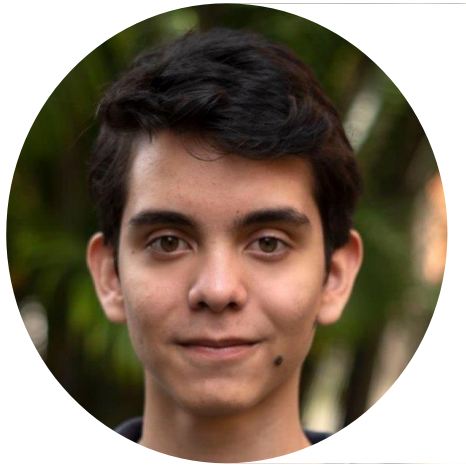
PREDICCIÓN DE RESULTADOS EN LAS PRUEBAS SABER PRO APARTIR DE LAS PRUEBAS SABER 11

UNIVERSIDAD EAFIT

Presentación del Equipo

Alejandra Vélez

Laura Zapata



Miguel Correa

Mauricio Toro

http://github.com/kattezapata/ST0245-001/proyecto/

UNIVERSIDAD EAFIT

Diseño del Algoritmo

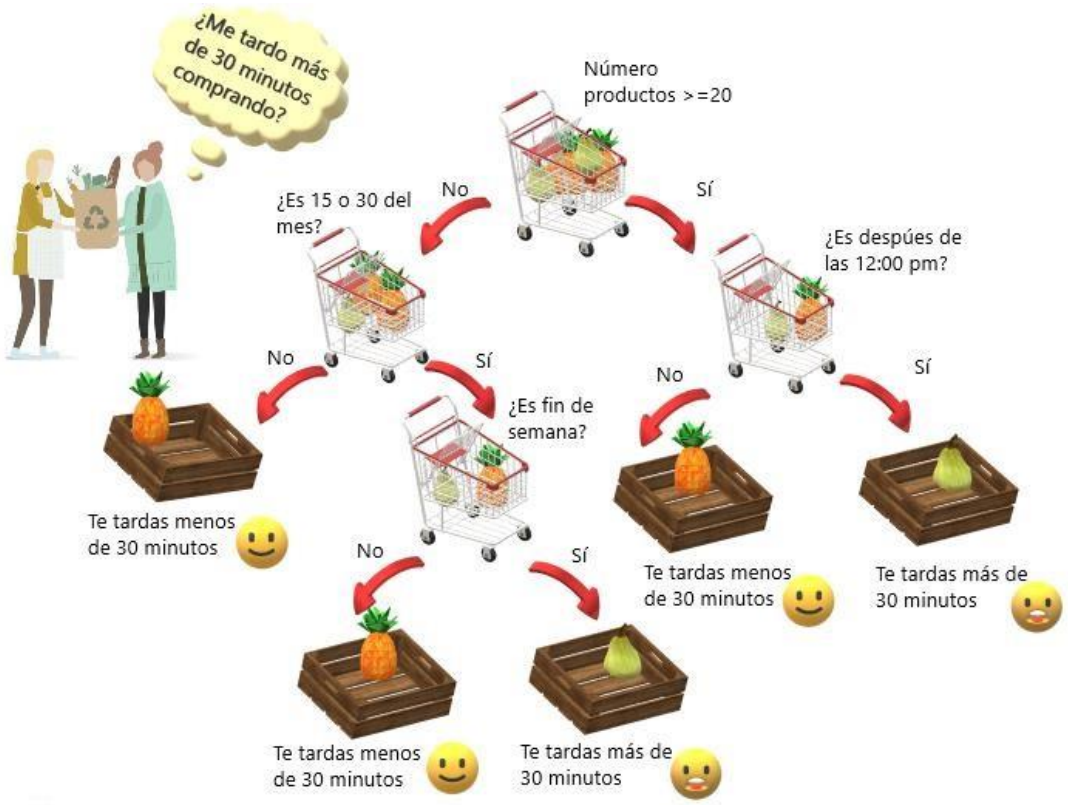


Figura 1: Algoritmo para construir un árbol binario de decisión usando CART. En este ejemplo, mostramos un modelo para predecir si una persona tardará mas de 30 minutos haciendo sus compras. Donde la condición “número de productos >= 20” es la que tiene mayor ganancia de información, es decir es la que mejor divide los datos.

