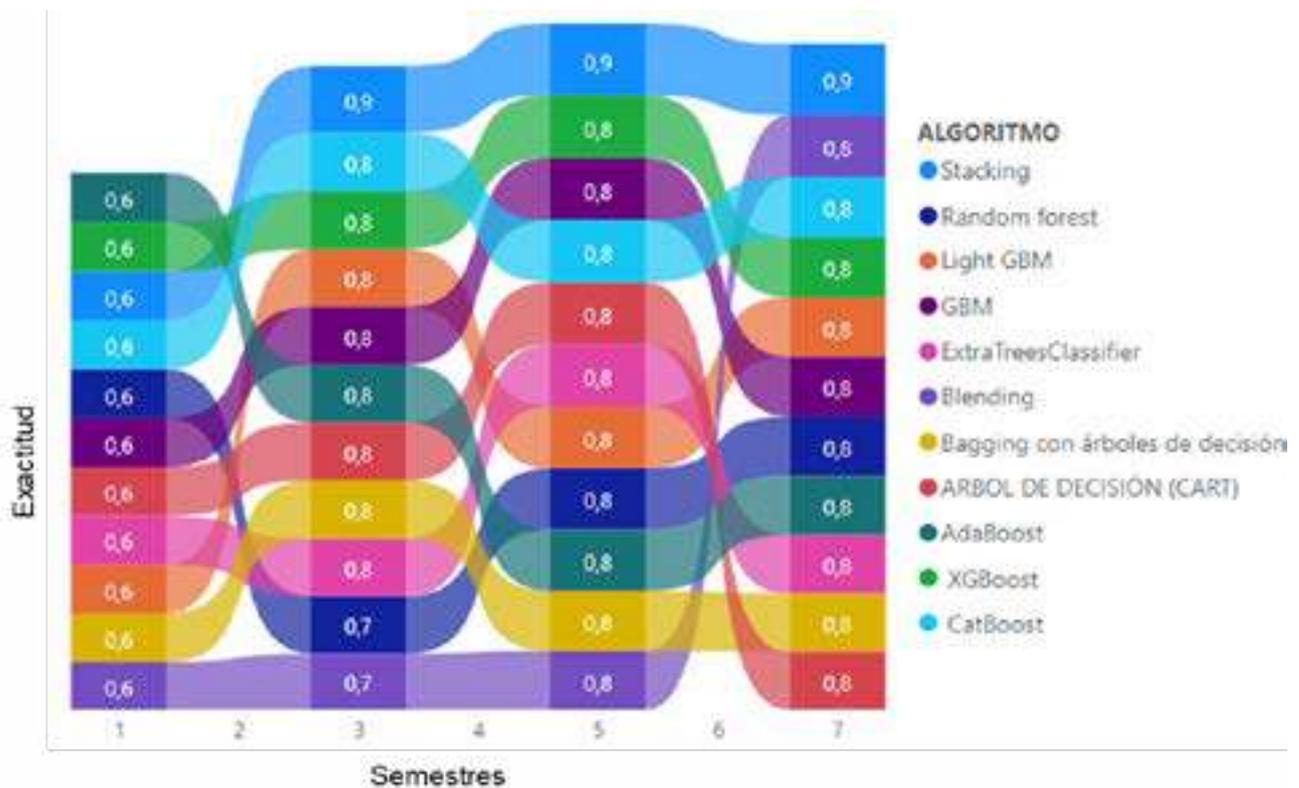


Figura 7. Resumen de resultados de mejores algoritmos de Machine Learning para predicción del rendimiento académico

