PREDICCIÓN DE RESULTADOS EN LAS PRUEBAS SABER PRO A PARTIR DE LAS PRUEBAS SABER 11

Maria Alejandra Vélez Clavijo Universidad Eafit Colombia mavelezc1@eafit.edu.co	Laura Katterine Zapata Rendón Universidad Eafit Colombia lkzapatar@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

RESUMEN

En busca de contribuir con las predicciones en el éxito académico de la educación superior colombiana, se plantea generar un algoritmo basado en árboles de decisiones el cuál permita analizar los diversos factores (socioeconómicos, académicos, etc.) que influyen en el proceso de los estudiantes a la hora de poner a prueba los conocimientos adquiridos.

Se pretende que este trabajo sea diferente a los antes realizados relacionados con el tema, ya que gran parte de los análisis que se han hecho son meramente estadísticos y dejan por fuera ciertas variables de gran influencia para el desempeño en las pruebas, cosas que con el trabajo con árboles de decisión efectivamente pueden ser evaluadas y tomadas en cuenta para la predicción esperada

Palabras clave
Árboles de decisión, aprendizaje automático, éxito académico, predicción de los resultados de los exámenes

1. INTRODUCCIÓN

En los próximos años el papel de la tecnología será un factor importante en la transformación digital de la educación en Colombia. Dicha transformación se conoce como Educación 4.0 y tiene como objetivo mejorar la calidad de la enseñanza aprovechando las herramientas que nos ofrecen las nuevas tecnologías. En relación a temas de educación en el pasado, se han estudiado qué factores influyen en la deserción académica, cuáles son sus causas y motivaciones, y se han utilizado algoritmos para predecir la deserción, sin embargo, es muy poco lo que se ha logrado para predecir el éxito académico en la educación superior.

El objetivo de este estudio es contribuir con el análisis comprensivo de características que influyen en los resulta- dos de los desempeños que obtienen los estudiantes colombianos en las pruebas de Estado, en este caso específica- mente Saber Pro. De esta manera se busca no solamente contribuir a la mejora y evaluación de la educación, sino a la proyección social y de desarrollo general del país.

1.1. Problema

El reto que se plantea es diseñar un algoritmo, basado en árboles de decisión y en los datos de las pruebas Saber 11,

para predecir si un estudiante tendrá un puntaje total por encima del promedio o no, en las pruebas Saber Pro.

En particular, las variables académicas y sociodemográficas que se tiene a disposición son: la edad, el ingreso de los padres, la carrera, los resultados en el Saber 11, el género, el estrato, las horas que invierte en internet, entre muchas otras. Con estas variables, el objetivo es crear un árbol de decisión que pueda predecir la probabilidad que tiene un estudiante de obtener un resultado por encima del promedio. Además de las variables sociodemográficas y académicas, para cada estudiante, se cuenta con una variable que dice si un estudiante obtuvo un resultado por encima del promedio o no, en el puntaje total, de las pruebas Saber Pro.

1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad y una muy buena precisión en las predicciones que se realizan con estos. Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad y no se tiene garantía de que su solución sea altamente confiable.

Como solución se implementa un algoritmo basado en un árbol de decisión para predecir el éxito de los estudiantes en las pruebas saber pro, por medio del algoritmo C4.5, ya que este puede funcionar tanto con datos discretos como continuos y permite trabajar con datos incompletos. C4.5 emplea la poda mitigando el sobreajuste de los datos, eliminando implícitamente las variables de menor aporte.

2. TRABAJOS RELACIONADOS

2.1 Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°

Se realizó un estudio con el fin de obtener un modelo que permitiera predecir para los estudiantes que presentarían futuras pruebas saber 11°; los factores asociados al buen o mal desempeño académico en dichas pruebas. Lo que se destaca de este estudio a comparación de otros relacionados, es que la mayoría de estos se basan en análisis estadísticos en los que dejan de lado muchas variables influyentes como lo son las interrelaciones, cosa que en este proceso se toma en cuenta, utilizando un tratamiento más complejo de los datos.

