Formación Universitaria
Vol. 13(5), 233-246 (2020)
http://dx.doi.org/10.4067/S0718-50062020000500233

Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático

Leonardo E. Contreras[1], Héctor J. Fuentes[2] y José I. Rodríguez[3]
(1) Facultad de ingeniería, Grupo de investigación DIMSI, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. (correo-e: lecontrerasb@udistrital.edu.co)
(2) Facultad de ingeniería, Grupo de investigación DIMSI, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. (correo-e: hjfuentesl@udistrital.edu.co)
(3) Facultad de ingeniería, Grupo de investigación GICOECOL, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. (correo-e: jirodriguezm@udistrital.edu.co)

Recibido Mar. 5, 2020; Aceptado May. 6, 2020; Versión final May. 25, 2020; Publicado Oct. 2020

Resumen

Esta propuesta plantea la selección de variables que influyen en la predicción del rendimiento en estudiantes de ingeniería industrial de la Universidad Distrital (Colombia) por diferentes metodologías: filtro, envoltura e integrados. Se implementaron algoritmos de clasificación a través del lenguaje de programación Python como árbol de decisión, K vecinos más cercanos, perceptrón y otros, los cuales son comparados para conocer el mejor resultado de predicción. El género y el puntaje ICFES (examen de estado en Colombia) para condición matemática se encuentran en el rango superior de todos los métodos de selección de características, y el algoritmo perceptrón arroja mejor exactitud con respecto a los otros algoritmos usados. Se concluye que las variables que más influyen en el rendimiento académico de los estudiantes de ingeniería son: edad, género, puntaje ICFES para aptitud matemática, puntaje global ICFES, valor de matrícula y puntaje ICFES para condición matemática y cohorte.

Palabras clave: análisis de datos; aprendizaje automático; educación en ingeniería; modelo; rendimiento académico

Academic performance prediction by machine learning as a success/failure indicator for engineering students

Abstract

This research study identifies variables that influence the prediction of performance in industrial engineering undergraduate students at the Universidad Distrital (Colombia) by three methodologies: filter, wrappers, and integrated. Python programming language classification algorithms such as decision tree, K nearest neighbors, and perceptron are implemented and they are compared to obtain the best prediction results. The results show that gender and the ICFES Score (Colombian nation-wide university admission exam) for mathematics were in the upper range in all the selection methods. The Perceptron algorithm is the most accurate of all the algorithms tested. It is concluded that the variables that most affect academic performance in engineering students are: age, gender, tuition fee, the overall ICFES score, and the ICFES scores for mathematical aptitude and cohort mathematics.

Keywords: analytics of data; machine learning; engineering education; model; academic performance

INTRODUCCIÓN

Hoy en día se están presentando diversos cambios en áreas como la medicina, la economía, y el comercio, entre otros, como producto de la incursión de la analítica de datos en ellos. Y tanto ha sido el auge de la analítica que ha permeado el campo de la educación, donde se empieza a procesar un gran volumen datos que contienen la información relacionada con los actores del proceso educativo. Y es aquí donde la ingeniera puede jugar un papel importante al dar su aporte para solucionar múltiples aspectos de índole académico - administrativo tales como mejorar el aprendizaje, la deserción, el abandono y el rendimiento académico.

Existen diversos factores que permiten medir la eficiencia del proceso educativo, muchos de ellos de carácter multidimensional, como es el caso del fenómeno de retención, definido como la diferencia entre el número de estudiantes que ingresan en primer semestre y los graduados por año (Salcedo, 2010); y del rendimiento académico definido como el principal indicador de éxito o fracaso del estudiante. Por tanto, su determinación, ha generado controversia, ya que no existe una teoría definitiva acerca de una metodología para su medición o un indicador para su valoración. Al ser multidimensional, el rendimiento académico depende de múltiples aspectos tales como los objetivos del docente, de la institución, del estudiante, etc., así mismo requiere realizar una integración de las diferentes técnicas y metodologías con el propósito de predecirlo (Khan y Choi, 2014). Existen otros indicadores como: Tasa de abandono (proporción de estudiantes que estando matriculados dos semestres atrás son clasificados como desertores un año después), Tasa de deserción por cohorte (contabiliza la deserción acumulada en cada semestre para un grupo de estudiantes que ingresaron a primer curso en un mismo periodo académico), Tasa de graduación (contabiliza el número total de graduados, sobre el total acumulado de los estudiantes que ingresaron a primer curso), y la Tasa de retención. Estos indicadores para nuestro país no son los mejores según el MEN - Ministerio de Educación de Colombia, al año 2017 (figura 1)

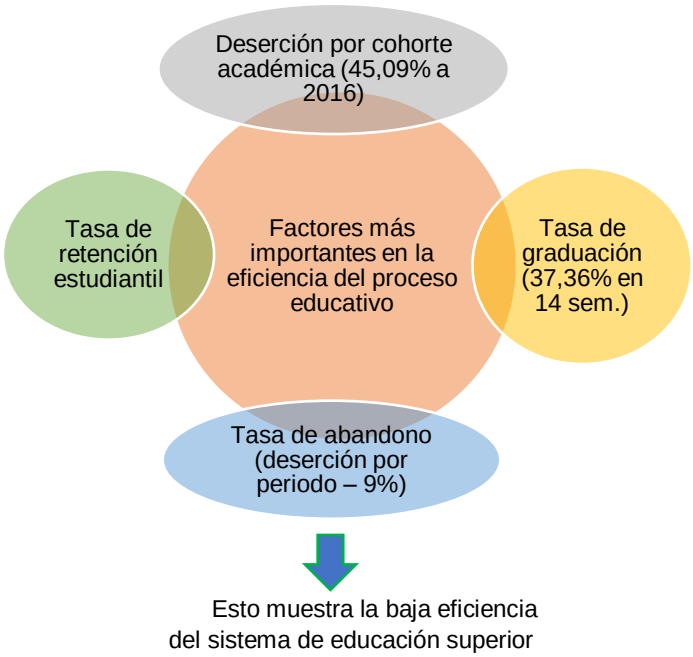


Fig. 1: Estadísticas de deserción y graduación año 2017. (Datos tomados de MEN, 2017)

La incursión de la tecnología en educación para maximizar la experiencia del aprendizaje y aspectos relacionados, ha dado pie para el establecimiento del aprendizaje mejorado por tecnologías (TEL - Technology Enhanced Learning) y dentro de este, se han acotado diversos conceptos relacionados a la analítica en el campo de la educación como son la minería de datos para la educación (EDM - Educational Data Mining), la analítica académica (AA - Academic Analytics), y la analítica del aprendizaje (LA - Learning Analytics) (Ayesha et al., 2010). Conceptos que van de la mano con el aprendizaje automático. Estos conceptos, convierten datos educativos en información útil que permite tomar acciones previas o fomentar la enseñanza y el aprendizaje (figura 2).

La Analítica Académica es usada por las instituciones para monitorear el progreso de los objetivos institucionales, eficiencia al terminar las carreras, impacto de la difusión, etc. La razón radica en que las instituciones educativas tienen que rendir cuentas ante organismos estatales que evalúan la eficiencia de la educación, por lo que están aplicando este tipo de análisis que les ha permitido cambiar su forma de tomar

decisiones (García, 2019). Finalmente, la Analítica del aprendizaje ayuda a personalizar la experiencia del estudiante, predecir tasas de deserción y abandono, y el diseño de acciones para mejorar el aprendizaje (Lonn et al., 2015). La analítica académica contiene a la analítica del aprendizaje.