## DISCUSIÓN

Variables como la edad de los estudiantes en estudios previos (Espinosa et al., 2020), (Awadalla et al., 2020), habían mostrado que incidían directa o indirectamente en el rendimiento académico de los estudiantes. Este hecho es contrario a lo que arrojaron los trece (13) métodos de selección de características aplicados en esta investigación ya que la variable edad no es una característica importante sobre la predicción del rendimiento académico universitario. Hecho que es similar a lo encontrado por (Bonsaksen, 2016) but also for the competitive status of the institutions providing their education. The literature is sparse concerning the factors of importance for occupational therapy students' academic performance in, and satisfaction with, their education programme. Method: This cross-sectional study explored sociodemographic, relationship, education and

work-related variables and their associations with the students' academic performance and satisfaction with the education programme. The data were analysed with multivariate linear regression. Results: Participants were 123 students (mean age 24 years, 80% women). Así mismo, se muestra fuerte influencia de la variable localidad del colegio en el rendimiento del estudiante como lo manifiestan (Grob et al., 2015) y (Muñoz-Comonfort et al., 2014)) ya que el rendimiento académico se ve afectado por el tipo de colegio en el que se estudió. De igual manera el transcurrir académico del estudiante y sus resultados muestra influencia en el rendimiento académico a través de las notas de algunas asignaturas mostradas anteriormente.

Al utilizar algoritmo supervisados para predecir el rendimiento académico el resultado promedio de métricas como la exactitud es del 74% para el mejor resultado conseguido mediante

el algoritmo árbol de decisión, el cual es superior a la exactitud estimada por (Adekitan & Noma-Osaghae, 2018) que fue de 49,078% de exactitud. Se debe posiblemente al no uso eficiente de los métodos de transformación, selección de características y optimización. La predicción del rendimiento al finalizar el primer año de carrera son prometedores cuando se trabajan con métodos de ensamble (80% de exactitud) utilizando las variables establecidas por (Yamao et al., 2018) debido a que su estudio implementó como método de selección de variables la regresión lineal.

Los resultados de los métodos de ensamble son superiores a los de los supervisados. Mediante métodos Bagging la exactitud es alrededor del 74.5% en promedio y con métodos Boosting 75% en promedio. Pero son aún superiores los resultados alcanzados con las variantes del método voting, especialmente mediante la opción Stacking (79% en promedio). De tal manera que las combinaciones de múltiples clasificadores utilizadas en los grupos de algoritmos Bagging, Boosting y voting ayudan a disminuir el sesgo y varianza que pueden presentarse cuando se emplean clasificadores débiles o únicos como árbol de decisión, KNN y SVC para analizar un conjunto de datos.

Intentar predecir el rendimiento de semestres futuros fue realizado por (Adekitan & Salau, 2019) aunque su mejor resultado fue mediante el algoritmo AdaBoost (86%) sigue siendo inferior a los porcentajes alcanzados en esta investigación, aun tomado en cuenta características académicas, demográficas y conductuales.

A diferencia del uso del método Voting de forma general en el que se usa el promedio simple o el promedio ponderado (en el mejor de los casos) para combinar las predicciones de los modelos base, el Blending y el Stacking utilizan metamodelos (que pueden ser lineales o no) para producir sus predicciones de ahí la mejora

en las métricas de rendimiento de los algoritmos al intentar predecir el rendimiento académico universitario para diferentes periodos de tiempo.

Este trabajo es un paso en la búsqueda de un modelo que permita por medio de analítica de datos establecer una de los indicadores de la educación superior como lo es el rendimiento académico universitario, con el cual se pretende hacer uso al máximo de los recursos técnico – digitales que permitan no solo el almacenamiento de la información creada por los autores del proceso educativo sino que esta pueda traducirse en información analizada que permita a educadores e instituciones tomar acciones acertadas sobre la vida académica del estudiante.

## CONCLUSIONES

De acuerdo con los resultados obtenidos, se pueden plantear las siguientes conclusiones principales:

- 1.- Los algoritmos de ensamble requieren información relevante de aquellas variables que los métodos de selección de características consideran que son las más influyentes en la variable de salida (rendimiento académico). De esta manera ofrecen mejores valores de métricas de rendimiento.
- 2.- Los algoritmos de ensamble se basan en establecer diferentes métodos que trabajan en conjunto con el fin de reducir los errores que proporcionarían cada uno de los algoritmos supervisados débiles.
- 3.- En el proceso de optimización de un modelo suele recurrir a hiperparámetros que controlan el proceso de aprendizaje del algoritmo. Aspecto que nos solo resulta interesante para los algoritmos débiles, sino que facilita la forma de entrenar a los algoritmos de ensamble.

