Análisis de datos para la predicción del éxito académico mediante árboles de decisión

Alejandra Palacio Jaramillo Universidad EAFIT Colombia apalacioj@eafit.edu.co	Valentina Moreno Ramírez Universidad EAFIT Colombia vmorenor@eafit.edu.co	Miguel Correa Universidad EAFIT Colombia macorream@eafit.edu.co	Mauricio Toro Universidad EAFIT Colombia mtorobe@eafit.edu.co

RESUMEN

El objetivo de este proyecto es crear un algoritmo que permita, mediante una estructura de datos, predecir el éxito académico de una persona en las Pruebas Saber Pro, teniendo en cuenta su resultado en las Pruebas Saber 11 y demás factores relacionados como su condición socioeconómica, el pregrado estudiado, entre otros. La importancia de solucionar este problema radica en que cada día el pronóstico de la información es de mayor relevancia, puesto que tener una idea de lo que sucederá a futuro permite actuar de manera rápida y eficiente frente a lo predicho, gracias a la implementación de las tecnologías emergentes. Con base en soluciones a problemas similares, se hallará la solución más adecuada.

Palabras clave

Árboles de decisión, aprendizaje automático, éxito académico, predicción de los resultados de los exámenes, minería de datos, rendimiento académico.

1. INTRODUCCIÓN

En la actualidad, con la llegada de la recolección de datos, se ha dado espacio a la aparición de problemas relacionados con la predicción y el análisis de los resultados en futuro eventos, con el fin de implementar estrategias que permitan mejorar y optimizar la calidad de estos en años próximos. Para Colombia, y en general para América Latina, es claro que uno de los mayores problemas es la incertidumbre que genera la calidad de la educación y la eficiencia del sistema educativo en relación con el aprendizaje del estudiante. El hecho de pensar en la creación de un algoritmo único que permita anticipar los resultados académicos de los estudiantes para encontrar las falencias que hacen de la educación latinoamericana inoperante respecto al resto del mundo, es la motivación para trabajar en este proyecto arduamente, junto con la idea de llevar a cabo planes que mejoren este sistema que ha estado fragmentado durante años.

1.1. Problema

De acuerdo con la información personal, familiar, académica, socioeconómica y sociodemográfica que se tiene a disposición de estudiantes que han presentado las Pruebas Saber 11, se desarrollará un algoritmo basado en el uso de árboles de decisión, el cual predecirá si los resultados totales en las pruebas Saber Pro están por encima del promedio o no. Teniendo en cuenta las herramientas enseñadas en el curso de Estructuras de Datos y Algoritmos I, el concepto y funcionamiento de los árboles de decisión y los conocimientos previos respecto a la temática, en resumen, se logrará presagiar el éxito académico en esta prueba. La importancia de resolver este problema se sintetiza en cuanto puede mejorar el rendimiento académico en las áreas de mayor dificultad si se tiene una idea previa (predicción) de los posibles resultados.

2. TRABAJOS RELACIONADOS

3.1 Predicción de factores relacionados al desempeño académico.

Como problema de la investigación se tenía el detectar los factores y patrones asociados al rendimiento académico de estudiantes colombianos de grado undécimo que presentaron las Pruebas Saber 11, teniendo en cuenta la información socioeconómica, académica e institucional de las bases de datos del ICFES. Para obtener esta predicción y así estimar las próximas condiciones de los estudiantes en futuras generaciones, se utilizó la herramienta WEKA y su algoritmo J48, el cual es una implementación del árbol de decisión C4.5, el cual tuvo una precisión del 67% en la clasificación correcta de los datos, pero a su vez tuvo un 33% de imprecisión al clasificar incorrectamente cierta cantidad de instancias. [3]

3.2 Predicción del rendimiento académico con minería de datos.

El problema planteado en esta investigación consistía en predecir el resultado final (ganado o perdido) de los estudiantes matriculados en un curso de Estadística General en la UNALM, tomando en cuenta su información académica previa, con el fin de contar con información adecuada para identificar, posteriormente, los aspectos influyentes en el aprendizaje. Para solucionar el problema, se utilizaron diferentes algoritmos como las redes bayesianas y los árboles de decisión. Como árbol de decisión, se aplicó el método C4.5 con un nivel de confianza de 0.15 y con poda, el cual tuvo una precisión 68.3%, 2.7% menos que el algoritmo que obtuvo mayor precisión (71%). [4]

