Predicción del éxito académico, en educación superior, utilizando árboles de decisión

Alejandro Villada Toro	Cristian Alzate Urrea	Miguel Correa	Mauricio Toro
Universidad Eafit	Universidad Eafit	Universidad Eafit	Universidad Eafit
Colombia	Colombia	Colombia	Colombia
avilladat@eafit.edu.co	calzateu@eafit.edu.co	macorream@eafit.edu.co	mtorobe@eafit.edu.co

RESUMEN

En este informe se pretende predecir el resultado de los estudiantes en las pruebas Saber Pro, mediante un árbol de decisión capaz de analizar los diferentes factores que pueden influir en la puntuación de dicho examen. Lo que se busca es encontrar y utilizar un árbol de decisión que permita predecir de manera efectiva los resultados.

Esta práctica es de suma de importancia porque va a permitir identificar los diferentes factores que influyen en este resultado y así tomar acciones para mejorarlos. Este problema se puede relacionar fácilmente con otros estudios cuyo objetivo es predecir los resultados de otras pruebas estandarizadas, como la prueba Saber 11.

Palabras clave

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

1. INTRODUCCIÓN

La prueba Saber Pro, es de gran importancia, ya que por medio de ella se puede observar y analizar los conocimientos y habilidades que adquirieron los estudiantes durante su proceso de aprendizaje, evaluando las capacidades que debería poseer. En este orden de ideas, es muy importante la predicción de esos resultados, ya que, por medio de ello, se puede conocer los factores que afectan el desempeño de un estudiante, y así, después de conocerlos, crear estrategias para mejorar los resultados.

La implementación de esas estrategias puede mejorar considerablemente la calidad educativa en las universidades y, en consecuencia, el rendimiento de los estudiantes, lo que les permitirá tener un mayor éxito en la inserción laboral y mayores opciones de estudios posteriores.

1.1. Problema

El problema al cual nos enfrentamos es la elaboración e implementación de un árbol de decisión que nos permita predecir de manera eficiente y efectiva el rendimiento de los estudiantes en la prueba Saber Pro.

La solución de este problema nos conducirá a la creación de estrategias que favorezcan el desempeño en dicha prueba, con lo cual, se contribuirá al mejoramiento de la educación y del éxito académico de los pregrados.

1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad ya que los "árboles de decisión son un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Es además un diagrama que representa en forma secuencial, condiciones y acciones" [16] es este último hecho es lo que lo vuelve fácil de explicar y entender. Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad, ya que, en el caso de los bosques aleatorios, "a diferencia de los árboles de decisión, la clasificación hecha por estos es difícil de interpretar" [17], sucede algo similar con los demás métodos mencionados.

Para llegar la solución a este problema hemos elegido el árbol de decisión ID3, ya que permite una fácil creación y análisis de este. Además, es sencilla la recopilación de datos e información sobre el algoritmo. El árbol clasifica los datos mediante una serie de atributos elegidos por medio del concepto de ganancia de información y entropía, que se obtiene calculando la impureza de Gini, la cual nos dice qué tan mezclados están los datos, y mientras más baja sea, se tendrá una mayor precisión.

1.3 Estructura del artículo

En lo que sigue, en la sección 2, presentamos el trabajo relacionado con el problema. Más adelante, en la sección 3, presentamos los conjuntos de datos y métodos utilizados en esta investigación. En la sección 4, presentamos el diseño del algoritmo. Después, en la sección 5, presentamos los resultados. Finalmente, en la sección 6, discutimos los resultados y proponemos algunas direcciones de trabajo futuras.