División de un nodo

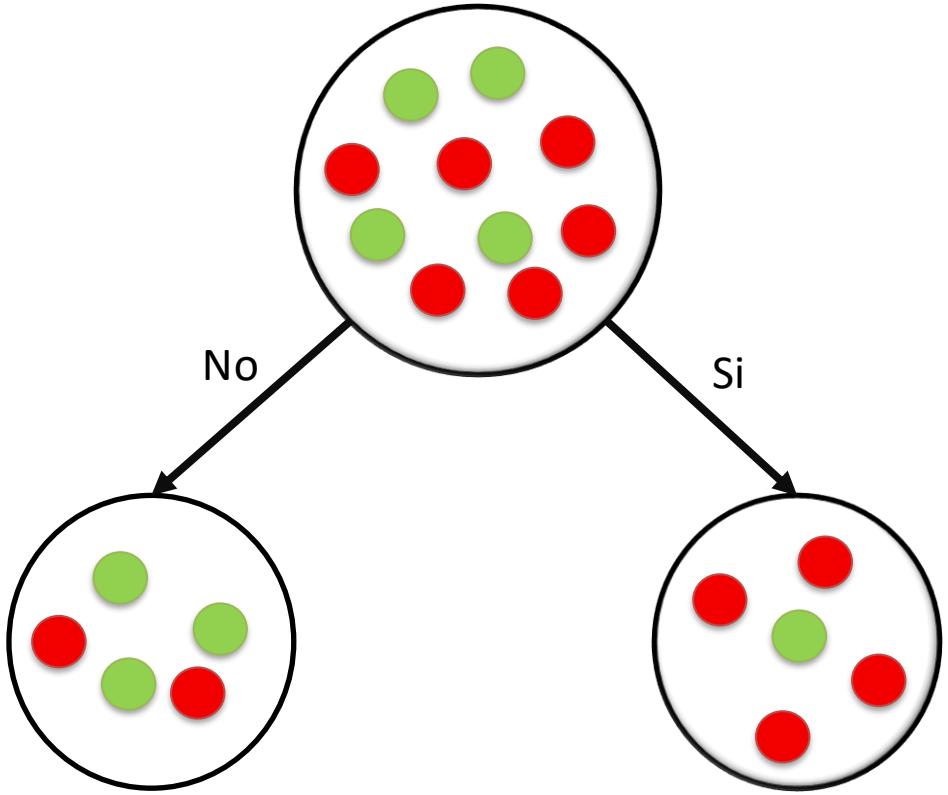


Figura 2: Esta división está basada en la condición “Puntaje lenguaje <= 65” Para este caso, la impureza Gini de la izquierda es 0.48, la impureza Gini de la derecha es 0.32 y la impureza ponderada es de 0.4.

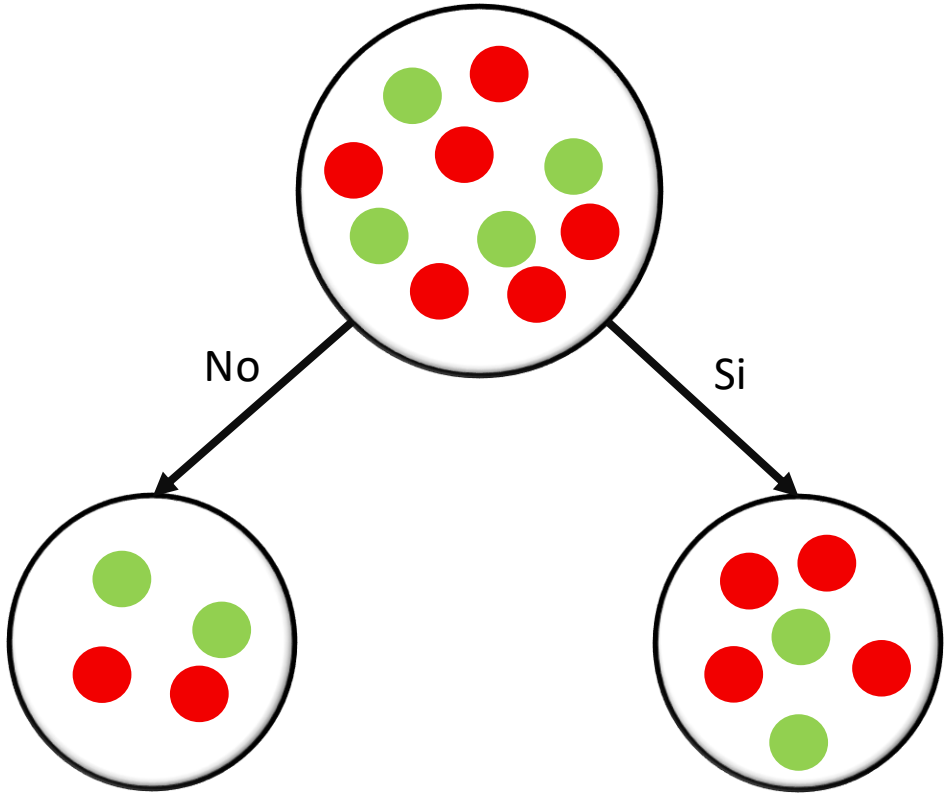


Figura 3: Esta división está basada en la condición “Puntaje de inglés <=70” Para este caso, la impureza Gini de la izquierda es 0.5, la impureza Gini de la derecha es 0.44 y la impureza ponderada es 0.46.