3.3 Comparación de técnicas de minería de datos para predecir resultados académicos.

La cuestión en desarrollo en esta investigación es la evaluación comparativa de la funcionalidad en tiempo de ejecución y precisión de algoritmos de clasificación para predecir el rendimiento académico de los estudiantes (bajo, medio y alto), con base en sus datos académicos previos (calificaciones). Dentro de los algoritmos comparados, se encuentran los árboles de decisión y se toma como modelo el algoritmo CART, el cuál obtuvo un 75% de precisión y un 25% de error. [5]

3.4 Predicción del rendimiento de los estudiantes utilizando dos tipos de árboles de decisión.

El problema tratado en esta investigación es predecir el rendimiento académico de estudiantes de primer año de ingeniería, con base en los resultados de los estudiantes de ingeniería que en ese momento cursaban el segundo año. Dentro de los datos se tenían el nombre, género, calificaciones en el semestre, calificación obtenida en el examen de ingreso, tipo de admisión, entre otros. Para la predicción, se utilizaron dos árboles de decisión: el algoritmo ID3 y el C4.5, los cuales arrojaron la misma precisión, tanto en las evaluaciones masivas como en las singulares, siendo esta, en promedio, del 75.27% de exactitud. [6]

3. MATERIALES Y MÉTODOS

En esta sección se explica cómo se recopilaron y procesaron los datos y, después, cómo se consideraron diferentes alternativas de solución para elegir un algoritmo de árbol de decisión.

3.1 Recopilación y procesamiento de datos

Obtuvimos datos del *Instituto Colombiano de Fomento de la Educación Superior* (ICFES), que están disponibles en línea en ftp.icfes.gov.co. Estos datos incluyen resultados anonimizados de Saber 11 y Saber Pro. Se obtuvieron los resultados de Saber 11 de todos los graduados de escuelas secundarias colombianas, de 2008 a 2014, y los resultados de Saber Pro de todos los graduados de pregrados colombianos, de 2012 a 2018. Hubo 864.000 registros para Saber 11 y 430.000 para Saber Pro. Tanto Saber 11 como Saber Pro, incluyeron, no sólo las puntuaciones sino también datos socioeconómicos de los estudiantes, recogidos por el ICFES, antes de la prueba.

En el siguiente paso, ambos conjuntos de datos se fusionaron usando el identificador único asignado a cada estudiante. Por lo tanto, se creó un nuevo conjunto de datos que incluía a los estudiantes que hicieron ambos exámenes estandarizados. El tamaño de este nuevo conjunto de datos es de 212.010 estudiantes. Después, la variable predictora binaria se definió de la siguiente manera: ¿El puntaje del estudiante en el Saber Pro es mayor que el promedio nacional del período en que presentó el examen?

Se descubrió que los conjuntos de datos no estaban equilibrados. Había 95.741 estudiantes por encima de la media y 101.332 por debajo de la media. Realizamos un submuestreo para equilibrar el conjunto de datos en una proporción de 50%-50%. Después del submuestreo, el conjunto final de datos tenía 191.412 estudiantes.

Por último, para analizar la eficiencia y las tasas de aprendizaje de nuestra implementación, creamos al azar subconjuntos del conjunto de datos principal, como se muestra en la Tabla 1. Cada conjunto de datos se dividió en un 70% para entrenamiento y un 30% para validación. Los conjuntos de datos están disponibles en https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets .

	Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3	Conjunto de datos 4	Conjunto de datos 5
Entrenamiento	15,000	45,000	75,000	105,000	135,000
Validación	5,000	15,000	25,000	35,000	45,000

Tabla 1. Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

3.2 Alternativas de algoritmos de árbol de decisión:

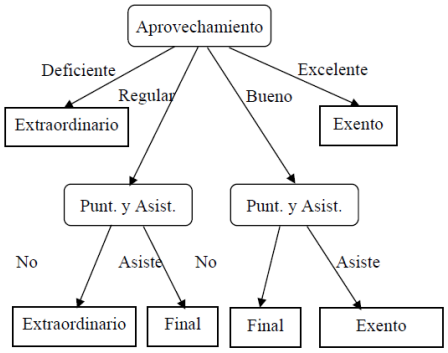
Un árbol de decisión es un clasificador que divide los datos de forma recursiva para formar grupos o clases.