## REFERENCIAS BIBLIOGRÁFICAS

- Acharya, A., & Sinha, D. (2014). Early Prediction of Students Performance using Machine Learning Techniques. *International Journal of Computer Applications*, 107(1), 37–43. <https://doi.org/10.5120/18717-9939>
- Adekitan, A. I., & Noma-Osaghae, E. (2018). Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technologies*, 24, 1527–1543. <https://doi.org/10.1007/s10639-018-9839-7>
- Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5(2), e01250. <https://doi.org/10.1016/j.heliyon.2019.e01250>
- Alyahyan, E., & Düştögör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(3). <https://doi.org/10.1186/s41239-020-0177-7>
- Anderson, H., Afshan, B., & Baker, R. (2019). *Predicting Graduation at a Public R1 University Predicting Graduation at a Public R1 University*. 2012(February), 1–4.
- Awadalla, S., Davies, E. B., & Glazebrook, C. (2020). A longitudinal cohort study to explore the relationship between depression, anxiety and academic performance among Emirati university students. *BMC Psychiatry*, 20(448). <https://bmcp psychiatry.biomedcentral.com/track/pdf/10.1186/s12888-020-02854-z.pdf>
- Bhutto, S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020). Predicting Students' Academic Performance Through Supervised Machine Learning. *ICISCT 2020 - 2nd International Conference on Information Science and Communication Technology*. <https://doi.org/10.1109/ICISCT49550.2020.9080033>
- Bonsaksen, T. (2016). Predictors of academic performance and education programme satisfaction in occupational therapy students. *British Journal of Occupational Therapy*, 79(6). <https://doi.org/10.1177/0308022615627174>
- Bourel, M. (2012). *Model aggregation methods and applications*. 10, 19–32.
- Campo-Ávila, D., Ramos-Jimenez, G. P., Morales-Bueno, R., & Baena-García, M. (2018). *Minería de datos educativos para la predicción personalizada del rendimiento académico*.
- 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*.
- Castrillón, O., Sarache, W., & Ruiz, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Revista Formación Universitaria*, 13(1), 93–102. <https://doi.org/10.4067/S0718-50062020000100093>
- Céspedes, R. C., Vara-Horna, A., Lopez-Odar, D., Diaz-Rosillo, A., & Asencios-Gonzalez, Z. (2018). Ausentismo, presentismo y rendimiento académico en estudiantes de universidades peruanas. *Propósitos y Representaciones*, 6(1), 83–133. <https://doi.org/10.20511/pyr2018.v6n1.177>
- Contreras, L., Fuentes, H., & Molano, J. (2021). Analítica académica: nuevas herramientas aplicadas a la educación. *Revista Boletín Redipe*, 10(3), 137–158.