Para detectar los agentes influyentes en el desempeño de los estudiantes colombianos que presentaron las pruebas saber 11° en años anteriores (2015 – 2016), se utilizó un modelo de clasificación basado en árboles de decisión, con información de las bases de datos del ICFES (información socioeconómica, institucional y académica). Para ello se empleó el algoritmo J48, el cual implementa al algoritmo C4.5, teniendo en cuenta que para este tipo de proyectos suele ser muy utilizado por su simplicidad y facilidad de entender los resultados.

Según los resultados del análisis realizado se concluyó que fue mayor el porcentaje de estudiantes que tienen un desempeño académico bajo, además se obtuvo que los patrones más influyentes en el desempeño de las pruebas, son el estrato socioeconómico, la jornada de estudio, el índice de TIC y la edad menor que 18 años. Es menester recalcar que estos patrones descubiertos también sirvieron de ayuda en los procesos de toma de decisiones del Ministerio de Educación Nacional, y demás instituciones relacionadas con la calidad de educación Colombiana.

2.2 Predicción de la Deserción Académica a través de árboles de decisión .

El objetivo del estudio realizado se centra en presentar clasificaciones basadas en árboles de decisión con parámetros optimizados (utilizados para una mayor precisión en los resultados) y de esta forma poder predecir la deserción de estudiantes universitarios, en este caso estudiantes de una universidad pública chilena.

Los atributos que se seleccionaron para hacer el análisis se relacionan a variables demográficas, antecedentes de ingreso a la universidad, situación económica y datos de rendimiento académico; para dicho análisis se utilizó la herramienta RapidMiner Studio 7.5 que implementa el algoritmo C4.5.

Para una mayor confiabilidad en los resultados, se hizo un proceso de optimización de parámetros, profundidad máxima del árbol y se halló el nivel de confianza utilizado para el cálculo de error pesimista de la poda.

En cuanto a los factores que impactan a la deserción estudiantil se encuentra que el promedio de notas es el mayor causante de dicha deserción, también tiene gran peso la cantidad de asignaturas aprobadas, años de avance en la carrera y el puntaje de ingreso a la universidad.

Los resultados de precisión en la predicción que obtuvieron superaron a los obtenidos en otras investigaciones que han utilizado C4.5 (se obtuvo una precisión de 87.27%), sin embargo se indica que con fines de mejorar la precisión se podrían utilizar otras técnicas de clasificación.

2.3 Un modelo basado en árboles de decisión para predecir la deserción estudiantil en la Educación Superior Privada.

Debido a la gran cantidad de datos que tienen las Instituciones de Educación Superior Universitaria en este trabajo de investigación se propone hacer uso de las técnicas de minería de datos para predecir la deserción o el abandono en la Educación Superior Privada. Para el desarrollo de proyecto se usó la metodología CRIPS-DM con la herramienta comercial spss clementine 12.0,para los cuales se hicieron uso de la técnica de minería de datos árboles de decisión, para lo cual se utilizaron 1761 datos de los estudiantes de la Universidad Privada César Vallejo, comprendidos del semestre 2009-I al semestre 2013-II de la Escuela profesional de Ingeniería de Sistemas con 27 atributos para cada uno de ellos que están relacionadas con la deserción del alumno, que fueron extraídos del área de registros académicos, Asuntos Estudiantiles y del área de Informática. Para el desarrollo del proyecto se hizo uso del algoritmo de árboles de decisión en donde se hizo el entrenamiento, validación y prueba con 100 datos nuevos en donde se obtuvo una precisión de 89%.[4]

2.4 Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos

Se analiza información académica con el objetivo de identificar factores que influyen sobre la deserción de los estudiantes de la carrera de Ingeniería en Informática de la Universidad Gastón Dachary en Argentina.