3.2.1 Algoritmo ID3:

Fue el primer árbol de decisión desarrollado por Ross Quinlan. Este construye un árbol de decisión para los datos dados de forma descendente, comenzando por un conjunto de objetos y una especificación de propiedades, recursos e información.[1]

En cada nodo del árbol, se revisan los atributos en función de maximizar la ganancia de información y minimizar el desordenamiento de los datos. Luego de maximizar la ganancia, se forma la ramificación. Cabe destacar que este proceso se realiza de forma recursiva hasta que el conjunto en un subárbol dado contenga objetos de la misma categoría. En sí, el algoritmo ID3 selecciona una prueba usando el criterio de ganancia de información, y no busca otras opciones. [1]

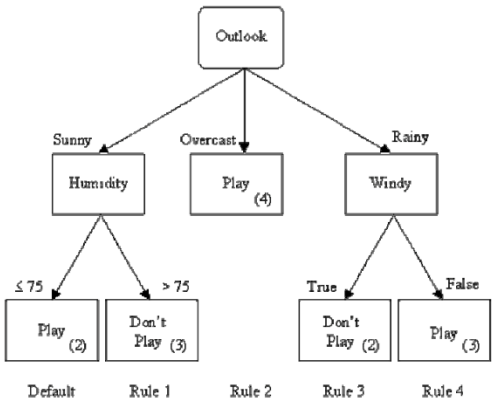
Para n atributos, hay 2^n filas y podemos considerar la salida como una funcion definida por 2^ n. Con esto hay 2^2^n posibles funciones diferentes para n atributos.



3.2.2 Algoritmo C4.5:

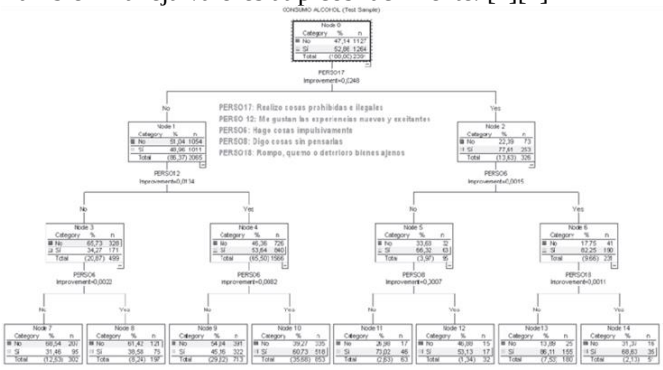
El algoritmo C4.5 es una versión mejorada del algoritmo ID3. C4.5 usa el proceso de *Shannon Entropy* (medir la incertidumbre de una fuente de información), para elegir características con la mayor ganancia de información. Una característica particular de este algoritmo es que construye los nodos con ramas vacías con valores de cero. [1]

El algoritmo considera los casos bases, luego todas las pruebas posibles que pueden dividir el conjunto de datos, posteriormente selecciona la prueba que resulta en la mayor ganancia de información, la toma como parámetro de decisión en el siguiente nodo y sigue así hasta obtener las posibles respuestas. Este proceso de particiones de datos se realiza recursivamente (se puede observar que tiene un comportamiento similar al "divide y vencerás" del algoritmo ID.3). A diferencia del algoritmo ID3, el C4.5 puede tomar tanto valores discretos como continuos, tiene menos errores en la poda, es más eficiente, etc. Cabe resaltar que, para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo, mientras que, para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos. Tomando en cuenta lo anterior, se puede afirmar que, en cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos. [2]



3.2.3 Algoritmo CART:

Los árboles de Clasificación y regresión CART (*Classification And Regression Trees*) se caracteriza por el hecho de construir el árbol a partir de particiones binarias recursivas (árboles binarios), es decir, cada nodo tiene exactamente dos bordes salientes. Las divisiones se seleccionan utilizando los criterios de dosificación y el árbol obtenido, se poda mediante poda de coste-complejidad. En otras palabras, también se puede decir que se selecciona la variable que minimice la impureza de Gini, para obtener valores más correctos. Es importante resaltar que CART puede manejar variables numéricas y categóricas (objetos). También maneja valores atípicos fácilmente. [1][7]



Como se mencionó anteriormente, el algoritmo utiliza el índice de Gini para calcular la medida de impureza:

$$G(Ai)=Mi\sum j=ip(Aij)G(C/Aij)$$

Siendo:

