Predicción del éxito académico en educación superior utilizando árboles de decisión

Mauricio Toro, Miguel A. Correa

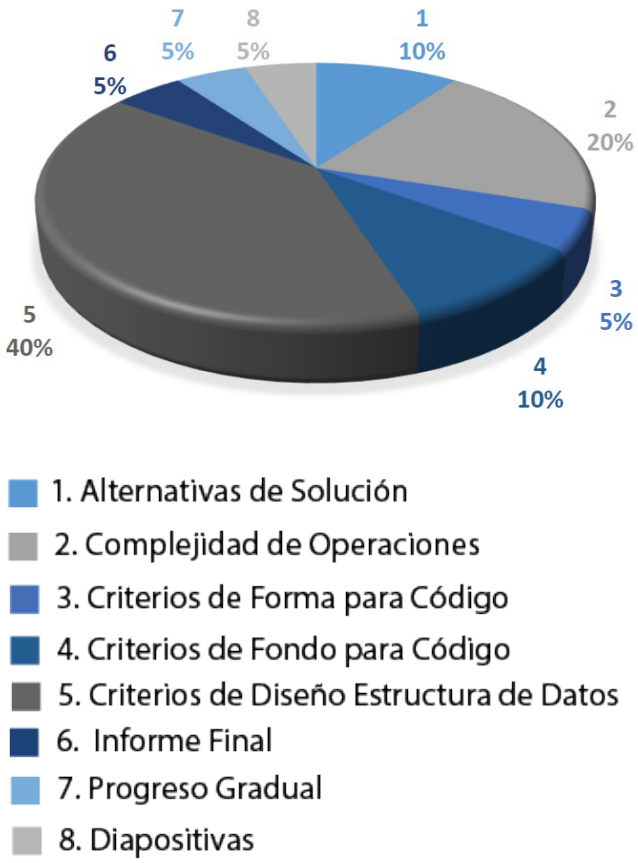
IMPORTANTE: El tema de árboles de decisión no es un tema del curso, no lo vemos en clase, debe ser consultado por cuenta de cada grupo. Por eso se eligió como proyecto.

Consideraciones iniciales

	Trabajo **en parejas**		Puntuación extra si lo escriben y sustentan en **inglés**
	Usar **plantilla ACM**		Entregar informe en **PDF** y código en **GIT**
	Informe **máximo en 4 páginas**		Detalles completos en **"Guía para la realización del proyecto final"**

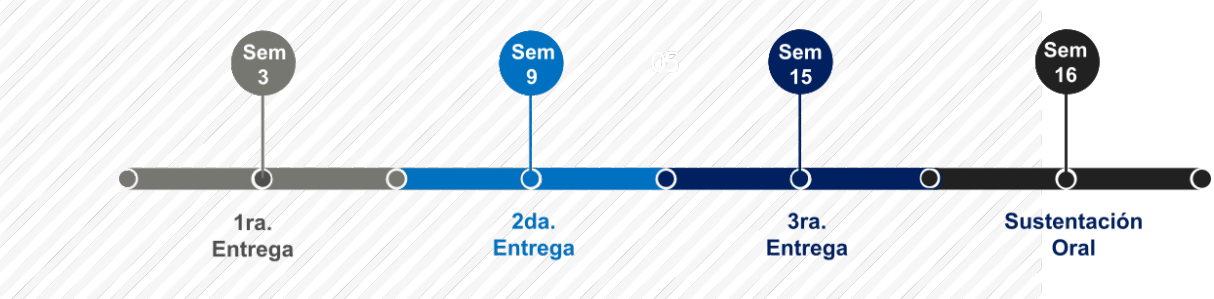
DOCENTE MAURICIO TORO BERMÚDEZ
Teléfono: (+57) (4) 261 95 00 Ext. 9473. Oficina: 19 - 627
Correo: mtorobe@eafit.edu.co

Criterios de evaluación para el proyecto



DOCENTE MAURICIO TORO BERMÚDEZ
Teléfono: (+57) (4) 261 95 00 Ext. 9473. Oficina: 19 - 627
Correo: mtorobe@eafit.edu.co

Tiempos de entrega en semanas académicas



Rúbricas de calificación

Lean la Sección 9 de la *"Guía para la realización del proyecto final de Estructuras de Datos 1"*

Intercambio de archivos

DOCENTE MAURICIO TORO BERMÚDEZ
Teléfono: (+57) (4) 261 95 00 Ext. 9473. Oficina: 19 - 627
Correo: mtorobe@eafit.edu.co

1. Motivación

En un futuro cercano, el papel de la tecnología será un factor clave en el proceso de transformación digital de la educación en Colombia. Esta transformación se conoce como *Educación 4.0*. 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. No obstante, es poco lo que se ha logrado para predecir el éxito académico en educación superior. El éxito puede medirse de muchas formas; por ejemplo, la empleabilidad del egresado, el salario de los egresados, la felicidad del trabajo de los egresados, entre otros.