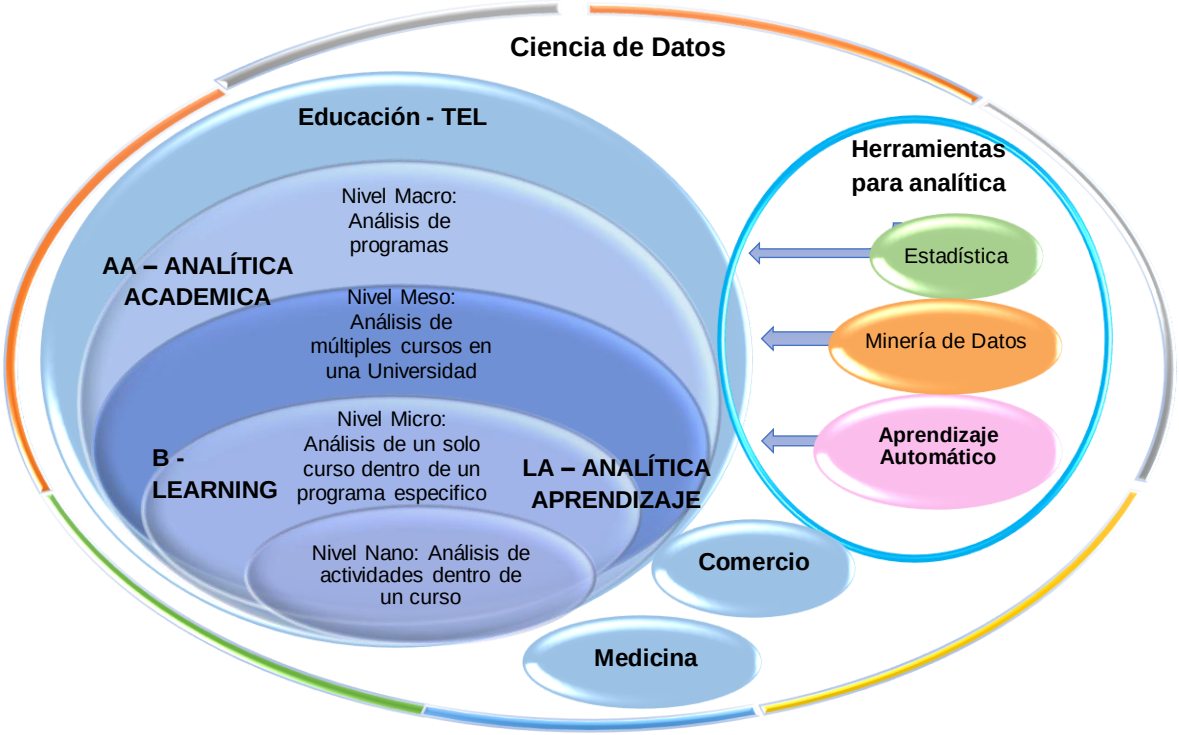


Fig. 2: Panorama de la Analítica en educación (TEL)

El Aprendizaje automático o Machine Learning (ML) aplicado a la educación puede definirse como la recolección, análisis y divulgación de datos sobre los actores educativos, con el propósito de comprender y optimizar aspectos relacionados del proceso enseñanza- aprendizaje (Dyckhoff et al, 2012). Se centra en herramientas y métodos para la exploración de datos provenientes de contextos educativos, lo que las hace considerar hoy en día como una de las técnicas que ayudarán a moldear el futuro de la educación superior. Cabe recordar que con estas herramientas no existe una única forma de resolver un problema, debido a que existen diferentes algoritmos que se pueden ajustar a los datos y el trabajo del investigador consiste en seleccionar el más adecuado teniendo en cuenta las métricas de evaluación. Este estudio tiene por objetivo conocer el mejor escenario para predecir el rendimiento de los estudiantes de Ingeniería Industrial de la Universidad Distrital utilizando diferentes métodos que permiten seleccionar cuales de las 30 variables iniciales, son las más relevantes en su determinación, e implementar modelos por medio de algoritmos de clasificación (árbol de decisión, KNN, SVC y Perceptrón). Los resultados a pesar de no considerar todas las variables que influyen según la literatura en el fenómeno son bastantes acertados: con buena precisión y métrica de los modelos planteados.

OTROS ANTECEDENTES

Existen diferentes definiciones sobre el rendimiento académico. Algunos como (Garbanzo y María, 2007) conceptúan que el rendimiento académico es la suma de diferentes y complejos factores que actúan en la persona que aprende. (Tourón, 1985) afirma que el rendimiento académico es un resultado del aprendizaje, suscitado por la intervención pedagógica del profesor o profesora y producido en el alumno; (Rojas,2013) conceptúa que son una serie de factores que giran alrededor de los resultados finales del esfuerzo hecho por el estudiante; y (García, 2019), lo considera como el principal indicador de éxito o fracaso del estudiante, por tal motivo ha sido considerado como uno de los aspectos importantes a la hora de analizar resultados sobre el proceso de enseñanza-aprendizaje.

Rendimiento académico

Otro aspecto, es el hecho de cómo medirlo. (Page et al, 1990) manifiestan que es el resultado aritmético de aprobar una asignatura, contrario piensa (Escudero, 1999) ya que, según él, las calificaciones son una medida de los resultados de la enseñanza, pero no estrictamente de su calidad. Otros autores expresan que las pruebas objetivas es el medio adecuado para determinarlo debido a que las respuestas son cortas y precisas, sin la influencia subjetiva del profesor. Una tercera forma es según el número de asignaturas aprobadas ya

que el número de asignaturas aprobadas por año es un indicador de rendimiento estudiantil más adecuado que el promedio (Porto y Gresia, 2005). Por último (Noel et al., 2011) manifiestan que debería medirse según la cantidad de créditos acumulados, ya que esto permite hacer una comparación entre los créditos acumulados por el alumno durante cierto tiempo de estudio y los créditos que, de acuerdo al plan de estudios debió acumular en el tiempo programado (Noel et al., 2011).

Aprendizaje automático

Se han utilizado diversos algoritmos de ML en el campo de la educación: de aprendizaje supervisado y de aprendizaje no supervisado. Los primeros son aquellos que aprenden a partir de datos de ejemplos y respuestas de destino, para luego predecir la respuesta correcta de un dato completamente nuevo. Ejemplos de este tipo de algoritmos son: Naive Bayes, Árbol de decisión, Regresión logística, K Vecinos cercanos, Máquina de vectores de soporte, etc. Los segundos, son aquellos que aprenden de ejemplos simples sin ninguna respuesta asociada, dejando que el algoritmo determine los patrones de datos por sí mismo; es decir, organiza los datos según algunos grupos de características. Cuando se tiene un dato nuevo, este tipo de algoritmo lo agrupará en un conjunto de datos que tenga las características más parecidas al dato inicial suministrado (no es una predicción).

Analítica en Educación

Algunas investigaciones relevantes publicadas en revistas de ámbito estadístico y educativo de alto impacto (palmer y Stuart, 2013), (García,2019), (Oblinger et al.,2007), así como trabajos relacionados con el campo del Machine Learning son mostrados en la figura 3. Los trabajos en el área estadística inician con el método de los mínimos cuadrados, pasando por métodos de regresión, hasta hoy en día en los cuales se han estado utilizando métodos multivariantes (multinivel y análisis logístico bivariante). La razón por la cual se usan estos métodos según (García, 2019) es porque los investigadores se han dado cuenta que los datos dentro del campo de la educación se encuentran anidados, es decir, un estudiante pertenece a un curso; este curso pertenece a un área; esta área pertenece a un proyecto curricular; este proyecto curricular pertenece a una facultad; y esa información es difícilmente separable. También se muestran algunos algoritmos de Machine Learning y las variables que se han utilizado con frecuencia para calcular el rendimiento académico (Fernandes et al., 2019), (Burgos et al., 2019), (Nithya et al., 2016), (Ramesh et al., 2013), (Osmanbegovic y Sulji´c, 2012), donde el promedio acumulado de materias es el atributo principal para predecir el rendimiento de los estudiantes, seguido de la evaluación interna (quiz, trabajos de clases, etc.);por variables demográficas y socio-económicas.