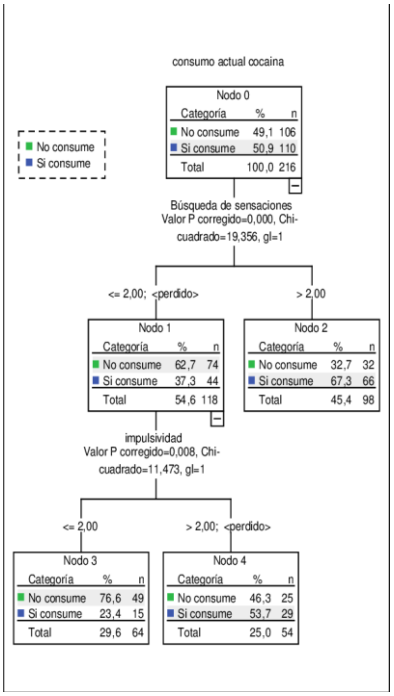
- **Aij:** es el atributo empleado para ramificar el árbol,
- **J** es el número de clases,
- **Mi** es el de valores distintos que tiene el **bv cfr** atributo AiAi
- **p(Aij)** constituye la probabilidad de que AiAi tome su j-ésimo valor y representa la probabilidad de que un ejemplo sea de la clase CkCk cuando su atributo AiAi toma su j-ésimo valor.

El índice de diversidad de Gini toma el valor cero cuando un grupo es completamente homogéneo y el mayor valor lo alcanza cuando todas las p(Aij) son constantes, entonces el valor del índice es (J−1)J.

3.2.4 Algoritmo CHAID

CHAID es un algoritmo rápido y eficaz que permite la creación de árboles de decisión mediante la significancia ajustada de los datos. Se caracteriza por el uso de la chi-cuadrado de Pearson para determinar si la variable predictora

tiene una interacción o ajuste significativo respecto a la variable dependiente. Si la interacción es significativa, entonces se dividirá el nodo con base a esa variable y, si no lo es, se fundirá la variable junto con las otras que no son significativas, es decir, junto al grupo al que más se parece la variable, respecto a la variable dependiente. Cabe destacar que este algoritmo solo trata variables discretas. [8][9]



REFERENCIAS

[1] Anónimo. *Cuáles son las diferencias entre ID3, C4.5 y CART?*. Organización PresMaryTwen, Abril del 2020, from https://presmarymethuen.org/es/dictionary/what-are-the-differences-between-id3-c4-5-and-cart/

[2] López, B. *AlGORITMO C4.5*. Instituto Tecnológico Nuevo Laredo, Nuevo Laredo, Tamaulipas, Noviembre del 2005, from http://www.itnuevolaredo.edu.mx/takeyas/apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf

[3] Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. *Revista de Investigación Desarrollo e Innovación*, *9*(2), 363-378, from https://revistas.uptc.edu.co/index.php/investigacion_duitama/article/view/9184/7721

[4] Menacho, C.H. Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales científicos*, *78*(1), 26-33, from https://dialnet.unirioja.es/servlet/articulo?codigo=6171237

[5] Ochoa, L.L., Rosas, K., Baluarte, C. *Evaluación de Técnicas de Minería de Datos para la Predicción del Rendimiento Académico*. Global Partnerships for Development and Engineering Education: Proceedings of the 15th LACCEI International Multi-Conference for Engineering, Education and Technology, Boca Raton, FL, Estados Unidos, 2017, from http://www.laccei.org/LACCEI2017-BocaRaton/full_papers/FP368.pdf

[6] Adhatrao, k., Gaykar, A. Dhawan, a., Jha, R. & Honrao, V. Predicting Students' Performance Using Id3 And C4.5 Classification Algorithms. *International Journal of Data Mining & Knowledge Management Process, 3*(5), 39-52, from https://github.com/mauriciotoro/ST0245-Eafit/blob/master/proyecto/problemas-relacionados/PREDICTING%20STUDENTS%E2%80%99%20PERFORMANCE%20USING%20ID3%20%26%20C4.5.pdf

[7] Anónimo. *Aprendizaje automatizado: árboles de clasificación.* Departamento de Sistemas e Informática, from https://www.dsi.fceia.unr.edu.ar/downloads/ing_conocimiento/Presentaciones/ArbDec09.pdf

[8] Teixeira, O. Integración del algoritmo CHAID y una adaptación de este, CHAID*, en la plataforma Weka. Universidad de País Vasco, 2016, from https://addi.ehu.es/bitstream/handle/10810/21808/TFG_TeixeiraMartin.pdf?sequence=1&isAllowed=y

[9] Berlanga, V., Rubio, M.J., Vilà, R. Cómo aplicar árboles de decisión en SPSS. *REIRE, 6*(1), 65-79, from https://revistes.ub.edu/index.php/REIRE/article/viewFile/5155/7229#:~:text=La%20funci%C3%B3n%20%C3%A1rboles%20de%20decisi%C3%B3n,se%20ajuste%20a%20nuestros%20datos.