Este trabajo se centra en el análisis de variables relacionadas directamente con los resultados académicos del estudiante y su interacción con la universidad, basado en los datos que se obtienen del trayecto del estudiante en la carrera. El problema abordado es complejo, debido a que los datos pueden presentar una alta dimensionalidad (muchas variables o características que pueden influir) y suelen estar desbalanceados (muchos estudiantes suelen aprobar y sólo algunos pocos desertan). El objetivo es detectar con anterioridad cuales son los estudiantes que presentan características relacionadas con la posibilidad de abandono y así proveer contención o ayuda especial y así evitar y/o disminuir los casos de deserción estudiantil.

El primer algoritmo utilizado es C4.5, que genera un árbol de decisión a partir de las variables disponibles, mediante particiones realizadas recursivamente.

El segundo algoritmo empleado se denomina Naïve Bayes aumentado a árbol (Tree Augmented Network (TAN)), como todos los clasificadores Bayesianos, se basan en el teorema de Bayes, conocido como la fórmula de la probabilidad de las causas.

Como último algoritmo se escogió a OneR, el cuál es uno de los algoritmos clasificadores más sencillos y rápidos; dado que simplemente identifica el atributo que mejor explica la clase de salida.

Con relación a la precisión obtenida para los casos que desertan ("Des") o no desertan ("NoDes"), podemos ver de qué manera general los 3 algoritmos clasifican con más exactitud los casos que no desertan, donde el algoritmo OneR obtuvo el porcentaje superior (82,3%) en relación a los demás; y para los casos de deserción, el algoritmo que obtuvo mayor precisión es J48 (79,7%).[5]

3.2 Alternativas de algoritmos de árbol de decisión

3.2.1 ADTree

Un Árbol de decisión alternativo (ADTree) es un método de clasificación proveniente del aprendizaje automático, el cual básicamente se fundamenta en reglas. Una sola regla consta de una condición previa, una condición y dos puntuaciones. A continuación se muestra la estructura básica de un ADTre:

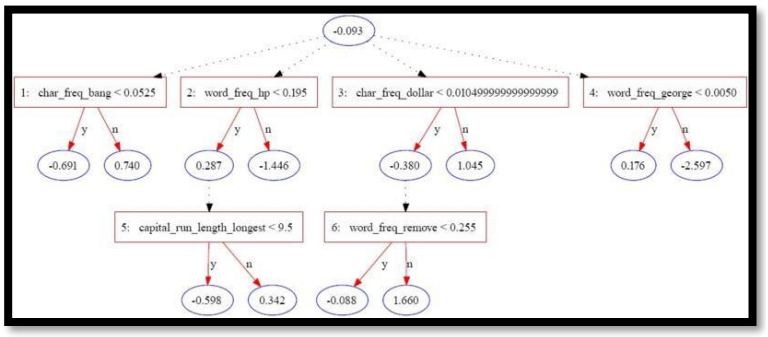

Figura 1: Árbol ADTree: el spam se codifica como 1 y el correo electrónico regular se codifica como -1. [1]

La estructura de un ADTree consiste en tener un nodo raíz de dirección que siempre va a ser un nodo de predicción; y ciertas unidades, cada unidad es una regla de desición que se compone de un nodo splitter (nodos de desición) y meramente de dos hijos (nodos de predicción), de esta forma cada paso positivo es seleccionado y se adiciona una nueva regla, de lo contrario no incrementa dichas reglas en la estructura, puesto que directamente se obtiene el valor de predicción. El valor de cada instancia se obtiene sumando las puntuaciones de los nodos de predicción por los cuales pasó la ruta seleccionada. A comparación de otros algorítmos requiere mucho menos iteraciones y es de fácil visualización.

3.2.2 ID3

El algoritmo ID3, es un algoritmo que permite crear árboles de desición y determinar variables que portan información valiosa y relevante para la solución de un problema dado. Aunque este algoritmo es computacionalmente caro (2n subconjuntos para n valores) y es aplicable generalmente sólo a problemas de clasificación y diagnóstico; es uno de los más usados en aplicaciones reales.