2. TRABAJOS RELACIONADOS

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

El objetivo del estudio era detectar factores asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media, que presentaron las pruebas Saber 11° en los años 2015 y 2016.[11]

El algoritmo que fue utilizado para realizar la construcción del árbol y encontrar la solución al problema fue el J48 de la herramienta WEKA. Se escogió este algoritmo porque para

esta clase de estudios, es uno de los más utilizados y populares gracias a su simplicidad y facilidad para entender los resultados finales. Además, se escogió WEKA por ser de software libre y tener herramientas muy prácticas que facilitan la resolución del proyecto.[11]

Los datos que se utilizaron para la realización del estudio fue toda la información socioeconómica, institucional y académica, las cuales están disponibles en la base de datos del ICFES, de los estudiantes que presentaron las pruebas Saber 11° entre los años 2015 y 2016.[11]

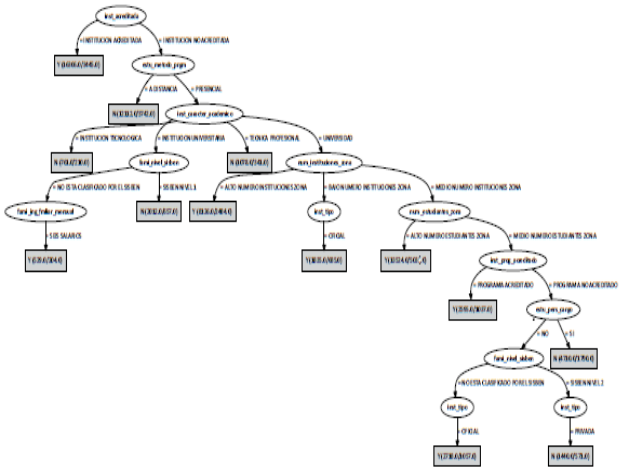
Al analizar los datos obtenidos se llegó a la conclusión de que el 67% (711.116) de los registros se clasifico de correcta, mientras que el restante 33% (350.564) fue clasificado de manera incorrecta. [11]

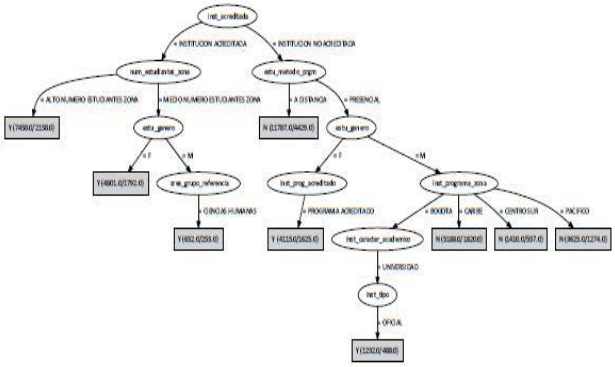
2.2 Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional

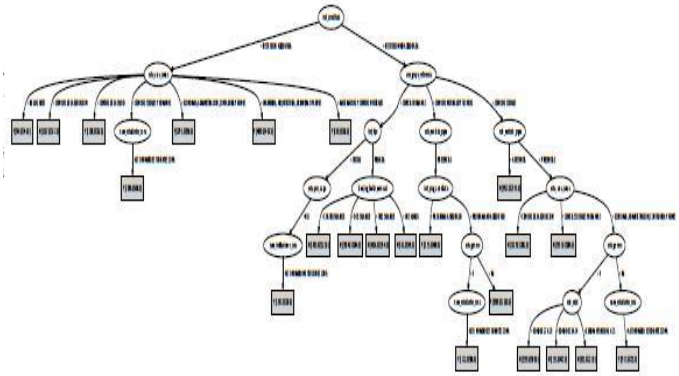
La meta del estudio era descubrir los factores que influyen en el desempeño de los futuros profesionales en las diferentes competencias de las pruebas Saber Pro, analizando las circunstancias económicas, institucionales, educativas y sociodemográficas.[1]
El algoritmo que les permitió realizar este análisis fue el J48, de la herramienta WEKA, el cual implementa el algoritmo C4.5. Dicho código funciona con el criterio de ganancia de información, que permite elegir el atributo con el cual se haga la mayor y más efectiva división de los elementos de un conjunto de datos.[1]
Una vez terminados los árboles de decisión para cada una de las materias, se obtuvieron las siguientes precisiones:

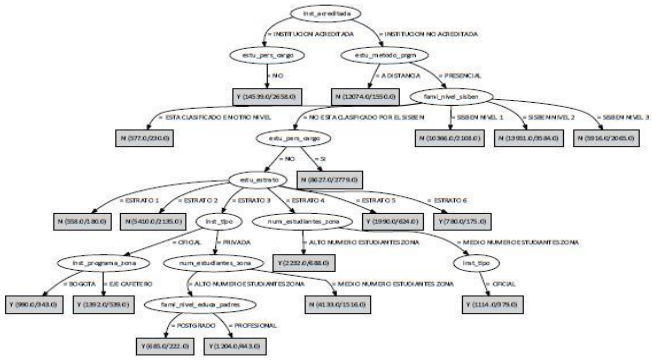
- Lectura crítica: 64.686%
- Comunicación escrita: 59.4407%
- Razonamiento cuantitativo: 66.5063%
- Inglés: 71.7925%

Los respectivos árboles de decisión son los siguientes:









[1]
Debería mencionar el problema que resolvieron, el algoritmo que usaron, la precisión que lograron y la citación.

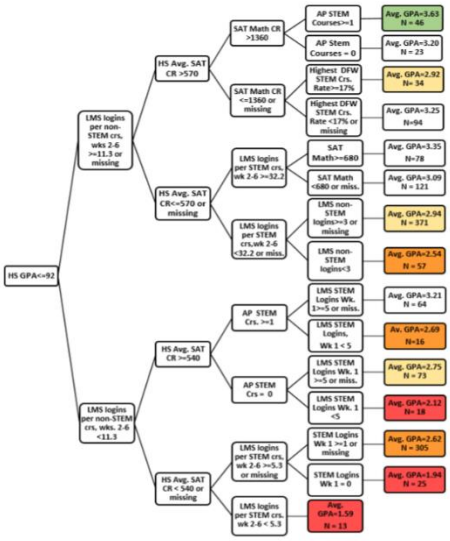
2.3 Uso de la minería de datos para predecir el rendimiento de los estudiantes universitarios de primer año

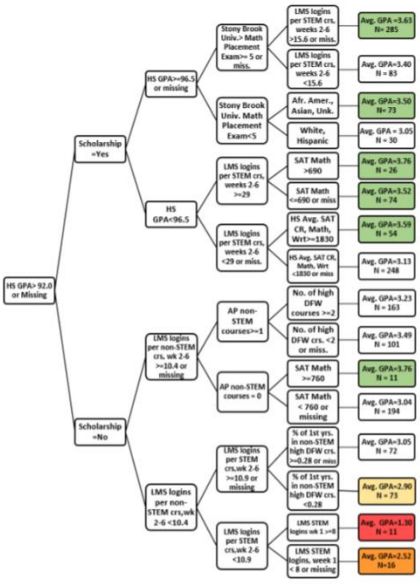
El objetivo del estudio es crear un modelo de para predecir, que, basándose en una serie de factores, consiga pronosticar de manera precisa y eficiente el rendimiento de los estudiantes universitarios de primer año.[4]

Para predecir estos resultados se utilizan factores como el resultado de ingreso a la universidad y el promedio escolar; por ejemplo, los estudiantes que ingresan a la universidad por primera vez con puntajes altos en el SAT y un GPA escolar

alto, pueden tener un mejor rendimiento que aquellos que ingresaron con un puntaje menor.[4]

Los datos en los que se basó el proyecto en un principio fueron el género, la etnia y el área geográfica; luego fueron añadidos los puntajes del GPA escolar y el SAT. El algoritmo que fue utilizado para la construcción e implementación del árbol fue el CART. El árbol resultante fue dividido en dos partes, para los que tenía un GPA menos o igual a 92 y los que lo tenían mayor a 92, quedando de la siguiente manera:





[4]