Predecir el rendimiento de los estudiantes se vuelve más desafiante debido al gran volumen de datos, así como a la gran cantidad de variables que según la literatura podrían influir en su determinación (alrededor de 105). 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, implementando diversas herramientas tecnológicas con el fin de predecir y/o favorecer por medio de herramientas de aprendizaje automático la toma de decisiones en el campo educativo tanto para estudiantes como para docentes. Este trabajo es un aporte en ese camino de búsqueda.

MÈTODO

Plantear una metodología de solución al problema resulta ser tan complejo como el hecho de determinar cuáles serían las variables que puedan afectar su determinación. La metodología empleada en el estudio se expone de manera resumida en los siguientes 7 pasos: (1) Participantes; (2) Variables; (3) Tratamiento de datos; (4) Estadísticas (5) Selección de Características; (6) Algoritmos de predicción; y (7) Métricas de evaluación.

Participantes

Este trabajo se enmarca dentro del campo de la analítica aplicada a la educación. El objetivo consiste en plantear un modelo mediante algoritmos de aprendizaje automático usando lenguaje de programación Python, con el fin de establecer la posibilidad de predicción del rendimiento académico de estudiantes de ingeniería Industrial de la Universidad Distrital (Colombia). Para ello, se tienen en cuenta los registros de un total de 1620 estudiantes del periodo 2008 - 2014.

Variables

El modelo Macro (futuro) a diseñar involucra diversas variables interrelacionadas para predecir el rendimiento académico entre otros aspectos de la vida académica del estudiante. Son variables que podrían agruparse

en diferentes niveles tales como: Factores académicos preuniversitarios, Factores demográficos preuniversitarios, Factores socio-culturales preuniversitarios, Factor socio-económicos preuniversitarios, Factores de Gestión Académica universitaria, Factores Tecnológicos, Factores de Biblioteca, Factores Institucionales, Factores pedagógicos, Factores Intelectuales y Factores afectivos (Castrillón, Sarache y Ruiz, 2020) (Contreras y Rodríguez, 2018). Las variables que se tuvieron en cuenta para iniciar el trabajo actual son mostradas en la tabla 1.

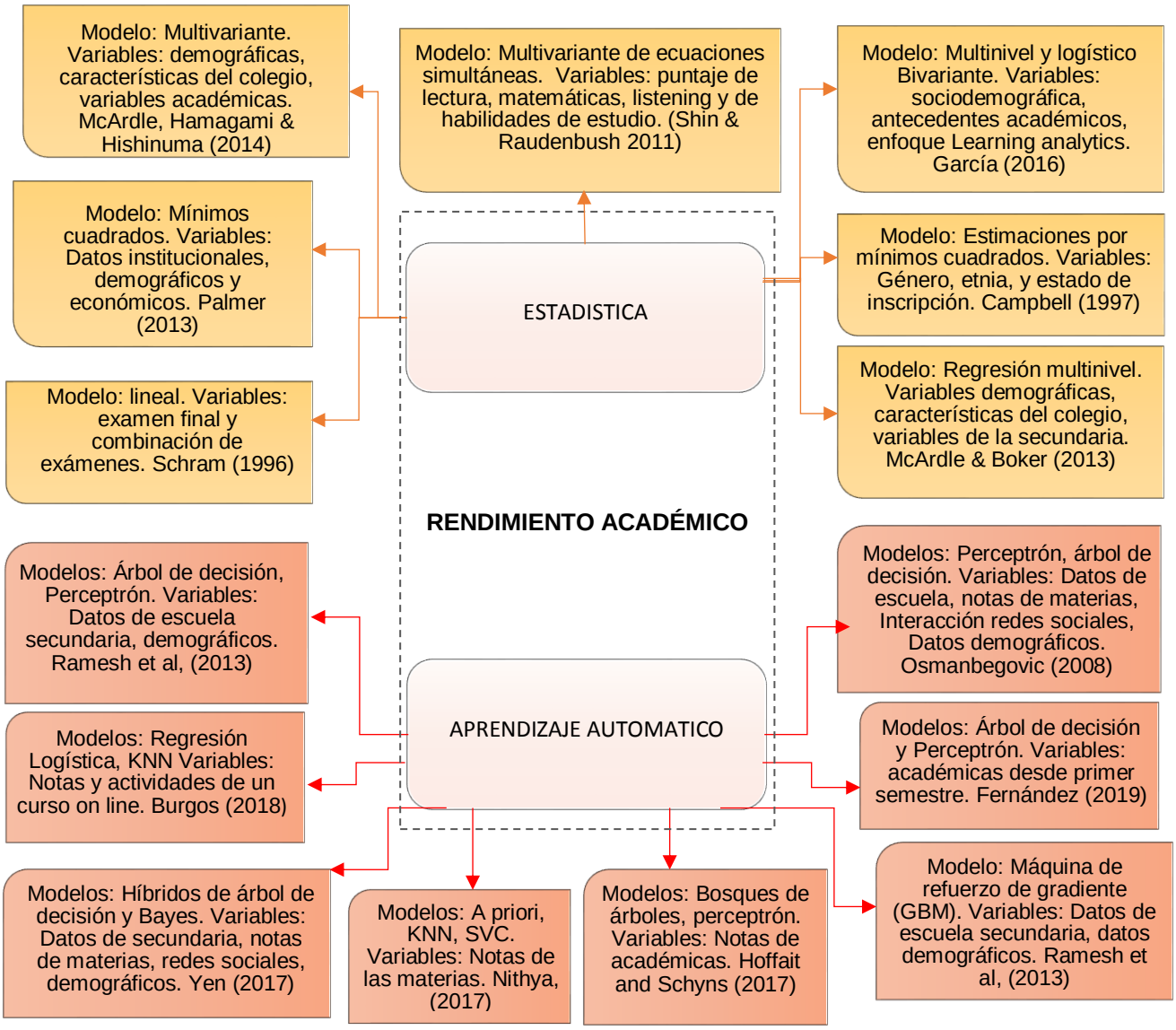


Fig. 3: Metodologías estadísticas, de Machine Learning y variables para la medición del rendimiento académico.

Tratamiento de datos

Inicialmente no se poseía la información o registros de los estudiantes a analizar. Solo se tenían muchos archivos .CSV (Comma Separated Values - Valores separados por coma) con información en cuanto a las variables a analizar. Haciendo uso de Microsoft Access (sistema de gestión de bases de datos) se importaron los diversos archivos .CSV para convertirlos en tablas de Acces. De esta manera se obtuvo una data de entrada para los modelos.

Como es posible que una variable independiente ejerza mayor influencia sobre la variable dependiente por el hecho de que su escala numérica es mayor que en otras variables, fue necesario realizar una normalización (tipo de transformación) del conjunto de datos, es decir, se busca que se eliminen los efectos de influencia, ya que son principalmente modificaciones sintácticas llevadas a cabo sobre datos sin que supongan un cambio para el algoritmo que se desea aplicar (García et al., 2019).

