

## Predicción del éxito académico en educación superior utilizando arboles de decisión

Juan Pablo Madrid Florez  
Universidad Eafit  
Colombia  
jpmadridf@eafit.edu.co

Abelino Sepúlveda Estrada  
Universidad Eafit  
Colombia  
asepulvede@eafit.edu.co

Andrés Gomez Arango  
Universidad Eafit  
Colombia  
agomeza10@eafit.edu.co

Mauricio Toro  
Universidad Eafit  
Colombia  
mtorobe@eafit.edu.co

### RESUMEN

Unas de las mayores preguntas que nos hacemos muchos antes de empezar la vida universitaria es saber si tendremos éxito académico en nuestra educación superior y si tendríamos un buen puntaje en las pruebas Saber Pro, pero esto es algo que no podemos saber a ciencia cierta. Lo que buscamos es seguir una serie de patrones que se relacionen con la vida académica del estudiante, tales como el desempeño en las pruebas saber 11, estrato socioeconómico, carrera, edad, género, entre otros factores, para predecir el éxito de las personas en dichas pruebas, definiendo si quedaría o no por encima del promedio.

Meditar este problema es muy importante porque saber cómo nos iría en las pruebas saber Pro, nos alentaría a sacar ese puntaje e inclusive mejorarlo, dependiendo de la situación, ya que pruebas como estas son factores influyentes en momento posgrado, es decir, ayudaría a tener un buen empleo y salario. Para hacer posible la predicción de este problema estaremos implementando árboles de decisión.

### PALABRAS CLAVE

Estructura de datos; ArrayList; Complejidad; Tiempos de Ejecución; memoria; operaciones; notación O

### PALABRAS CLAVE DE LA CLASIFICACIÓN DE LA ACM

Software and its engineering → software organization and properties → Operating systems → Memory management → Virtual memory

### 1. INTRODUCCIÓN

Hoy en día, se conoce que Colombia es el segundo país con la mayor deserción universitaria de Latinoamérica y esto es bastante preocupante ya que es casi la mitad de los jóvenes que cursan educación superior en el país. Este problema se origina desde la educación secundaria ya que no se prepara correctamente a los jóvenes dado varios factores, pero uno de los mas importantes es el socioeconómico.

Por lo mencionando anteriormente es necesario buscar una solución que nos pueda ayudar a predecir el éxito académico en las pruebas saber pro, para esto se utilizaran arboles de

decisión que mediante un análisis a las pruebas saber 11 y a variables socioeconómicas podremos definir el éxito de los estudiantes. Este es el objetivo por cumplir de este proyecto.

### 2. PROBLEMA

El problema que debemos resolver es hacer una predicción de éxito en las pruebas saber pro mediante el análisis de datos como las pruebas saber 11 y variables socioeconómicas en arboles de decisión y poder tener resultados certeros.

Resolver esta problemática sería de gran ayuda para saber en que falla actualmente el sistema de educación del país y hacer mejor para que la calidad de educación superior será de gran nivel y también la educación secundaria, a su vez ayudaría para acabar con la deserción universitaria tan grande que se vive en el país.

### 3. TRABAJOS RELACIONADOS

#### 3.1 Algoritmo ID3

Se usa principalmente para la búsqueda de hipótesis o reglas en el árbol, dado un conjunto de ejemplos.

El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno siendo atributos, en el que uno de ellos, (el atributo a clasificar) es el objetivo, el cual es de tipo binario (positivo o negativo, sí o no, válido o inválido, etc.).

De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo. Como se muestra en la figura 1.

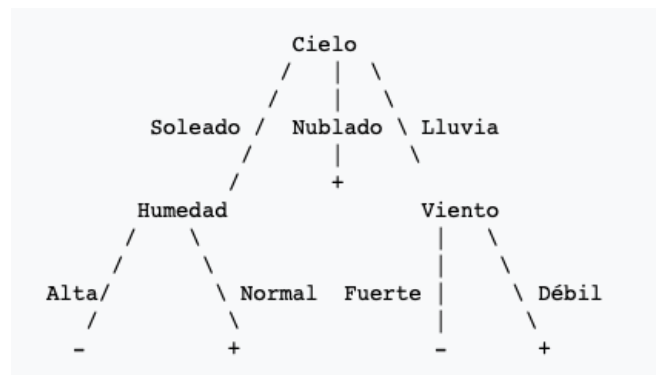


Figura 1. Algoritmo ID3. Ejemplo.

Los elementos son:

Nodos: Los cuales contendrán atributos.

Arcos: Los cuales contienen valores posibles del nodo padre.

Hojas: Nodos que clasifican el ejemplo como positivo o negativo.

### 3.2 Árbol de decisión alternativo

Un Árbol de decisión alternativo es un método de clasificación proveniente del aprendizaje automático conocido en inglés como Alternating Decision Tree (ADTree).

Un ADTree consiste en una alternancia de nodos de decisión, que especifican una condición determinante, y predicción, que contienen un solo número.

Una instancia es clasificada por un ADTree siguiendo todos los caminos para que todos los nodos de decisión sean verdaderos, y suma algún nodo de predicción que es recorrida.

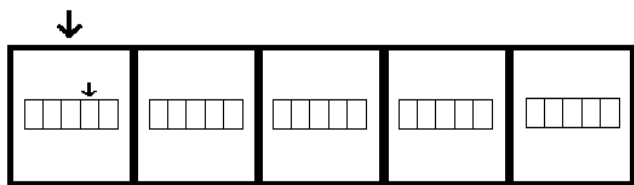
### 3.3 C 4.5

C4.5 es un algoritmo usado para generar un árbol de decision desarrollado por Ross Quinlan. C4.5 es una extensión del algoritmo ID3 desarrollado anteriormente por Quinlan. Los árboles de decisión generados por C4.5 pueden ser usados para clasificación, y por esta razón, C4.5 está casi siempre referido como un clasificador estadístico.