2.4 Predicción del rendimiento académico aplicando técnicas de minería de datos

La finalidad de la investigación es predecir el rendimiento de los estudiantes, mediante la aplicación de regresión logística, arboles de decisión, redes bayesianas y redes neuronales, usando los datos académicos de los estudiantes de Estadística General de la UNALM entre los años 2013 y 2014; comparar los resultados de los diferentes métodos y decidir cuál es el mejor.[10]

El algoritmo que fue utilizado para la creación y ejecución del árbol de decisión fue el C4.5 que se implementó en la herramienta WEKA con el algoritmo J48. La simplicidad y facilidad de la herramienta WEKA permitió un eficiente desarrollo del estudio, además una correcta interpretación de los resultados.[10]

La precisión con la funciono el árbol en el momento de analizar los datos, fue de un 71% de correcta clasificación.[10]

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

En lo que sigue, presentamos diferentes algoritmos usados para construir automáticamente un árbol de decisión binario, entre los que se incluyen los algoritmos ID3, C4.5, C5.0 y CART.

3.2.1 Algoritmo ID3

Este algoritmo fue creado por J. Ross Quinlan, tiene la capacidad de tomar decisiones de manera eficiente y con gran precisión. ID3 significa “inducción mediante árboles de decisión”.[14]

El funcionamiento del algoritmo ID3 consiste en determinar el árbol de decisión mínimo por medio de la mejor secuencia de preguntas de clasificación, que ayudan a segregar un conjunto. De esta manera la información permanece entendible, organizada y clara. [14]

La cantidad de salidas de este tipo de árbol varía dependiendo la función, por ejemplo, para una función booleana el árbol da dos salidas representadas por “si” o “no”; para funciones distintas, la cantidad de salidas puede ser mayor. [14]

Para su correcta ejecución, el ID3 se apoya en diversas técnicas matemáticas y estadísticas. Los atributos seleccionados, de los cuales se basan las preguntas, deben ser aquellos que disminuyan en mayor medida la entropía. La entropía no es más que la medida de desorden o incertidumbre de los datos. [14]

El criterio con el que se selecciona el atributo que disminuya en mayor medida la entropía es el concepto de ganancia de información, el cual es una medida de segregación que clasifica los atributos. [14]

El esquema del algoritmo ID3, definido por su autor, cuenta con la siguiente estructura:

1.Calcular la entropía para todas las clases.

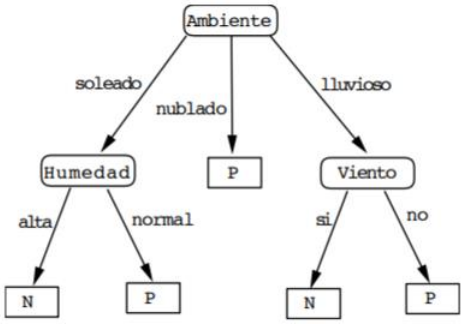
2.Seleccionar el mejor atributo basado en la reducción de la entropía.

3.Iterar hasta que todos los objetos sean clasificados.

El algoritmo ID3 sigue una estrategia de hill-climbing, es decir, no utiliza backtracking. Al no realizar la “vuelta atrás” y al tender por la construcción de árboles pequeños, pero ganancia de información puede llegar menos complejo que otras soluciones. [14]

La complejidad de construir por medio de este algoritmo está dada por: $O(mn\log n) + O(n(\log n)^2)$ donde m es el número de datos y n es el número de atributos. [14]

Un árbol puede ayudarnos a conocer si salir o no a jugar dependiendo del estado del clima. [14]



[14]

3.2.2 Algoritmo C4.5

C4.5 es un algoritmo creado por Ross Quinlan usado para generar un árbol de decisión. C4.5 es la versión mejorada del algoritmo ID3 desarrollado previamente por Quinlan. C4.5

está casi siempre definido como un clasificador estadístico, ya que los árboles generados por medio de él pueden ser usados para clasificar. [15]