Tabla 1. Variables iniciales del conjunto de datos de los estudiantes

Factor	*Variables*
Factores académicos Pre-Universidad	Puntaje global ICFES – Colombia (PIT), Año de presentación de prueba ICFES (APPI), Puntaje ICFES para aptitud Matemática (PIAM), Puntaje ICFES para condición Matemática (PICM), Puntaje ICFES para área de Física (PIAF), Puntaje ICFES para área de Química (PIAQ), Puntaje ICFES para área de Lenguaje (PIAL), Puntaje ICFES para área de Biología (PIAB), Puntaje ICFES para área de Ciencias sociales (PIACS), Puntaje ICFES para área de Filosofía (PIAF), Puntaje ICFES área de Aptitud verbal (PIAV), Puntaje ICFES área de idiomas (PIAI)
Factores demográficos	Edad (ED), Estrato (E), Género (G), Estado civil (EC), Tipo de colegio donde culminó la secundaria (TC), Calendario del colegio (CC), Carácter del bachillerato (CB), Cohorte (C), Año de cohorte (AC), Código de localidad del colegio (CLC)
Factor socio-económicos	Tipo de ingreso a la Universidad (TIU), Nivel educativo de los padres (NEP), número de miembros de la familia (NMF), Valor de matrícula (VM), Desplazado (D), Minoría Étnica (ME), Mejores Bachilleres (MB), Numero de pensum (NM)

Estadísticas

Antes de proceder a la aplicación de métodos para analizar los datos, según diversos autores como (Santosh, 2020) es conveniente realizar algunas estadísticas descriptivas, lo que permitirá conocer más el marco de datos que se pretende trabajar. Una forma práctica de comprender algunas estadísticas y mejorar la comprensión de los datos antes de aplicar algoritmos de aprendizaje automático es por medio de gráficos individuales de las variables, entre variables independientes, así como entre estas y la variable de salida.

Selección de características

El objetivo principal de los métodos de selección de características es seleccionar un subconjunto de atributos de entrada, eliminando aquellas características dan menos información predictiva (Zaffar et al., 2018). Los métodos de selección de características pueden ayudar a identificar y eliminar atributos innecesarios, que no contribuyen a la precisión de un modelo. Estos métodos se subdividen en: métodos de filtro, métodos de envoltura y métodos integrados.

Algoritmos de Predicción

En este trabajo se tuvieron en cuenta diversos algoritmos de clasificación, algunos de los cuales se describen brevemente: i) Árbol de decisión: Es un algoritmo que permite crear un árbol invertido permitiendo dividir inicialmente los datos en dos conjuntos teniendo en cuenta el valor del diferenciador más significativa entre todas las variables de entrada. Cada uno de esos resultados crea nodos adicionales, que se ramifican en otras posibilidades. Esto le da una forma similar a la de un árbol. El número de condiciones que tenga el árbol dependerá de la cantidad de variables de entrada (independientes) de lo que se pretenda modelar. Con el fin de establecer cuál es la mejor partición del nodo se usan diferentes metaheurísticas las cuales buscan minimizar la entropía, la cual es una medida generalmente usada para determinar el grado de incertidumbre de una variable (Radhwan et al., 2017) ii) Máquinas de vectores de soporte (SVM): Este algoritmo permite la búsqueda de un hiperplano que modele la tendencia de los datos de entrenamiento y según ella predecir cualquier dato en el futuro. SVM ofrece un enfoque basado en principios matemáticos y en la teoría del aprendizaje estadístico. Funciones tradicionales del hiperplano para este algoritmo pueden ser: lineal, polinomial, Gauss, sigmoide y series de Fourier. iii) Red Neuronal (Perceptrón): Es un algoritmo que tiene la capacidad de detectar todas las posibles interacciones entre las variables predictoras. El modelo se basa en un conjunto de unidades neuronales simples (neuronas artificiales), de forma aproximadamente análoga al comportamiento observado en los axones de las neuronas en los cerebros biológicos. Este algoritmo es considerado debido a la robustez que posee la técnica para el manejo de datos, adaptabilidad y su reconocida capacidad de generalización (Sánchez y García, 2017). iv) K-Vecinos más cercanos (KNN): Es un algoritmo al cual se le debe proporcionar un hiper paramero K y de esta manera él puede predecir la clase dependiendo de los K datos cercanos, es decir, se predice a que grupo pertenece el dato en cuestión, dependiendo cual es la clase mayoritaria más cercana. Independiente del tipo de Kernel que se utilice en el algoritmo, este se enfoca a la búsqueda de los coeficientes de la función del hiperplano lo cual lo consigue mediante métodos de optimización (De la Hoz, De la Hoz y Fontalvo, 2019).

Métrica de evaluación

A los modelos se les debe evaluar la calidad de la clasificación. Según (Nieto et al.,2018), puede hacerse por cuatro métricas diferentes: Exactitud, precisión (especificidad), Recall (sensibilidad) y medición puntaje F1. Valores que se determinan a partir de la matriz de confusión (tabla 2)

Tabla 2. Matriz de confusión

		Predicción	
		Positivo	Negativo
Actual	Positivo	Verdaderos positivos (TP)	Falsos Negativos (FN)
	Negativo	Falsos positivos (FP)	Verdaderos Negativos (TN)

La exactitud se define como el número de instancias correctamente predichas sobre el número total de registros. La precisión es la relación de instancias positivas predichas correctamente a el total de casos positivos predichos. La sensibilidad se calcula como la relación el número de instancias predichas correctamente sobre el número total de positivos. El puntaje F1 es el promedio ponderado de precisión y sensibilidad. Por lo tanto, esta puntuación tiene en cuenta tanto los falsos positivos como los falsos negativos.

$$\text{Exactitud} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{FN} + \text{VN}} \quad (1)$$

$$\text{Precisión} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (2)$$

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (3)$$

$$\text{Puntaje F1} = \frac{2 * \text{Precision} * \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}} \quad (4)$$

RESULTADOS

Al aplicar la metodología descrita anteriormente se obtuvieron diversos resultados para los pasos (4) Estadísticas (5) Selección de Características; (6) Algoritmos de predicción; y (7) Métricas de evaluación.

En cuanto a estadísticas

A partir del conjunto de datos se realizaron diversas estadísticas descriptivas de las variables, con el fin de conocer el conjunto de datos que se desea analizar. Fueron realizados Histogramas y diagramas de bigotes como medio de representación visual para describir varias variables importantes, así como la dispersión y simetría. Ejemplo de algunas estadísticas obtenidas mediante las bibliotecas Pandas y Numpy de Python son las mostradas en la figura 4.