- Contreras, L., Fuentes, H., & Rodríguez, J. (2020). Application of automatic learning as a prediction strategy for academic dropout in universities. *Sylwan Journal*, 164(6). <http://sylwan.ibles.org/archive.php?v=164&i=6>
- Contreras, L., & López, I. (2020). Academic Performance Prediction in Universities using Ensemble Algorithms: A Literature Review. *International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)*, 10(5), 797–810. [http://www.tjprc.org/view\\_paper.php?id=14682](http://www.tjprc.org/view_paper.php?id=14682)
- Costa, E., Fonseca, B., Almeida, M., & Ferreira, F. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73, 247–256. <https://doi.org/10.1016/J.CHB.2017.01.047>
- De La Hoz, E. J., De La Hoz, E. J., & Fontalvo, T. J. (2019). Methodology of Machine Learning for the classification and Prediction of users in Virtual Education Environments. *Informacion Tecnologica*, 30(1), 247–254. <https://doi.org/10.4067/S0718-07642019000100247>
- Durán, C., & Rosado, A. (2019). La comprensión lectora y el rendimiento académico en estudiantes de Ingeniería. *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, 1(33). <https://doi.org/10.24054/16927257.v33.n33.2019.3317>
- Espinosa, J., Hernández, J., Rodríguez, J., Chacín, M., & Bermúdez, V. (2020). Influencia del estrés sobre el rendimiento académico. *AVFT-Archivos Venezolanos de Farmacología y Terapéutica*, 39(1). <https://bonga.unisimon.edu.co/bitstream/handle/20.500.12442/6322/PDF.pdf;jsessionid=BCAE649A5ED0968F81C3D9B047C8039E?sequence=1>
- Ferreira, M., Botero, J., Haimovich, P., & Urzúa, S. (2017). *Momento decisivo La educación superior en América Latina y el Caribe*. <https://openknowledge.worldbank.org/bitstream/handle/10986/26489/211014ovSP.pdf?sequence=5&isAllowed=y>
- Garbano, & María, G. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación*, 31(1), 43–63. <http://www.redalyc.org/articulo.oa?id=44031103>
- García, G. (2014). *Modelo de Machine Learning para la Clasificación de pacientes en términos del nivel asistencial requerido en una urgencia pediátrica con Área de Cuidados Mínimos*. 103.
- García, J., Sánchez, P., Orozco, M., & Obredor, S. (2019). Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia Knowledge Capture for the Prediction and Analysis of Results of the Quality Test of Higher Education in Colombia. *Revista Formación Universitaria*, 12(4), 55–62. <https://doi.org/10.4067/S0718-50062019000400055>
- Gareth, J. (2013). *An introduction to statistical learning : with applications in R* (Springer (ed.); 1st ed., Vol. 1). Springer.
- Grob, M., Becerra, D., Rodríguez, A., Cristiane, J., Ramirez, V., & Sabag, N. (2015). Relación entre Puntaje de Prueba de Selección Universitaria y Nota Enseñanza Media, y el Rendimiento Académico de la Asignatura de Morfología en Alumnos

de Primer Año de Odontología de la Universidad de Los Andes. *International Journal of Morphology*, 33(2), 527–531.

- Guizado, G., Valenzuela, M., & Vallejo, P. (2020). Desempeño docente y el rendimiento académico de los estudiantes de la Facultad de Tecnología en la Universidad Nacional de Educación de Perú. *Revista Conrado*, 16(72). <https://orcid.org/0000-0002-7852-458X>
- Guleria, P., & Sood, M. (2018). Predictive data modeling: Educational data classification and comparative analysis of classifiers using python. *PDGC 2018 - 2018 5th International Conference on Parallel, Distributed and Grid Computing*, 74 Guleria, P., Sood, M. (2018). Predictive data. <https://doi.org/10.1109/PDGC.2018.8745727>
- Hernández, C. (2016). Diagnóstico del rendimiento académico de estudiantes de una escuela de educación superior en México. *Revista Complutense de Educación*, 27(3), 1369–1388. <https://revistas.ucm.es/index.php/RCED/article/view/48551/48839>
- Jahangiri, A., & Rakha, H. A. (2015). Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data. *IEEE Transactions on Intelligent Transportation Systems*, 16(5), 2406–2417. <https://doi.org/10.1109/TITS.2015.2405759>
- Jalota, C., & Agrawal, R. (2019). Analysis of Educational Data Mining using Classification. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Perspectives and Prospects, COMITCon 2019*, 243–247. <https://doi.org/10.1109/COMITCon.2019.8862214>
- Joshika, P. and Rajeshwari. (2019). *STUDENTS' PERFORMANCE ANALYSIS USING SIMPLE K-MEANS*. 21(14), 990–995.
- Kaunang, F. J., & Rotikan, R. (2018). Students' academic performance prediction using data mining. *Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018*, 1–10. <https://doi.org/10.1109/IAC.2018.8780547>
- Kostopoulos, G., Kotsiantis, S., Pierrakeas, C., Koutsonikos, G., & Gravvanis, G. A. (2018). Forecasting students' success in an open university. *International Journal of Learning Technology*, 13(1), 26–43. <https://doi.org/10.1504/IJLT.2018.091630>
- Kumar, V. Krishna, A. Neelakanteswara, P. Basha, C. (2020). Advanced Prediction of Performance of a Student in an University using Machine Learning Techniques. *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, 121–126. <https://doi.org/10.1109/ICESC48915.2020.9155557>
- Lamas, H. (2015). Sobre el rendimiento escolar. *Prósitos y Representaciones: Revista de Psicología Educativa*, 3(1), 313–386.
- Lenskiy, A., Shariat, R., & Seol, S. (2020). The effect of academic breaks on undergraduate academic performance. *The International Journal of Electrical Engineering & Education*, 0(0), 1–12. <https://doi.org/10.1177/0020720920922518>
- Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *Anales de Psicología*, 30(3), 1151–1169. <https://doi.org/10.6018/analesps.30.3.199361>