El algoritmo consiste en probar primero el atributo con mayor ganancia de información y menos entropía, de esta forma particiona los ejemplos. Cada subconjunto generado es un nuevo problema con menos ejemplos y un atributo menos, empleando de esta forma la recursividad. A continuación, un ejemplo de árbol ID3:

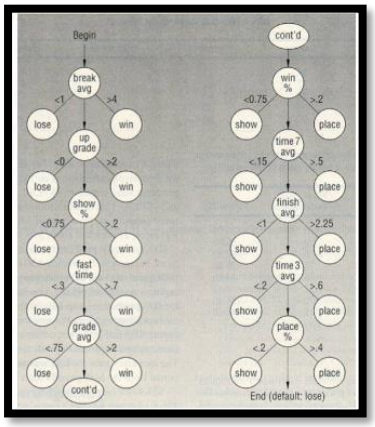

Figura 2: Ejemplo de árbol ID3. [11]

Este algoritmo es útil para el trabajo con variables discretas, pues de lo contrario implicaría la existencia de muchos más lugares para dividir los datos en los atributos continuos, y buscar el mejor valor para dividir puede llevar mucho tiempo, factor que no favorecería a una óptima solución del problema.

3.2.3 C4.5

El algoritmo C4.5 es una extensión del ID 3, y al igual que este último busca construir árboles de decisión a partir de un conjunto de datos de entrenamiento utilizando el concepto de entropía de la información.

En cada nodo del árbol, C 4.5 elige el atributo de los datos que divide de manera más efectiva su conjunto de muestras en subconjuntos enriquecidos en una clase u otra. El criterio de división es la ganancia de información normalizada es decir la diferencia de entropía. El algoritmo C 4.5 se repite en las sublistas particionadas. Es decir, el algoritmo C4.5 crea un árbol a partir de subconjuntos de casos extraídos del conjunto total de datos de entrenamiento.

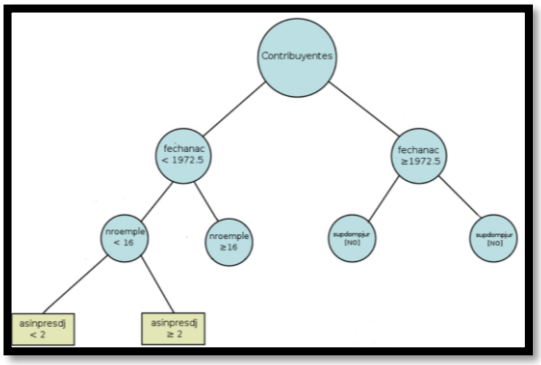

Figura 3: Ejemplo de árbol C4.5.[6]

Este algoritmo tiene algunas mejoras respecto al ID 3. Una de ellas es el manejo de atributos tanto continuos como discretos, para los valores continuos se crea un umbral y divide la lista en aquellos que sean mayores a este umbral y aquellos que sean menores o iguales a este. También, a diferencia del ID 3, permite un manejo de los datos de entrenamiento con valores faltantes, simplemente no empleándolos al momento de calcular la ganancia y la entropía. Otra diferencia es que debido a que el algoritmo realiza el procesamiento de los datos por ciclos generando nuevos árboles de acuerdo al subconjunto de datos del conjunto de entrenamiento, para la construcción del árbol no requiere terminar de clasificar los datos en todas las posibles categorías o clases posibles.

3.2.4 CART

Árbol de clasificación y regresión CART es una técnica de aprendizaje de árbol de decisión no paramétrica que produce árboles de decisión o regresión dependiendo si la variable dependiente es categórica o numérica respectivamente. En este algoritmo un nodo en el árbol de decisión solo puede dividirse en dos grupos. CART utiliza el índice de Gini como medida de impureza para seleccionar el atributo. El atributo con la mayor reducción

de impurezas se utiliza para dividir los registros del nodo. CART acepta datos con valores numéricos o categóricos y también maneja valores de atributos faltantes. Utiliza la poda de complejidad de costos y también genera árboles de regresión. Se trata de una partición binaria recursiva. En cada iteración se selecciona la variable predictiva y el punto de separación que mejor reduzcan la impureza.