Al igual que ID3, C4.5 se aprovecha del concepto entropía de información para construcción del árbol, partiendo desde un conjunto de datos llamados datos de entrenamiento. Los datos de entrenamiento son un grupo de ejemplos ya han sido clasificados. Cada grupo de ejemplos es un conjunto de atributos o características de este. Los datos de entrenamiento pueden ser aumentados con un conjunto cuyos elementos son la clase donde pertenece cada muestra. [15]

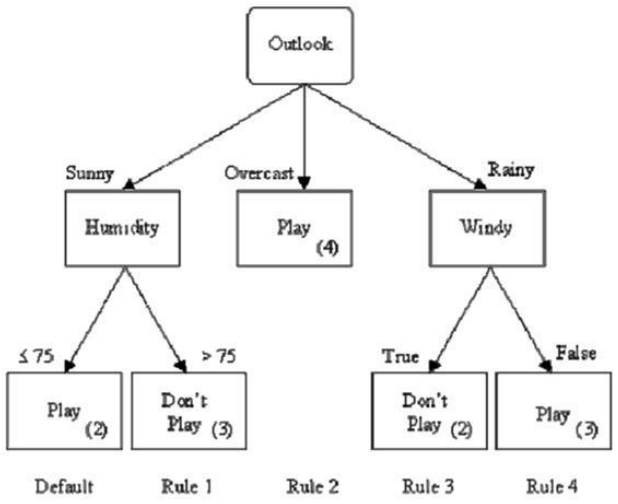
Para cada nodo del árbol, C4.5 selecciona un atributo de los datos que mejor dividen el conjunto de muestras en subconjuntos en una clase u otra. La manera de elegir los atributos es por medio de la ganancia de información, es decir, la diferencia de entropía. El atributo con la mayor ganancia de información se elige el parámetro de decisión. El algoritmo C4.5 divide de manera recursiva a subconjuntos más pequeños. [15]

El algoritmo C4.5 se ejecuta de la siguiente manera:

1. Comprobar los casos base.
2. Para cada atributo encontrar la ganancia relacionada al mismo.
3. Identificar el atributo con la mayor ganancia de información.
4. Crear un nodo de decisión que divida a dicho atributo.
5. Repetir en los subconjuntos obtenidos por la división del atributo y agregarlos como hijos del nodo.

Al ser un algoritmo basado en el ID3, su complejidad es muy similar, casi igual. [15]

Un ejemplo de algoritmo C4.5 es:



[15]

3.2.3 Algoritmo C5.0

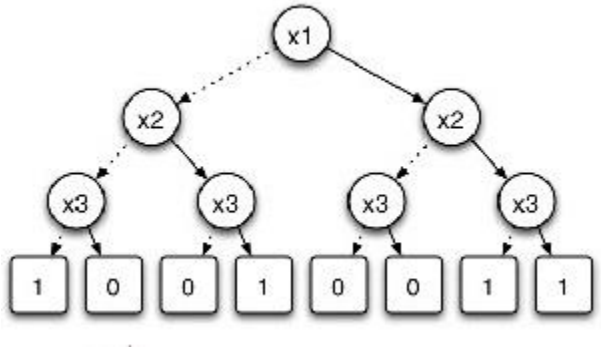
El algoritmo C5.0 es una mejora con respecto a su antecesor, el C4.5, y sigue conservando su estructura principal, pero con algunas optimizaciones. A su vez, este algoritmo fue desarrollado por el mismo personaje de los anteriores, J. Ross Quinlan. [15]

Además, su principal funcionalidad es generar árboles de decisión sencillos para las divisiones que se pueden hacer en un conjunto de datos. Entre sus puntos fuertes se encuentran el análisis de problemas en los que hay valores perdidos o muchas entradas; también, son bastante rápidos y fáciles de comprender, para así poderlos aplicar de la mejor forma.[2]

Sus mejoras con respecto al C4.5 son:

- Velocidad - C5.0 es significativamente más rápido que el C4.5 (varios órdenes de magnitud)
- El uso de memoria - C5.0 es más eficiente que el C4.5
- Árboles de decisión más pequeños - C5.0 obtiene resultados similares a C4.5 con árboles de decisión mucho más pequeños.
- Soporte para boosting - Boosting mejora los árboles y les da una mayor precisión.
- Ponderación - C5.0 le permite ponderar los distintos casos y tipos de errores de clasificación.
- Winnowing - una opción automática de C5.0 consiste en aplicar un algoritmo de clasificación (algoritmo Winnow) a los atributos para eliminar aquellos que sean de poca ayuda. [15]

Un esquema del funcionamiento del algoritmo es el siguiente:[9]



3.2.4 Algoritmo CART

El árbol CART, también llamado Classification And Regression Trees (Árboles de Clasificación y de Regresión), fue elaborado por Breiman et al. (1984). Este es un algoritmo que permite realizar problemas tanto de regresión como de clasificación.[3]

Para calcular la medida de impureza, este algoritmo se usa el índice de impureza Gini, el cual está dado por:

$$G(A_i) = \sum_{j=i}^{M_i} p(A_{ij}) G(\frac{C}{A_{ij}})$$

Donde $G\left(\frac{C}{A_{ij}}\right) = -\sum_{j=i}^{M_i} p(\frac{C_k}{A_{ij}})(1 - p\left(\frac{C_k}{A_{ij}}\right))$

[3]

Siendo

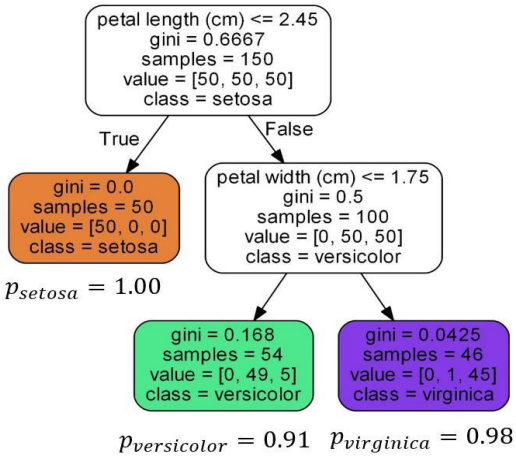
- Aij es el atributo empleado para ramificar el árbol,
- J es el número de clases,
- Mi es el de valores distintos que tiene el atributo Ai
- p(Aij) constituye la probabilidad de que Ai tome su j-ésimo valor
- p(Ck/Aij) representa la probabilidad de que un ejemplo sea de la clase Ck cuando su atributo Ai toma su j-ésimo valor.[3]

El mínimo valor que puede alcanzar la impureza de GINI es cero, sucede cuando los datos son completamente homogéneos, es decir, pertenecen todos a la misma clase, por el contrario, la máxima cantidad se obtiene cuando los datos son totalmente heterogéneos.[3]

La ejecución del algoritmo se da por los siguientes tres pasos:

- construcción del árbol saturado
- Elección del tamaño correcto
- Clasificación de nuevos datos a partir del árbol construido[13]

Un ejemplo de árbol de decisión generado por este algoritmo es:



[5]

4. DISEÑO DE LOS ALGORITMOS

En lo que sigue, explicamos la estructura de los datos y los algoritmos utilizados en este trabajo. La implementación del algoritmo y la estructura de datos se encuentra disponible en GitHub[1].

4.1 Estructura de los datos

Un árbol de decisión binario es una estructura de datos que es utilizado para representar de manera clara operaciones booleanas, en las que las únicas opciones son true o false[19], y de allí se deriva el nombre de binario, ya que solo existen dos opciones. Para nuestro caso el árbol diría true si predice que el estudiante analizado obtendrá un puntaje superior al promedio, y diría false si sucede lo contrario.