- López-Aguado, M., & Gutiérrez-Provecho, L. (2019). Cómo realizar e interpretar un análisis factorial exploratorio utilizando SPSS. *REIRE Revista d'Innovació i Recerca En Educació*, 12(2), 1–14. <https://doi.org/10.1344/reire2019.12.227057>
- Martínez-Rodríguez, R. A., Álvarez-Xochihua, O., Mejía Victoria, O. D., Jordan Aramburo, A., & González Fraga, J. A. (2019). Use of Machine Learning to Measure the Influence of Behavioral and Personality Factors on Academic Performance of Higher Education Students. *IEEE Latin America Transactions*, 17(4), 633–641. <https://doi.org/10.1109/TLA.2019.8891928>
- Mavrou, I. (2015). Análisis factorial exploratorio: Cuestiones conceptuales y metodológicas. *Revista Nebrija*, 19, 71–80. <https://www.nebrija.com/revista-linguistica/analisis-factorial-exploratorio.html>
- Mengash, H. A. (2020a). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, 8(1), 55462–55470. <https://doi.org/10.1109/ACCESS.2020.2981905>
- Mengash, H. A. (2020b). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, 8, 55462–55470. <https://doi.org/10.1109/ACCESS.2020.2981905>
- Minichil, W., Eskindir, E., Demilew, D., & Mirkena, Y. (2020). Magnitude of premenstrual dysphoric disorder and its correlation with academic performance among female medical and health science students at University of Gondar, Ethiopia, 2019: a cross-sectional study. *BMJ Open*, 10(e034166). <https://doi.org/10.1136/bmjopen-2019-034166>
- Montero, E., Villalobos, J., & Valverde, A. (2007). Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico en la Universidad de Costa Rica: un análisis multinivel. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, 13(2), 215–234. [www.uv.es/RELIEVE/v13n2/RELIEVEv13n2\\_5.htm](http://www.uv.es/RELIEVE/v13n2/RELIEVEv13n2_5.htm)[www.uv.es/RELIEVE](http://www.uv.es/RELIEVE)pag.215
- Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2020). Student Engagement Level in an e-Learning Environment: Clustering Using K-means. *American Journal of Distance Education*, 34(2), 137–156. <https://doi.org/10.1080/08923647.2020.1696140>
- Muñoz-Comonfort, A., Leenen, I., & der Goes, T. I. F. (2014). Correlación entre la evaluación diagnóstica y el rendimiento académico de los estudiantes de medicina. *Investigación En Educación Médica*, 3(10), 85–91. <https://www.sciencedirect.com/science/article/pii/S2007505714727310>
- Murnion, P., & Helfert, M. (2013). Academic Analytics in quality assurance using organisational analytical capabilities A User-level Usage Analytics in Cloud Based Applications View project Insight View project. In U. Oxford (Ed.), *Annual Conference of the UK Academy of Information Systems (UKAIS)*. <https://doi.org/10.13140/2.1.3368.1600>
- Nieto, Y., García, V., Montenegro, C., González, C., & González, R. (2019). Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions. *IEEE Access*, 7, 75007–75017. <https://doi.org/10.1109/ACCESS.2019.2919343>
- Ochoa, L. L., Rosas Paredes, K., & Baluarte Araya, C. (2017). Evaluación de técnicas de minería de datos para la predicción

del rendimiento académico. *Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology, 2017-July*(January). <https://doi.org/10.18687/LACCEI2017.1.1.368>

Orihuela Maita, G. Y. (2019). Aplicación de Data Science para la Predicción del Rendimiento Académico de los Estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú. *Universidad Nacional Del Centro de Perú*, 114.