El algoritmo consiste en considerar los datos pertenecientes a la muestra de entrenamiento. A partir de ella buscaremos dividir al conjunto de datos en dos conjuntos lo más homogéneos posibles, pero teniendo en cuenta únicamente el valor de una sola de las variables en cada paso.

La simplicidad de CART no solo produce un algoritmo de clasificación muy rápido para clasificar nuevas observaciones, sino que en general conduce a un modelo mucho más simple para explicar.

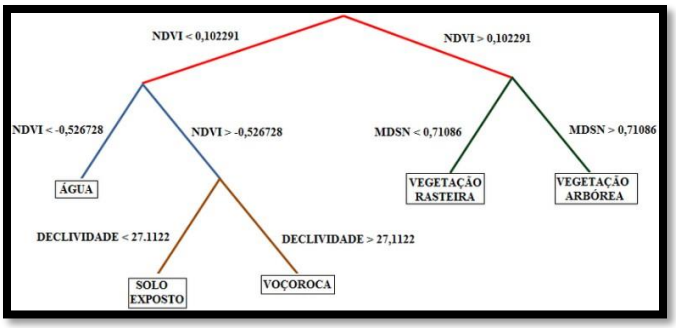

Figura 4: Árbol de decisión generado con el algoritmo CART [9]

4. DISEÑO DE LOS ALGORITMOS

4.1 Estructura de los datos

La estructura de datos utilizada en este trabajo es un árbol de decisión binario, el cuál tiene como objetivo clasificar un conjunto de datos para realizar una predicción.

La estructura del árbol de decisión binaria se asemeja a un mapa de los posibles resultados de un conjunto de decisiones relacionadas, por lo general comienza con un único nodo (nodo raíz el cual divide mejor los datos) y luego se hacen dos ramificaciones con los resultados posibles, de la misma manera cada resultado crea otros dos nodos adicionales que se ramifican en otras posibilidades hasta llegar al resultado definitivo.

En un árbol existen tres tipos de nodo: nodos de probabilidad que muestran las probabilidades de ciertos resultados; nodos de decisión que representan una condición y nodos terminales que muestran la decisión final de la ruta.

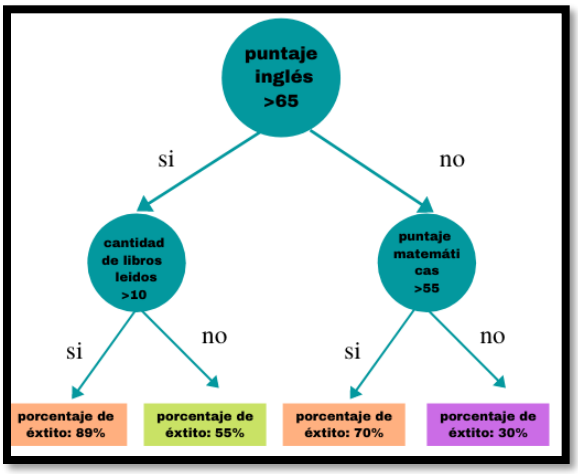

Figura 5: Árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos naranja representan a aquellos con una alta probabilidad de éxito, los verdes con una probabilidad media y los violeta con una baja probabilidad de éxito.

4.2.1 Entrenamiento del modelo

El algoritmo separa al conjuto de datos en dos grupos de acuerdo al criterio de ganancia de información, la condición que mayor ganancia de información aporte será la que determine el nodo de decision, de esta manera se va a repetir recursivamente hallando la variable que mejor divide el subgrupo.

Se construye el árbol organizando de forma descendente las variables, poniendo de nodo raíz a la variable que mejor divide los datos, creando así dos ramas. Análogamente sucede con los subgrupos hasta llegar al resultado de la predicción.