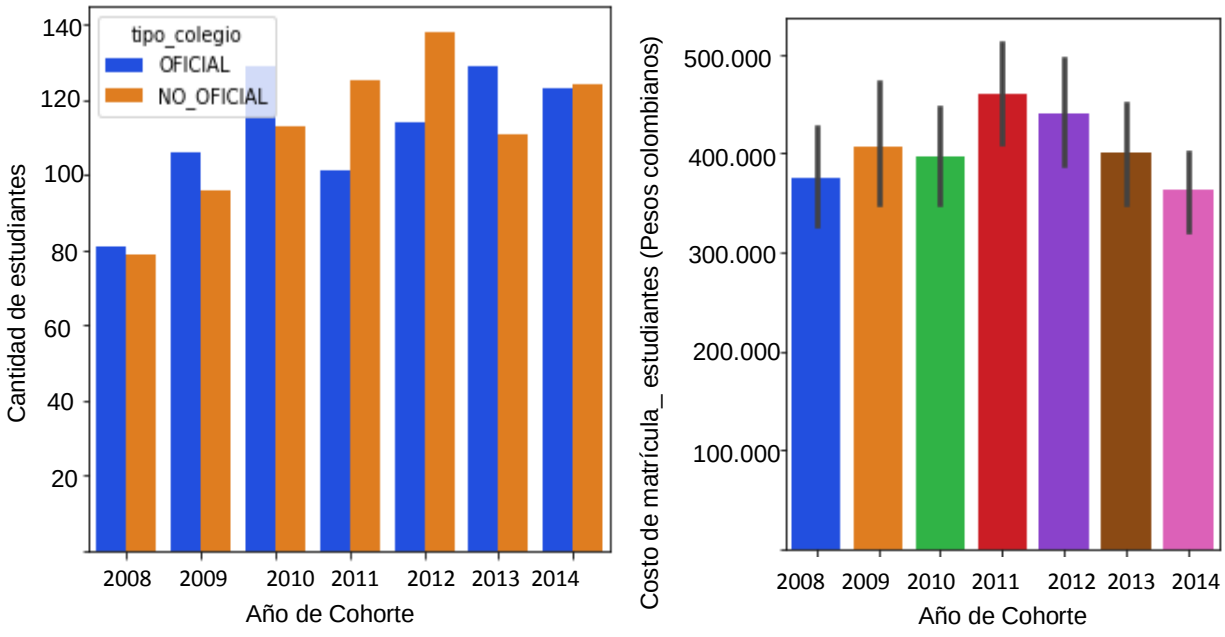
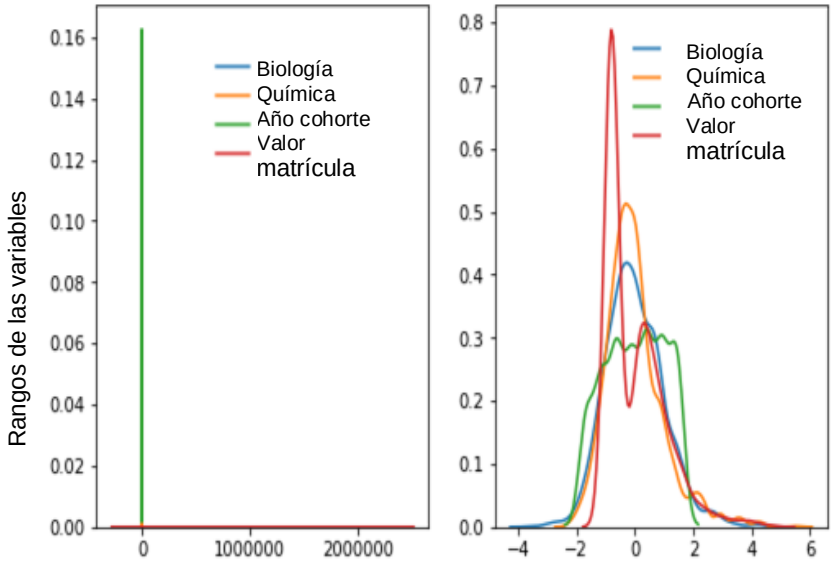


Fig.4: Estadísticas. (Izq.) Tipo de colegio de donde provienen los estudiantes. (Der.) Costo promedio del valor de la matrícula de un estudiante en pesos colombianos

En cuanto a Estandarización y Selección de características

La data inicial (1620 registros) debió depurarse, eliminando los registros de datos erróneos de la base de datos, quedando un conjunto de datos con 1571 registros y con 30 variables o características, los cuales debieron normalizarse como se muestra en la figura 5a), ya que muchos algoritmos de aprendizaje automático funcionan mejor cuando las características están en una escala relativamente similar y están cerca de la distribución normal (Jahangiri y Rakha, 2015). Como es necesario establecer si todas las variables (30) deben entrar en los modelos o cuáles de las variables son las más influyentes en la clase objetivo del conjunto de datos (rendimiento académico), en esta parte del estudio fueron implementados diferentes métodos de selección de características o variables. Algunos de los métodos fueron: Chi cuadrado, Anova, correlación de Pearson, eliminación de características recursivas con Regresión Logística y eliminación de características recursivas con validación cruzada (RFECV), entre otros. En la figura 5b), se muestran resultados del método eliminación de características recursivas con validación cruzada (RFECV)



Forma de la distribución para algunas variables antes de Normalizar

Forma de la distribución para algunas variables después de Normalizar

Fig. 5: (a) Ejemplo de Estandarización de variables

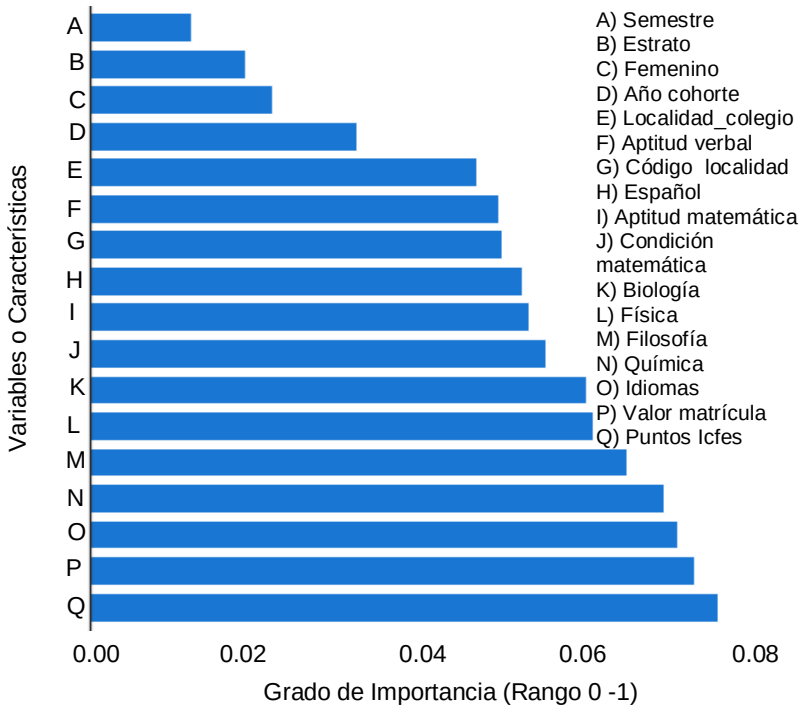


Fig. 5: (b) Selección de características por RFECV en Python.

De un total de 30 variables que se analizaron (mostradas en la Tabla 1), los métodos de selección de características arrojaron 10 características que tienen una gran influencia en el rendimiento académico del grupo de factores analizados. Estas se agrupan según la importancia para cada método (tabla 3). Para los tres primeros métodos de la tabla, la selección de características se realiza según puntajes estadísticos que tienden a determinar la correlación de las características con la variable de salida (rendimiento académico) Los otros métodos proporcionan resultados bastantes similares en cuanto a características relevantes para los modelos. Otro método ampliamente usado en Machine Learning, es el de análisis de componentes principales (PCA) que utiliza álgebra lineal para transformar el conjunto de datos en una forma comprimida realizando un cambio de dimensiones de la que tienen originalmente los datos. El resultado de este, es una matriz transformada que pude servir para aplicar algoritmos de Machine Learning y tal vez de buenos resultados. Pero no será posible saber cuáles son las características que más influyen en la variable de respuesta.

A partir de los métodos de selección de variables mencionados anteriormente, se determinó cuáles son las 7 características más influyentes sobre la variable de respuesta.