Para efectos de este proyecto, vamos a definir el éxito académico como la probabilidad que tiene un estudiante de obtener un puntaje total, superior al promedio de su cohorte, en la pruebas Saber Pro. Las pruebas Saber Pro son las pruebas estandarizadas que realiza el gobierno colombiano al final de la carrera, son la versión universitaria de las pruebas Saber 11.

2. Problema

De acuerdo a la motivación, el problema que tenemos es **diseñar un algoritmo, basado en árboles de decisión y en los datos del ICFES, para predecir si un estudiante tendrá un puntaje total, en las pruebas Saber Pro, por encima del promedio o no**. En particular, las variables académicas y sociodemográficas que tenemos 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 estrudiante, contamos 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.

Quienes deseen un reto adicional, después de diseñar la solución con árboles de decisión, utilicen bosques aleatorios.

3. Entradas y Salidas esperadas

Los algoritmos deben funcionar para conjuntos de datos con n filas y m columnas. De igual manera, la complejidad se debe expresar en términos de m y de n. No suponer que uno de ellos es constante. NO puede ser sólo en términos de n. ¡NO!

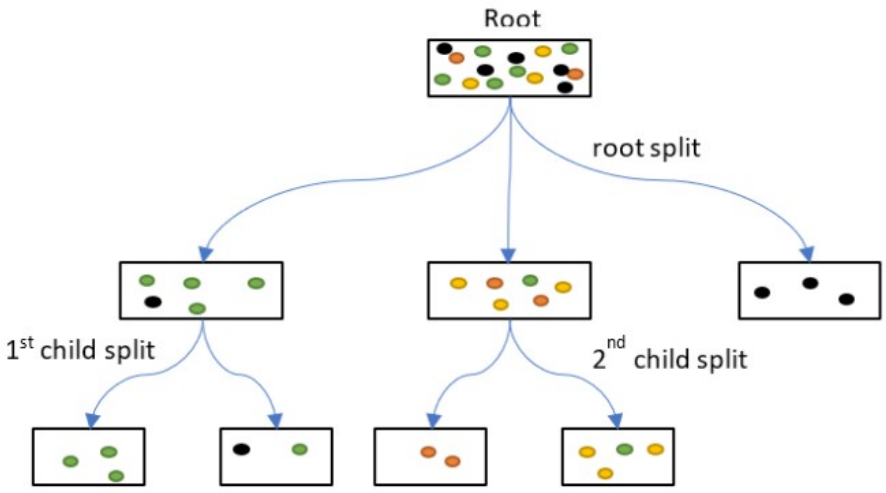
Para resolver este problema, se deben diseñar dos algoritmos. El primer algoritmo recibe como entrada un conjunto de datos, en un formato de *archivo separado por comas* (CSV), y se debe generar un árbol de decisión. A los datos que recibe este algoritmo se le llaman datos de entrenamiento, en Inglés, *training dataset*.

El segundo algoritmo recibe como entrada otro conjunto de datos (con el mismo formato que el anterior) y el árbol de decisión construído en el