[1]https://github.com/avilladat/ST0245-001/tree/master/proyecto/codigo

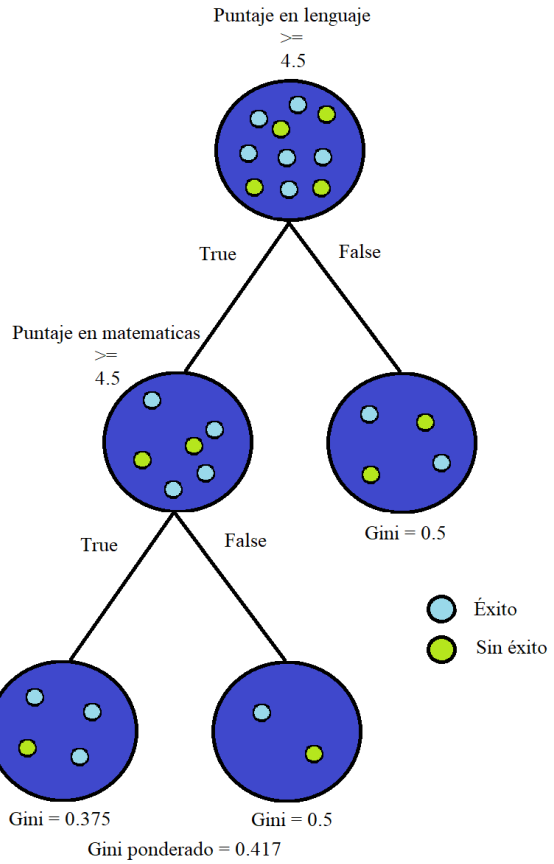


Figura 1: Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los círculos azules representan aquellos estudiantes que obtuvieron un resultado superior al promedio, los verdes los que obtuvieron un resultado inferior al promedio. En este ejemplo, mostramos un modelo con los puntajes de lenguaje y matemáticas.

4.2 Algoritmos

El primer algoritmo permite que el árbol clasifique los datos mediante una serie de atributos elegidos por medio del concepto de ganancia de información y entropía, que se obtiene calculando la impureza de Gini, la precisión obtenida es mayor si esta impureza es menor, si es igual a cero la precisión será del 100%.

El segundo algoritmo es el encargado de leer los datos que se le ingresen y procesarlos con el árbol de decisión ya entrenado.

4.2.1 Entrenamiento del modelo

La generación del árbol de decisión binario funciona en base al concepto de ganancia de información y entropía, que se refiere a que tan mezclados están los datos en uno de los nodos; esto se puede conocer fácilmente por medio de la impureza de Gini. Los atributos que generen impurezas de Gini bajas serán los elegidos, en que caso de que el atributo genere una impureza de Gini muy alta será descartado y se pasara al siguiente.

En cada nodo de decisión, el árbol analizara atributo por atributo generando otros dos nodos, y después a estos nodos generados se les obtiene la impureza de Gini y la impureza de Gini ponderada, si estos valores son muy altos el atributo se descartará, se eliminarán los nodos y se prosigue analizando el siguiente atributo; si estos valores son bajos se prosigue con el siguiente atributo.

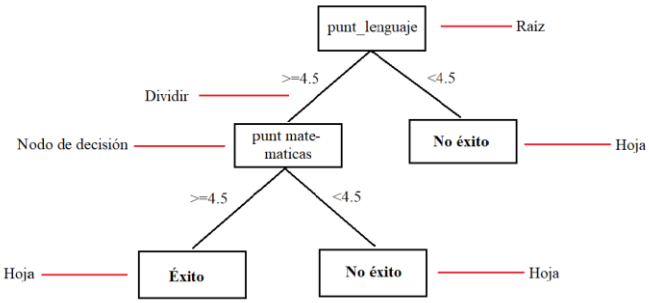


Figura 2: Entrenamiento del árbol de decisión binario para predecir los resultados de la prueba saber pro a partir de los resultados de la prueba saber 11°. En este ejemplo, mostramos un modelo con los puntajes de lenguaje y matemáticas.