Tabla 3. Resultados de Métodos de selección de características

Método de selección de características	*Características seleccionadas*
Chi - Cuadrado	PIT, PIAM, PICM, PIAQ, PIAB, TIU, VM, E, NP, PIAV
Anova	CLC, PIAQ, PIAM, PIAE, PIAI, PIT, E, VM, NP, D, MB
Correlación de Pearson	CLC, PIAQ, PIAM, PIAE, PIAI, PIT, E, VM, NP, D, MB
Eliminación de características RFE: Regresión Logística	E, CC, G, TC, TIU, PIT, PIAM, PICM, PIAE, D, ME
Eliminación de características recursivas: Regresión Lineal	PIT, PIAM, PICM, PIAF, PIAI, G, CLC, C, AC, E
Eliminación de características recursivas RFE: SVM	E, CC, G, TC, PIT, PIAM, PICM, PIAF, PIA, PAI
Bosques aleatorios	VM, PIT, PIAI, PIAM, E, G, PIAF, PIAQ, PIAF, PICM
Eliminación hacia atrás	VM, C, AC, PIAF, CL, E, G, PIAM, PIAE, PICM, PIAQ
Arboles de decisión: Extra Trees	PIAB, PIAQ, PIAF, E, I, PIT, CC, PIAM, PICM

En cuanto a Predicción mediante Algoritmos

En el paso 6 se seleccionaron diversos algoritmos de aprendizaje automático para tareas de clasificación. Por cuestión de espacio solo se muestran imágenes de los resultados de algunos de ellos, los cuales son: 1) SVC; 2) Perceptrón; 3) KNN y 4) Árbol de decisión. Cada algoritmo requiere de un conjunto de datos para aprender y otro conjunto de datos para evaluar su rendimiento. En este estudio se seleccionó el denominado método aleatorio que consiste en tomar el conjunto de datos originales y dividirlo en dos grupos: unos datos para entrenamiento (70%) que corresponde a 1134 registros y unos datos para prueba (30%) que corresponden a 486 registros.

Modelo de clasificación (SVC)

Cada uno de los algoritmos usados tiene su propio hiperparámetro, que son valores que hay que identificar y modificar para lograr obtener los mejores resultados del algoritmo que se trabaja. Este algoritmo tiene como hiperparámetro el valor Max_iter con el cual se busca para cada una de las corridas del algoritmo cual sería el valor que proporciona una mejor precisión del modelo (se determina mediante el código realizado en Python). En la figura 6, se muestra como varia la precisión el modelo al variar el hiperparámetro tanto para los datos de prueba como para los datos de entrenamiento. Los mejores resultados se obtienen con un valor del hiperparámetro Max_iter= 6.

Modelo de clasificación (Perceptrón) y (KNN)

El algoritmo Perceptrón, es el algoritmo más simple dentro del grupo de algoritmos que usan redes neuronales como forma de trabajo. Es usado en problemas donde los datos son linealmente separables. Al igual que otros métodos busca un hiperplano de separación o medio para clasificar los datos del problema. La Fig. 7 muestra los resultados para el algoritmo perceptrón implementado usando datos de entrenamiento y datos de prueba. El mejor hiperparámetro (valor a identificar y modificar para lograr obtener los mejores resultados del algoritmo) con el que se trabaja, fue Max_iter=2.

El algoritmo K-vecinos más cercanos (KNN), es uno de los algoritmos básicos de los clásicos de machine Learning, este busca clasificar un dato teniendo en cuenta la distancia de separación de este dato puntual con respecto a sus vecinos. La distancia más usada es la euclidiana, pero también hay otras medidas que pueden ser más adecuadas para un entorno dado e incluyen la distancia de Manhattan y Minkowski. El

hiperparámetro a evaluar es n_neighbors (cantidad de vecinos más cercanos). Valor que se sugiere modificar para obtener mayor precisión tal como se muestra en la figura 7. Se observa que buenos valores podrían ser 10, 12 y 15 vecinos ya que según la curva arroja mayores valores de exactitud (métrica de valuación del algoritmo).

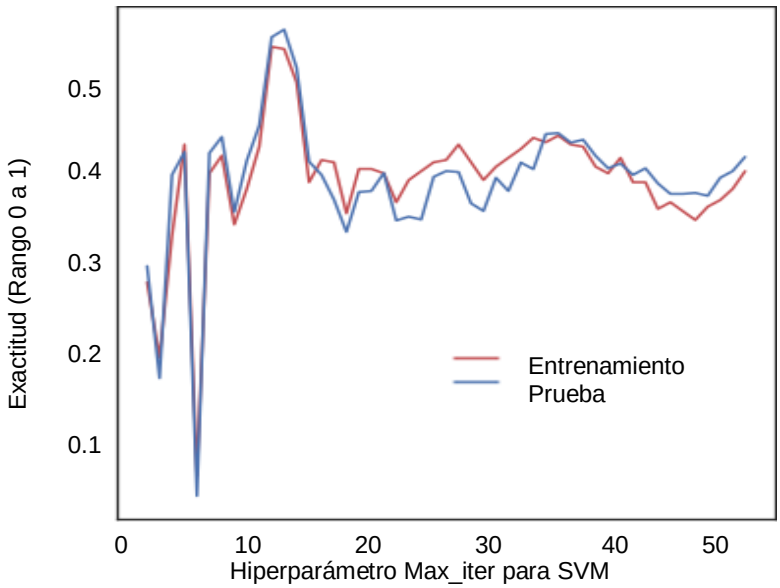


Fig. 6: Búsqueda del mejor Hiperparámetro para SVC

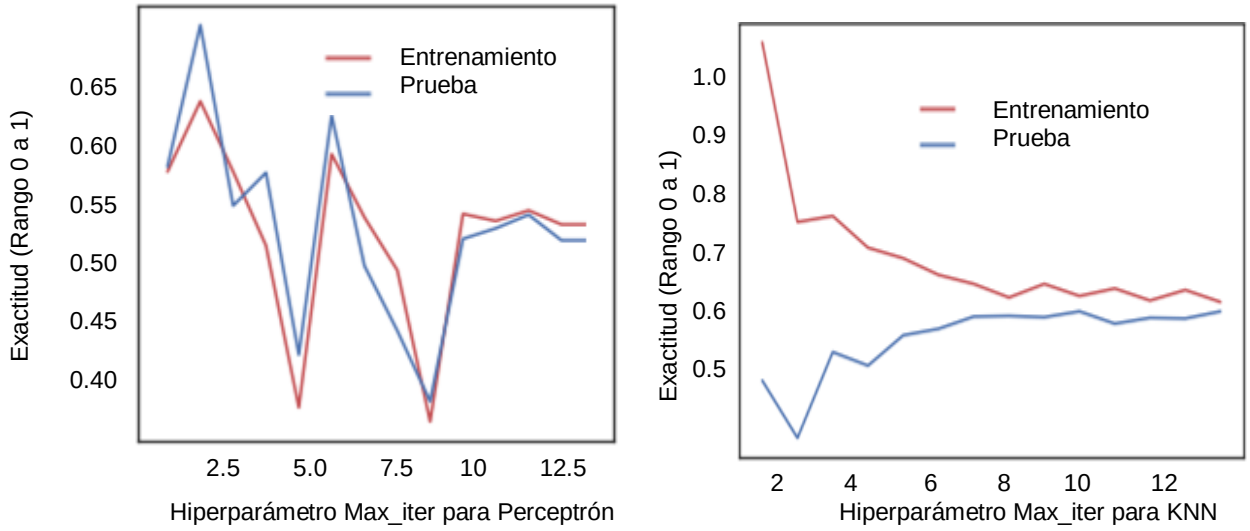


Fig. 7: (Izq.) Búsqueda del mejor hiperparámetro para Perceptrón (Der.) Búsqueda del mejor hiperparámetro para KNN

Modelo de clasificación (Árbol de decisión)

Fue necesario variar el hiperparámetro max_depth en el algoritmo con el fin de buscar cual sería el mejor para este caso; la mejor opción fue max_depth=7 (profundidad del árbol) se muestra en la figura 8.