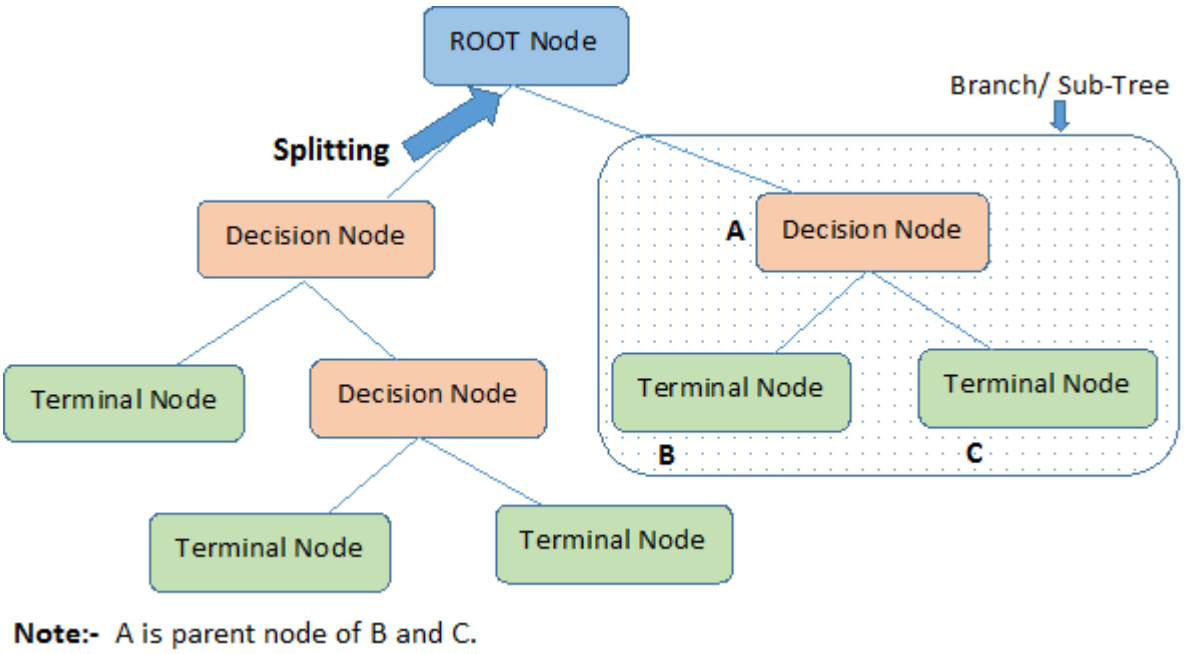
algoritmo anterior. A este conjunto de datos se le llama datos de validación, en Inglés, *test dataset*. Este algoritmo, para cada fila o registro, debe determinar la etiqueta, utilizando el árbol de decisión; es decir, si está o no por encima del promedio. Como el conjunto de datos ya tiene la respuesta, el algoritmo compara la respuesta que da el árbol de decisión con la respuesta que está en el conjunto de datos y determina cuál es el porcentaje de acierto del árbol de decisión.

Los conjuntos de datos se encuentran en GitHub en la carpeta *datasets*.

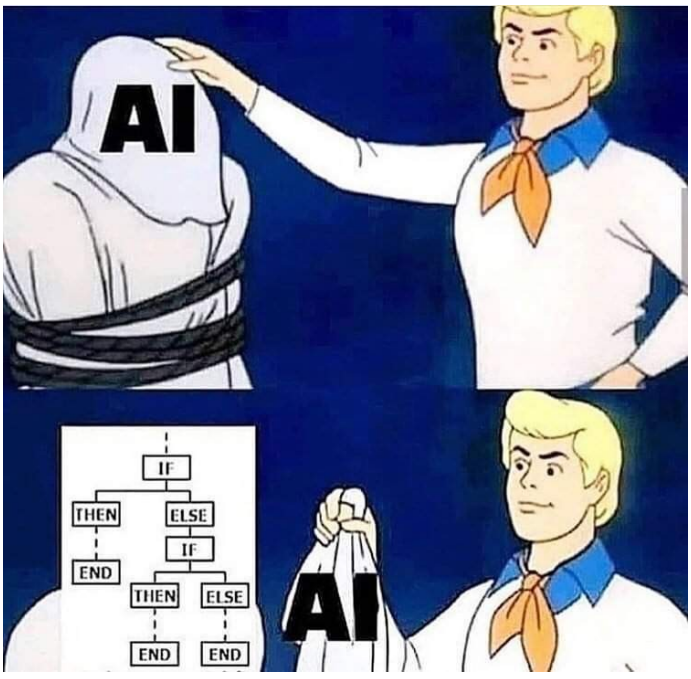
Como un ejemplo, en la Gráfica 2 se muestra cómo las decisiones que se toman en un árbol de decisión separan el conjunto de datos en subconjuntos donde son más homogéneas las clases, por ejemplo, donde están más diferenciados los registros en que un estudiante está por debajo del promedio de un estudiante que está por encima del promedio. Como otro ejemplo, la Gráfica 3 muestra cómo es la estructura de un árbol de decisión: Hay un nodo raíz, cuyos hijos son nodos de decisión y las hojas del árbol son nodos terminales.