### 3.4 CART

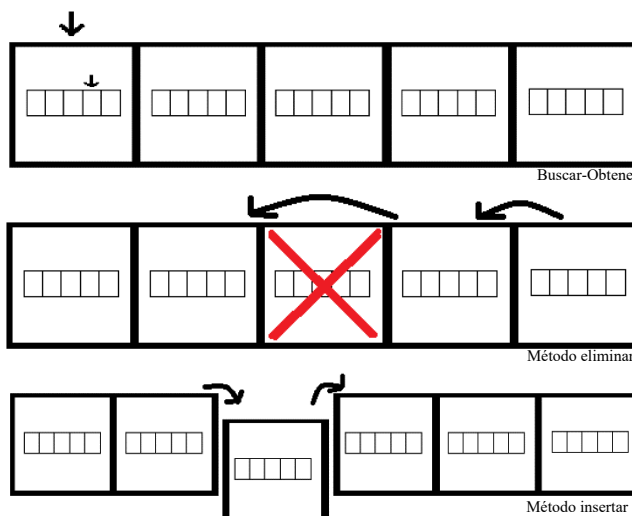
Aprendizaje basado en árboles de decisión utiliza un árbol de decisión como un modelo predictivo que mapea observaciones sobre un artículo a conclusiones sobre el valor objetivo del artículo. Es uno de los enfoques de modelado predictivo utilizadas en estadísticas, minería de datos y aprendizaje automático.

## 4. ESTRUCTURA DE DATOS



**Gráfica 1:** Se implementa un ArrayList de ArrayList, en el cual cada posición del ArrayList principal es una persona diferente y dentro de cada posición hay un nuevo ArrayList en el cual cada posición hace referencia a un dato de la persona

### 4.1 OPERACIONES DE LA ESTRUCTURA DE DATOS



**Gráfica 2:** Análisis de métodos como buscar-acceder, eliminar e insertar

### 4.2 Criterios de diseño de la estructura de datos

Elegimos esta estructura de datos debido a que es una que puede dar solución de una buena manera a los objetivos que se tienen en el trabajo.

Además, este es un método eficiente para el almacenamiento y manejo de datos de forma eficaz comparado a otro tipo de estructuras.

Otra de las ventajas que presenta esta estructura de datos es permite un manejo muy bueno de los datos, los métodos para buscar y eliminar datos es muy eficaz ya que no requiere de una complejidad alta.

### 4.3 Análisis de Complejidad

ArrayaList de ArrayList	Complejidad
Eliminar elemento	$O(n*m)$
Añadir elemento	$O(n*m)$
acceder	$O(1)$
Buscar elemento	$O(n)$

**Tabla 1:** Tabla que reporta el análisis de complejidad(donde n es el número de personas y m es la cantidad de datos)

### 4.4 TIEMPOS DE EJECUCIÓN

Dataset	1	2	3	4	5	6	Promedio
Runtime (s)	0.48	1.44	2.21	2.86	6.55	4.82	3.06

Datasets: 1(1500), 2(45000), 3(75000), 4(105000), 5(135000), 6(57765)

#### 4.5 MEMORIA

Dataset	1	2	3	4	5	6	Promedio
Memoria(mb)	67.5	197.93	328.71	458.68	589.6	252.91	315.90

Datasets: 1(1500), 2(45000), 3(75000), 4(105000), 5(135000), 6(57765)

#### 4.6 ANÁLISIS DE LOS RESULTADOS

Es evidente en ambas tablas que a mayor número de datos, el tiempo de creación de la estructura crece y el espacio que esta ocupa en la memoria también. También podemos ver que la estructura que hemos escogido, hasta ahora, ha demostrado un buen rendimiento, pues para una buena cantidad de datos (5700\*78) resulta un tiempo de creación justo

Dataset	Runtime(s)	Memoria(mb)
1(1500)	0.48	67.5
2(45000)	1.44	197.93
3(75000)	2.21	328.71
4(105000)	2.86	458.68
5(135000)	6.55	589.6
6(57765)	4.82	252.91
Promedio	3.06	315.90

**Table 4:** Análisis de los resultados obtenidos con la implementación de la estructura de datos

3. Wikipedia. Arbol de decisiones alternativos. Retrieved February 7, 2020. [https://es.wikipedia.org/wiki/Árbol\\_de\\_decisión\\_alternativo](https://es.wikipedia.org/wiki/Árbol_de_decisión_alternativo)
4. Adobe Acrobat Reader 7, Asegúrense de justificar el texto. <http://www.adobe.com/products/acrobat/>.
5. Fischer, G. and Nakakoji, K. Amplifying designers' creativity with domainoriented design environments. in Dartnall, T. ed. Artificial Intelligence and Creativity: An Interdisciplinary Approach, Kluwer Academic Publishers, Dordrecht, 1994, 343-364.

#### REFERENCIAS

1. Casas, P. El problema no es solo plata: 42 % de los universitarios deserta. Retrieved February 6,2020, from El espectador. <https://www.elspectador.com/noticias/educacion/el-problema-no-es-solo-plata-42-de-los-universitarios-deserta-articulo-827739>
2. Wikipedia. Algoritmo ID3. Retrieved February 7, 2020. [https://es.wikipedia.org/wiki/Algoritmo\\_ID3](https://es.wikipedia.org/wiki/Algoritmo_ID3)