En cuanto a evaluación de algoritmos

Existen diferentes métricas para determinar si un modelo ofrece buena precisión entre otras. En la tabla 4 se muestran los valores de exactitud que se puede definir como el número de predicciones correctas realizadas por el modelo para el número total de registros (filas del conjunto de datos). Si el valor es cercano o igual a 1 indica que todas las predicciones son correctas (100%). En esta tabla se muestra que cuando los datos que se pasan como material de análisis a cada uno de los algoritmos son los datos originales sin ninguna transformación, y mucho menos métodos de selección de características aplicado, los resultados en cuanto a la métrica de evaluación (exactitud) son mucho menores que cuando se aplica algún método de selección de características. Los algoritmos usados proporcionan exactitudes del orden de 0.54 (54%) frente al uso de métodos de selección de características del orden de 0.65 (65%).

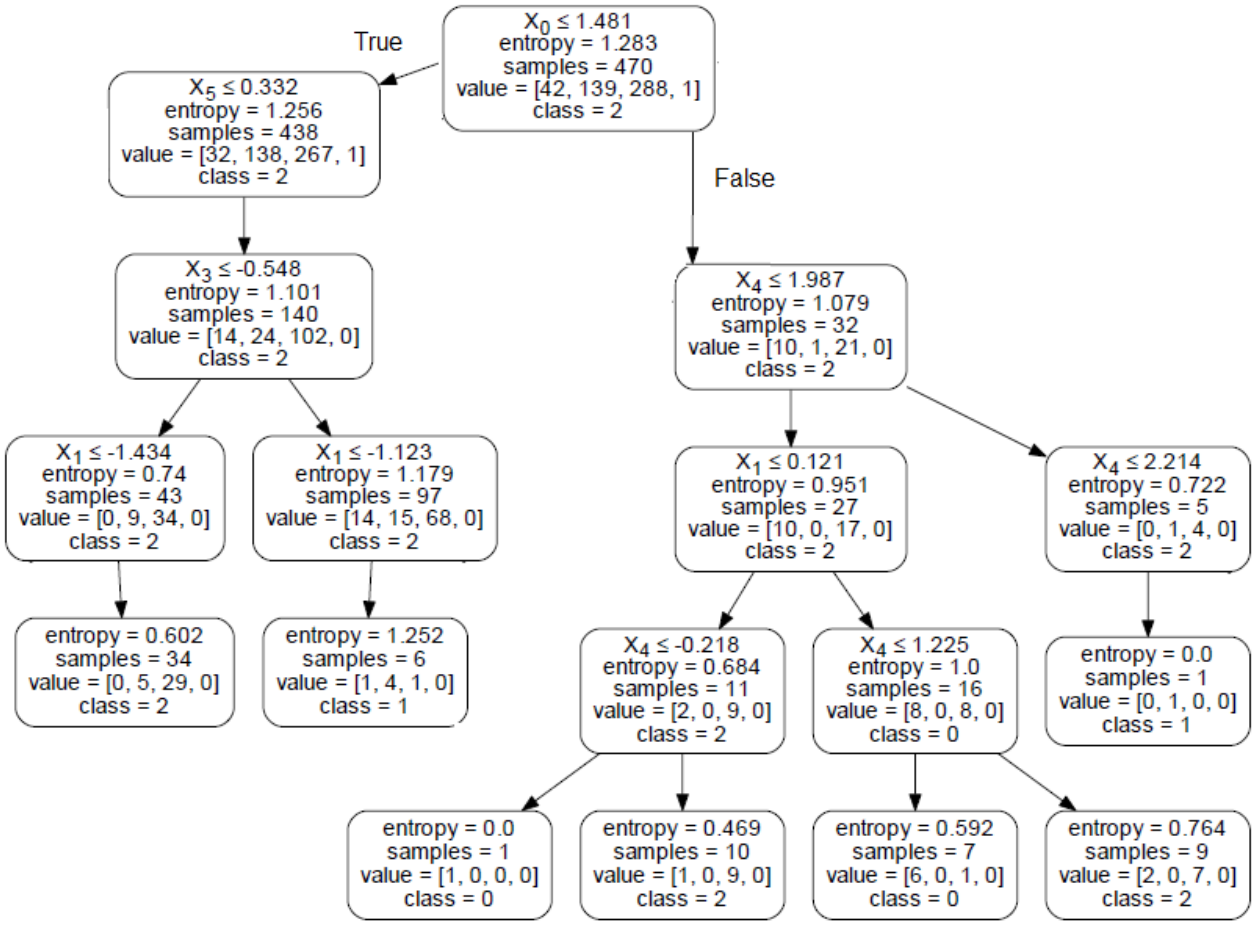


Fig. 8: Parte del árbol de decisión generado por Python

Tabla 4. Resumen de métrica de evaluación de Modelos de ML

Algoritmo	*Exactitud con todas las características*	*Exactitud con características seleccionadas*	*Precisión*	*Recall*	*F1*
SVC	0.54	0.66	0.61	0.66	0.54
KNN	0.54	0.64	0.55	0.64	0.57
Perceptrón	0.53	0.6624	0.61	0.66	0.54
Árbol de Decisión	0.55	0.65	0.55	0.64	0.57

DISCUSIÓN

En la última década según la investigación referencial ha existido una fuerte tendencia en los centros educativos por tratar de explicar las causas de los problemas que enfrenta el sector de la educación superior (abandono, deserción y rendimiento académico), haciendo uso de herramientas tecnológicas que permitan recopilar, analizar y visualizar información importante que no es evidente dentro de un conjunto de datos producidos por el estudiante y docentes durante el proceso de enseñanza-aprendizaje. Por esto, este proyecto inicial marca el camino para la investigación de predicción mediante herramientas de análisis de datos usando machine Learning ya que proporciona los algoritmos facilitadores para la búsqueda de información relevante contenida en los que son almacenados por las instituciones de educación superior y poco análisis se les ha realizado.

Según los resultados las mejores exactitudes se consiguen si se realiza el análisis de selección de atributos (características), situación similar fueron los resultados encontrados por (Nieto et al., 2018) que usaron algoritmos similares. En este trabajo los resultados son mejores para el algoritmo SVC (66.24%) y Perceptrón (66.0%) respectivamente., es decir de 1099 registros a los que debía predecir la clase, lo hizo correctamente para 728. Así mismo el SVC arroja otras métricas con valores relativamente buenos de precisión y sensibilidad (Recall) teniendo en cuenta que solo se analizaron 30 variables de las 105 encontradas en la literatura que pueden influir en la determinación del rendimiento académico en la educación superior.

Construir un modelo que contenga la mayoría de variables para determinar el rendimiento es un proceso que requiere mucho más análisis. Pero este trabajo arrojó resultados interesantes en el sentido de que se inicia con la selección de variables por diferentes métodos y la prueba con diferentes algoritmos (análisis comparativo). En este sentido las pruebas ICFES (pruebas de estado en Colombia) son un referente que ayuda a su determinación al igual que el género según la selección de características. Así mismo, el hecho de seleccionar las mejores variables para los modelos ayuda a mejorar la métrica de evaluación de los mismos. Los algoritmos usados arrojan resultados similares en la predicción de la clase (rendimiento académico) calificado como bajo, medio alto y superior (criterio establecido por el MEN – Ministerio de educación de Colombia).