Gráfica 2. Ejemplo del funcionamiento de un árbol de decisión



Gráfica 3. Ejemplo de la estructura de un árbol de decisión



Gráfica 4. Relación entre inteligencia artificial y árboles de decisión

5. Preliminares matemáticos

Los algoritmos para construir árboles de decisión, usualmente, funcionan de arriba a bajo, escogiendo, en cada paso, una variable que divide mejor el conjunto de elementos. Una de las métricas para escoger la "mejor" variable es la impureza de Gini.

Impureza de Gini

La impureza Gini es una medida de qué tanto un elemento escogido aleatoriamente del conjunto de datos se clasificará con la etiqueta incorrecta (por ejemplo, decir que estará por encima del promedio, cuando en realidad está por debajo del promedio). Para el caso de este proyecto, tenemos sólo dos etiquetas: La etiqueta 0 significa que el estudiante está por debajo del promedio y la etiqueta 1 significa que el estudiante está por encima del promedio. En consecuencia obtenemos la siguiente ecuación para calcular le impureza de Gini (I_G):

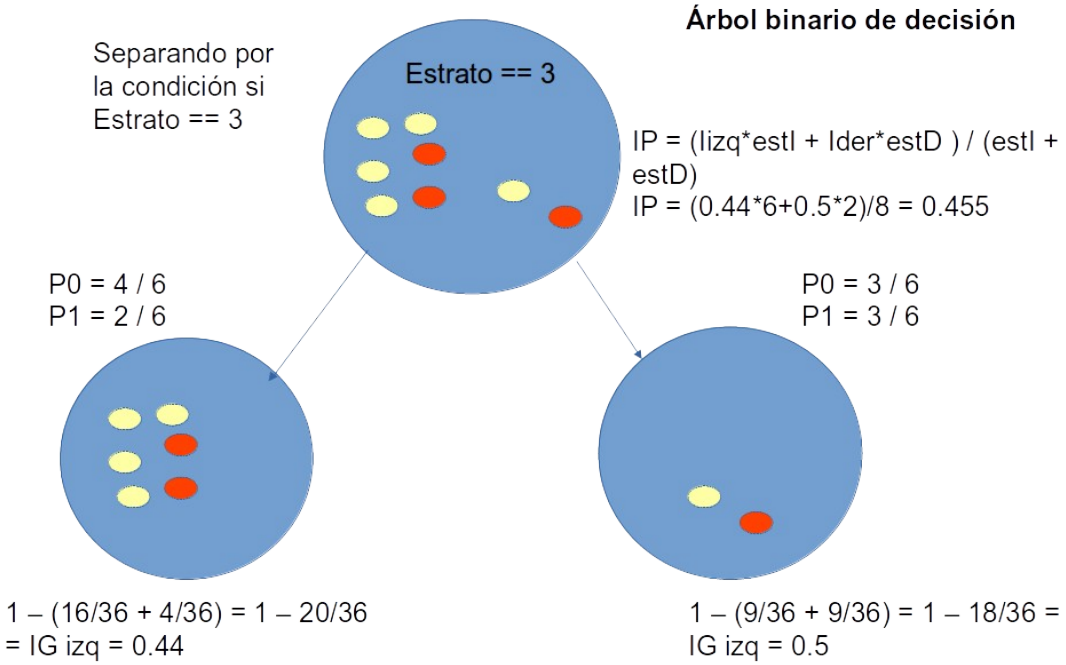
$I_G = 1 - (p_0^2 + p_1^2)$, donde p_0 es la proporción de estudiantes con la etiqueta 0 y p_1 es la proporción de estudiantes con la etiqueta 1. Las proporciones se definen como $p_0 = n_0 / (n_0 + n_1)$ y $p_1 = n_1 / (n_0 + n_1)$, donde n_i es el número de estudiantes con la etiqueta i.

La ecuación anterior describe la impureza de Gini para un conjunto de datos asociado a un nodo del árbol de decisión. Para saber cuál es la variable que divide un conjunto de datos de la mejor manera, debemos calcular la impureza de Gini del nodo de la izquierda (I_I), la impureza de Gini del nodo de la derecha (I_D) y, finalmente, la impureza ponderada (I_P) para cada una de las variables predictoras, así:

$I_P = (n_I.I_I + n_D.I_D) / (n_I + n_D)$, donde n_I es el número de estudiantes en el nodo de la izquierda y n_D es el númerode estudiantes en el nodo de la derecha.