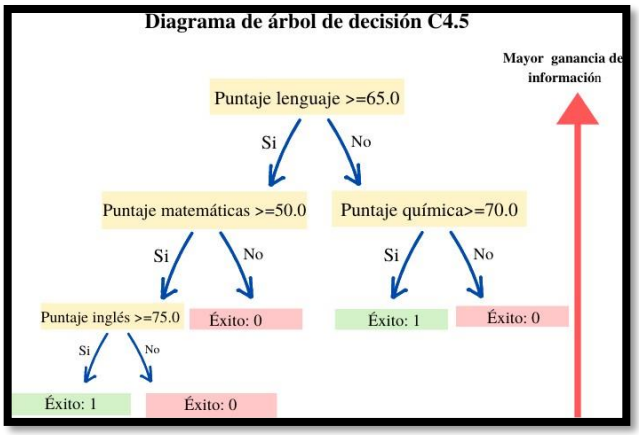

Figura 6: Entrenamiento de un árbol de decisión binario usando C4.5. En este ejemplo, mostramos un modelo para predecir el si el estudiante estará por encima o no del promedio en los resultados saber Pro.

4.2.2 Algoritmo de prueba

Este algoritmo recibe un nuevo conjunto de datos de prueba diferente al conjunto de datos de entrenamiento que utilizó el algoritmo para construir el árbol de decisión. Determina la etiqueta de cada persona a partir del al árbol de decisión, prediciendo si está o no por encima del promedio en los resultados saber Pro. Posteriormente compara los resultados que arrojó el algoritmo de entrenamiento con los resultados reales y calcula el porcentaje de precisión del árbol de decisión.

REFERENCIAS

1. Colaboradores de Wikipedia. Árbol de decisión alterno. ed. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/w/index.php?title=%C3%81rbol_de_decisi%C3%B3n_alternativo&oldid=1

2. Colaboradores de Wikipedia. Algoritmo ID3. ed. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/Algoritmo_ID3

3. Colaboradores de Wikipedia. Algoritmo C4.5. ed. Wikipedia, la enciclopedia libre. https://en.wikipedia.org/wiki/C4.5_algorithm

4. Daza, A. Un modelo basado en árboles de decisión para predecir la deserción estudiantil en la Educacion Superior Privada.: 2017. http://UCV - SCIENTIA, ISSN 2077-172X, Vol. 8, Nº. 1, 2016, págs. 59-73. Accessed: 2020-.

5. Eckert, K. and Suénaga, R. 2015. Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. Scielo. http://dx.doi.org/10.4067/S0718-50062015000500002.

6. Lopez, Rodrigo. 2014. Ingeniería de explotación de la información aplicada a la investigación tributaria fiscal. ResearchGate. https://www.researchgate.net/figure/Arbol-de-induccion-por-algoritmo-TDIDT-C45_fig2_262143491

7. Morales, E. and Escalante, H. Árboles de decisión. INAOE. https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/sbl.pdf

8. Ramírez, Patricio E., and Grandón, Elizabeth E. 2018. Predicción de deserción de estudiantes en una universidad pública chilena mediante clasificación basada en árboles de decisión con parámetros optimizados. Scielo. https://dx.doi.org/10.4067/S0718-50062018000300003

9. Tedesco, A. and Felippe, A. 2015. Integración de OBIA, árboles de decisión y clasificación jerárquica para mapeo de garganta. ResearchGate. https://www.researchgate.net/figure/Figura-7-Arvore-de-decisao-gerada-com-o-algoritmo-CART_fig4_276204366

10. Timarán-Pereira, R., Caicedo-Zambrano, J., and Hidalgo-Troya, A. 2019. Árboles de decisiones para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11°. Uptc. https://revistas.uptc.edu.co/index.php/investigacion_duitama/article/view/9184

11. Vizcaino, P. 2008. Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de weka (waikato environment for knowledge analysis). Docplayer. https://docplayer.es/5563204-Aplicacion-de-tecnicas-de-induccion-de-arboles-de-decision-a-problemas-de-clasificacion-mediante-el-uso-de-weka-waikato-environment-for-knowledge.html