Patacsil, F. F. (2020). Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models. *Universal Journal of Educational Research*, 8(9), 4036–4047. <https://doi.org/10.13189/ujer.2020.080929>

Rivera, E. E., Becerra, S. C., Cotrina, A. R., & Acero, A. C. (2020). Empatía y rendimiento académico en estudiantes universitarios. *Educare*, 24(2), 26. <https://revistas.investigacion-upelipb.com/index.php/educare/article/view/1319/1289>

Rodriguez, M., & Ruíz, M. (2009). Indicadores de rendimiento de estudiantes universitarios: calificaciones versus créditos acumulados. *Revista de Educación*, 355, 467–492. [http://www.revistaeducacion.educacion.es/re355/re355\\_20.pdf](http://www.revistaeducacion.educacion.es/re355/re355_20.pdf)

Sajjadi, S. Shapiro, B. Mckinlay, C. Sarkisyan, A. Shubin, C., & Osoba, E. (2018). Finding bottlenecks: Predicting student attrition with unsupervised classifier. *2017 Intelligent Systems Conference, IntelliSys 2017, 2018-Janua*, 1166–1172. <https://doi.org/10.1109/IntelliSys.2017.8324279>

Santosh, K. C. (2020). AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal

Data. *Journal of Medical Systems*, 44(5), 1–5. <https://doi.org/10.1007/s10916-020-01562-1>

Santoso, L. W., & Yulia. (2019). The Analysis of Student Performance Using Data Mining. In *Advances in Intelligent Systems and Computing* (Vol. 924). Springer Singapore. [https://doi.org/10.1007/978-981-13-6861-5\\_48](https://doi.org/10.1007/978-981-13-6861-5_48)

Sweeney, M., Rangwala, H., Lester, J., & Johri, A. (2016). *Next-Term Student Performance Prediction: A Recommender Systems Approach*. 1–27. <https://doi.org/10.5281/zenodo.3554603>

T.Velmurugan, & Anuradha, C. (2016). Performance Evaluation of Feature Selection Algorithms in Educational Data Mining. *International Journal of Data Mining Techniques and Applications*, 5, 131–140. <http://www.hindex.org/2016/article.php?page=1176>

Vega García, J. F. (2019). *Modelo de pronóstico de rendimiento académico de alumnos en los cursos del programa de estudios básicos de la Universidad Ricardo Palma usando algoritmos de Machine Learning*.

Viloria, A., García Guliany, J., Niebles Núñez, W., Palma, H. H., & Niebles Núñez, L. (2020). Data Mining Applied in School Dropout Prediction. *Journal of Physics*, 1432, 12092. <https://doi.org/10.1088/1742-6596/1432/1/012092>

Yamao, E., Saavedra, L., Campos, R., & Huancas, V. (2018). Prediction of academic performance using data mining in first year students of peruvian university. *Revista USMP - Campus*, 23(26), 151–160.

Zaffar, M., Hashmani, M. A., Savita, K. S., & Rizvi, S. S. H. (2018). A Study of Feature Selection Algorithms for Predicting Students

Academic Performance. *International Journal of Advanced Computer Science and Applications*, 9(5), 541–549. <https://doi.org/10.14569/IJACSA.2018.090569>

Zárate, E., Lavado, B., & Pomahuacre, W. (2020). Competencia comunicativa intercultural y rendimiento académico en lenguas extranjeras. *Revista Conrado*, 16(74). <https://orcid.org/0000-0002-2924-6771>

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: methods and applications* (Springer (ed.); 2nd ed.).

Zhi-Hua Zhou. (2012). *Ensemble learning: foundations and algorithms* (1st ed., Vol. 1). Chapman & Hall/CRC.