CONCLUSIONES

De acuerdo al trabajo presentado y a los resultados obtenidos, se pueden plantear las siguientes conclusiones principales:

1.- Las siete variables que más influyen en el rendimiento académico de los estudiantes de ingeniería a partir de los factores analizados son: Edad, Genero, Puntaje ICFES para aptitud Matemática, Puntaje global ICFES, valor de matrícula, Puntaje ICFES para condición Matemática y Cohorte

2.- Los algoritmos Máquina de vectores de soporte (SVM) y Perceptrón (red neuronal) fueron elegidos en base no solo a la popularidad observada en investigación referencial relativa a este trabajo, sino también a su precisión, hecho que se vio reflejado en que obtuvieron los mejores resultados en cuanto a métricas de evaluación. El modelo de perceptrón muestra ser exitoso para la determinación del rendimiento académico con una exactitud del 66.4%

3.- Con el fin de conseguir un mejor resultado en las métricas de evaluación de los algoritmos llevándolos por encima del 90% se hace necesario utilizar las variables de otros factores que influyen en el rendimiento académico tales como: factores de gestión académica universitaria, tecnológicos, de Biblioteca, Institucionales, pedagógicos e intelectuales. Esto sin demeritar los resultados obtenidos hasta el momento en el que fueron analizados factores académicos pre-universidad, demográfico y socio-económicos.

REFERENCIAS

Burgos, C., Campanario, M. y otros cuatro autores, Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout, https://doi.org/10.1016/j.compeleceng.2017.03.005, Computers and Electrical Engineering, 66, 541–556 (2018)

Castrillón, O., Sarache, W., y Ruiz, S., Predicción del rendimiento académico por medio de técnicas de inteligencia artificial, https://doi.org/10.4067/S0718-50062020000100093, Revista Formación Universitaria, 13, 93–102 (2020)

Contreras, L., y Rodriguez, J., Big data: An exploration toward the improve of the academic performance in higher education, https://doi.org/10.1007/978-3-319-93803-5_59. Lecture Notes in Computer Science, 10943, 627–637 (2018)

De La Hoz, E., De La Hoz, E., y Fontalvo, T. Methodology of Machine Learning for the classification and Prediction of users in Virtual Education Environments, https://doi.org/10.4067/S0718-07642019000100247, Información Tecnológica, 30, 247–254 (2019)

Dyckhoff, A., Zielke, D. y otros tres autores, Design and implementation of a learning analytics toolkit for teachers, Educational Technology and Society, 15, 58–76 (2012)

Escudero, T., Indicadores del rendimiento académico una experiencia en la Universidad de Zaragoza - Ministerio de Educación y Cultura, 1° edición, 251-262. Centro de Publicaciones, España (1999)

Fernandes, E., Holanda, M., y otros tres autores, Educational data mining: Predictive analysis of academic performance of public-school students in the capital of Brazil, https://doi.org/10.1016/j.jbusres.2018.02.012, Journal of Business Research, 94, February 2018, 335–343. (2019)

Garbanzo R y María, G., Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública, Revista Educación, 31,1, 43–63 (2007)

García, J., Sánchez, P., Orozco, M., y Obredor, S., Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia, https://doi.org/10.4067/S0718-50062019000400055, Revista Formación Universitaria, 12, 4, 55–62 (2019)

García, K., Learning Analytics as an analysis factor of university academic performance, CEUR Workshop Proceedings, 2231, 42-50 (2019)

Jahangiri, A., y Rakha, H., Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data, https://doi.org/10.1109/TITS.2015.2405759, IEEE Transactions on Intelligent Transportation Systems, 16 ,5, 2406–2417 (2015)

Khan, I. A., y Choi, J. T. An Application of Educational Data Mining (EDM) Technique for Scholarship Prediction, https://doi.org/10.14257/ijseia.2014.8.12.03, International Journal of Software Engineering and Its Applications, 8 ,12, 31–42 (2014).

Lonn, S., Aguilar, S. y Teasley, S., Investigating student motivation in the context of a learning analytics intervention during a summer bridge program, https://doi.org/10.1016/J.CHB.2014.07.013, Computers in Human Behavior, 47, 90–97 (2015).

MEN-Sistema Nacional de Información de la Educación Superior. MEN-Sistema Nacional de Información de la Educación Superior, (2017)

Nieto, Y., García, V., Montenegro, C., y Crespo, R., Supporting academic decision making at higher educational institutions using machine learning-based algorithms, https://doi.org/10.1007/s00500-018-3064-6, Soft Computing, 23, 4145–4153 (2018)

Nithya, P., Umamaheswari, B., y Umadevi, A. A Survey on Educational Data Mining in Field of Education, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 16,1, 145-153 (2019)

Noel, M., Ayán, R., Ángel, M., y Díaz, R., Indicadores de rendimiento de estudiantes universitarios, http://10-4438/1988-592X-RE-2011-355-033, Revista de Educación, 355, 467– 492 (2011)

Oblinger, G., Campbell, J., y otros dos autores, Academic Analytics: A New Tool for a New Era, Research in Higher Education, 1(2), 727–742 (2007)

Osmanbegović, E., y Suljić, M., Data mining approach for predicting student performance, Journal of Economics and Business, 10, 1, 20–30 (2012).

Page, M., Gaviria, J., y Gómez, C., Hacia un modelo causal del rendimiento académico, Ministerio de educación, 1° edición, 25-230. Centro de publicaciones y secretaria General, España, (1990)

Palmer, S., y Stuart., Modelling engineering student academic performance using academic analytics, The International journal of engineering education,29,1,132-138 (2013)

Porto, A., y Gresia, L. Di., Performance of University students and their determinants. Revista de economia y estadística, 42,1, 93-113 (2005)

Radhwan A., Abbas A, y Ali S., Popular Decision Tree Algorithms of Data Mining Techniques, International Journal of Computer Science and Mobile Computing, 6,6, 133–142 (2017)

Ramesh, V., Parkavi, P., y Ramar, K., Predicting Student Performance: A Statistical and Data Mining Approach, https://www.ijcaonline.org/archives/volume63/number8/10489-5242, International Journal of Computer Applications, 63,8,35-39 (2013).

Rojas, L., Validez predictiva de los componentes del promedio de admisión a la universidad de costa rica utilizando el género y el tipo de colegio como variables control, https://revistas.ucr.ac.cr/index.php/aie/article/view/11707/18183, Revista Electrónica Actualidades Investigativas En Educación, 13(1), 17–25, (2013)

Salcedo, A., Desertion in Colombian Universities, http://www.alfaguia.org/alfaguia/files/1319043663_03.pdf, Revista Academia y Virtualidad, 3(1), 50–60 (2010)

Sánchez, P. y García, J., A new methodology for neural network training ensures error reduction in time series forecasting, https://doi.org/10.3844/jcssp.2017.211.217, Journal of Computer Science, 13, 211–217 (2017)

Santosh, K., AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data, https://doi.org/10.1007/s10916-020-01562-1, Journal of Medical Systems, 44(5), 1–5 (2020)

Tourón, J., La predicción del rendimiento académico: Procedimientos y resultados, http://dadun.unav.edu/handle/10171/18774, Revista Española de Pedagogía, 1(25), 168–182 (1985)

Zaffar, M., Hashmani, M. A., Savita, K. S., y otros tres autores, A Study of Feature Selection Algorithms for Predicting Students Academic Performance, https://10.14569/IJACSA.2018.090569 International Journal of Advanced Computer Science and Applications, 9(5), 541-549 (2018).