4.2.2 Algoritmo de prueba

El algoritmo de prueba es el encargado de ejecutar y utilizar el árbol para hacer predicciones sobre estudiantes. El árbol fue entrenado con los datos de 135000 estudiantes que presentaron la prueba saber 11 y la prueba saber pro entre los años 2012 y 2019. Para utilizarlo se ingresarían los datos de los nuevos estudiantes que hayan realizado la prueba saber 11° y se les desee predecir el resultado de la prueba saber pro; con esta información, el árbol analizara los atributos que considere pertinentes para realizar su predicción, y genera un resultado booleano.

REFERENCIAS

1. Alvarado- Pérez, J. C. Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Ediciones Universidad Cooperativa de Colombia, Bogotá, 2016

2. BOOKDOWN. Árbol C5.0. Recuperado Agoto 15, 2020. https://bookdown.org/content/2274/metodos-de-clasificacion.html#arbol-c5.0

3. BOOKDOWN. Árbol CART (Classification and Regression Trees). Recuperado Agosto 16, 2020. https://bookdown.org/content/2274/metodos

-de-clasificacion.html#arbol-cart-classification-and-regression-trees

4. Galambos N. USING DATA MINING TO PREDICT FRESHMEN OUTCOMES. Recuperado el 15 de agosto de 2020. https://www.stonybrook.edu/commcms/irpe/reports/_presentations/DataMiningFreshmanOutcomes_Galambos_paper2015_11_03.pdf
5. IArtificial.net. Árboles de Decisión con ejemplos en Python. Recuperado el 15 de agosto del 2020. https://iartificial.net/arboles-de-decision-con-ejemplos-en-python/
6. IBM Knowledge Center. Nodo C5.0. Recuperado agosto 16, 2020. https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/c50node_general.html
7. Ibrahim Z, Rusli D. Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. Recuperado el 16 de agosto del 2020. https://www.researchgate.net/publication/228894873_Predicting_Students%27_Academic_Performance_Comparing_Artificial_Neural_Network_Decision_Tree_and_Linear_Regression
8. López B. ALGORITMO C4.5. Recuperado el 14 de agosto del 2020. http://www.itnuevolaredo.edu.mx/takeyas/apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf
9. Machine Learning con R**.** Ensamble de RandomForest + SVM + C5. Recuperado Agosto 14, 2020. http://apuntes-r.blogspot.com/2015/07/ensamble-de-randomforest-svm-c5.html
10. Menacho C. Predicción del rendimiento académico aplicando técnicas de minería de datos. Recuperado el 16 de agosto de 2020. https://dialnet.unirioja.es/servlet/articulo?codigo=6171237
11. Timarán R, Caicedo J, Hidalgo A. Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. Recuperado el 15 de agosto del 2020. https://revistas.uptc.edu.co/index.php/investigacion_duitama/article/view/9184/7721
12. Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y
13. Universitat de València. Árboles de clasificación y regresión. Recuperado el 14 de agosto del 2020. https://www.uv.es/mlejarza/actuariales/tam/arbolesdecision.pdf
14. Wikipedia. Árbol de decisión (modelo de clasificación ID3). Recuperado el 14 de agosto de 2020. https://es.wikipedia.org/wiki/Árbol_de_decisión_(modelo_de_clasificación_ID3)#Representación_de_un_Árbol_de_decisión
15. Wikipedia.C4.5. Recuperado el 15 de agosto del 2020. https://es.wikipedia.org/wiki/C4.5
16. EcuRed. Árbol de decisión. Recuperado el 10 de octubre de 2020. https://www.ecured.cu/Árbol_de_decisión
17. Wikipedia. Random Forest. Recupera el 10 de octubre de 2020. https://es.wikipedia.org/wiki/Random_forest
18. GitHub. Proyecto. Recuperado el 10 de octubre de 2020. https://github.com/avilladat/ST0245-001/tree/master/proyecto/codigo
19. Wikipedia. Diagrama de decisión binario. Recuperado el 10 de octubre de 2020. https://es.wikipedia.org/wiki/Diagrama_de_decisión_binario