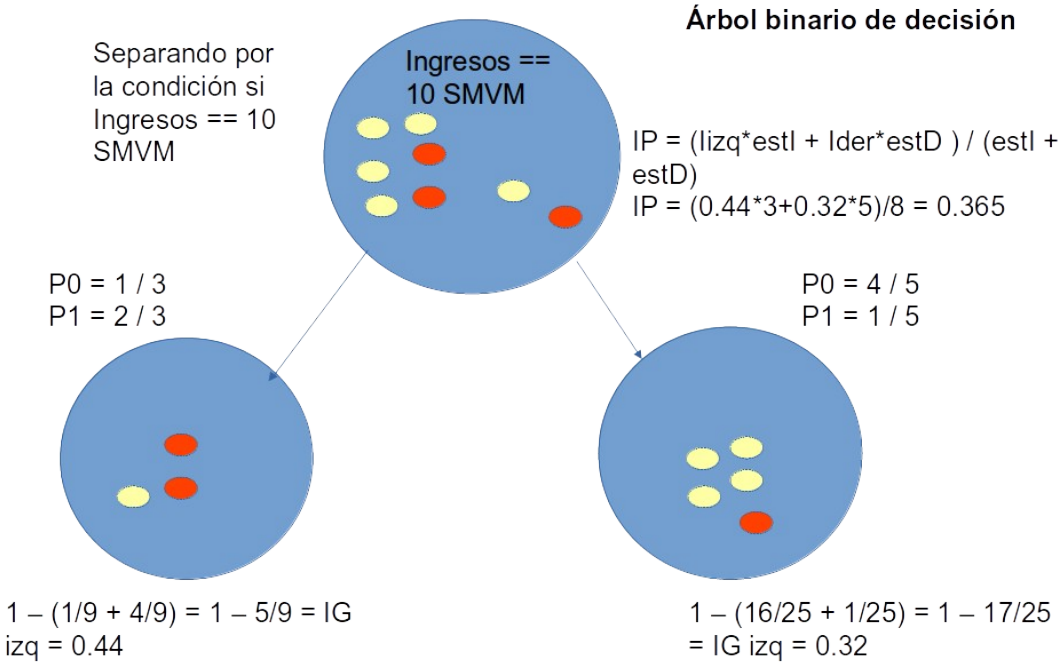
DOCENTE MAURICIO TORO BERMÚDEZ
Teléfono: (+57) (4) 261 95 00 Ext. 9473. Oficina: 19 - 627
Correo: mtorobe@eafit.edu.co

Las Gráficas 5 y 6 muestran ejemplos del cálculo de la impureza de Gini para el nodo izquierdo y derecho de un árbol, y de la impureza ponderada de Gini para el nodo raíz, así como las proporciones de quienes tienen y no tienen éxito.



Gráfica 5. Cálculo de impurezas de Gini en un árbol de decisión cuya raíz es la condición estrato == 3.

DOCENTE MAURICIO TORO BERMÚDEZ
Teléfono: (+57) (4) 261 95 00 Ext. 9473. Oficina: 19 - 627
Correo: mtorobe@eafit.edu.co



Gráfica 6. Cálculo de impurezas de Gini en un árbol de decisión cuya raíz es la condición ingresos == 10 SMVLM.

DOCENTE MAURICIO TORO BERMÚDEZ
Teléfono: (+57) (4) 261 95 00 Ext. 9473. Oficina: 19 - 627
Correo: mtorobe@eafit.edu.co

6. Algunos problemas relacionados

Para obtener información similar al problema aquí planteado, se sugiere ver los siguientes problemas relacionados:

Estructuras de datos y algoritmos relacionados

https://www.youtube.com/watch?v=LDRbO9a6XPU (video recomendado)

https://en.wikipedia.org/wiki/Decision_tree_learning

https://en.wikipedia.org/wiki/Recursive_partitioning

https://en.wikipedia.org/wiki/ID3_algorithm

https://en.wikipedia.org/wiki/C4.5_algorithm

https://en.wikipedia.org/wiki/Chi-square_automatic_interaction_detection

https://en.wikipedia.org/wiki/Alternating_decision_tree

DOCENTE MAURICIO TORO BERMÚDEZ
Teléfono: (+57) (4) 261 95 00 Ext. 9473. Oficina: 19 - 627
Correo: mtorobe@eafit.edu.co