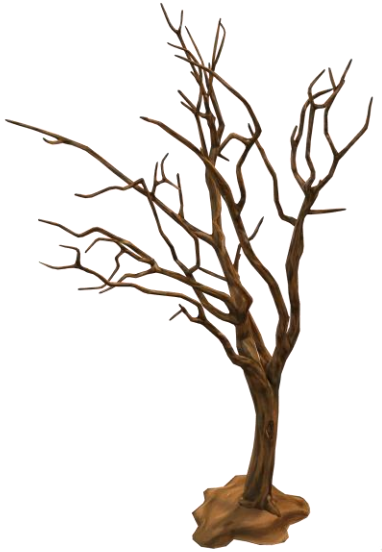
Diseño del Algoritmo

NO

NO

SÍ

NO

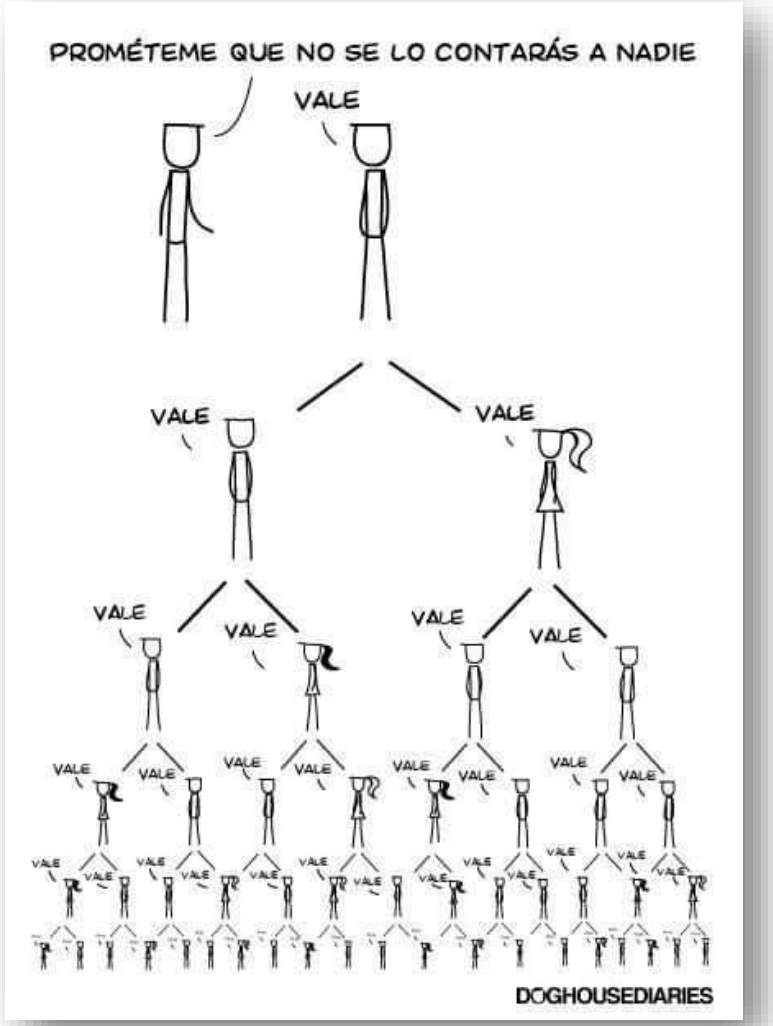


NO

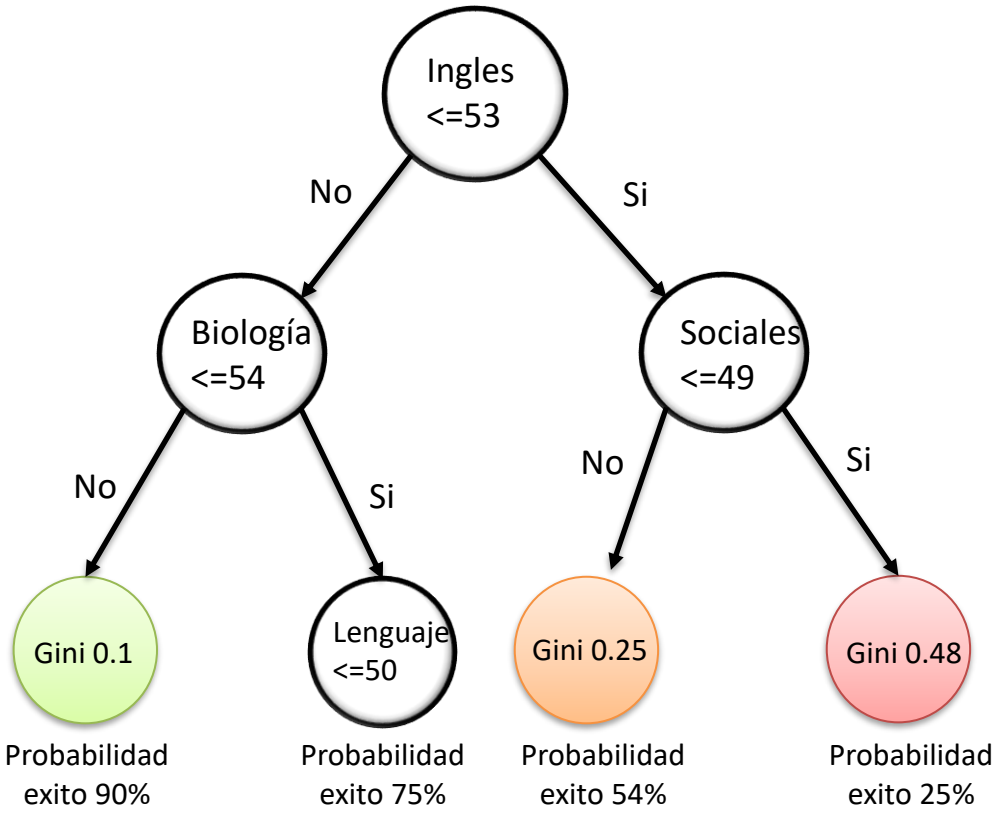
SÍ

¿Tendré éxito en el amor?

Modelo de Árbol de Decisión



Modelo de Árbol de Decisión



Características Más Relevantes

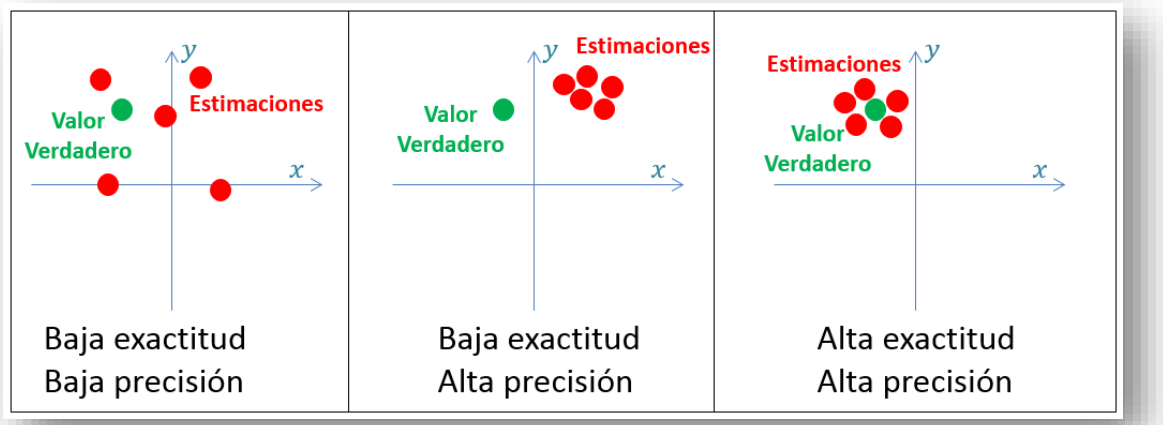
Inglés

Ciencias sociales

Biología

Un árbol de decisión para predecir el resultado del Saber Pro usando los resultados del Saber 11. Verde representa nodos con alta probabilidad de éxito; naranjado media probabilidad; y rojo baja probabilidad.

Métricas de Evaluación



Representación gráfica del conjunto de datos, aciertos y desaciertos en las predicciones

$$\textbf{\textit{Sensibilidad}} = \frac{predicción\ éxito\ acertada}{total\ de\ predicciones\ acertadas}$$

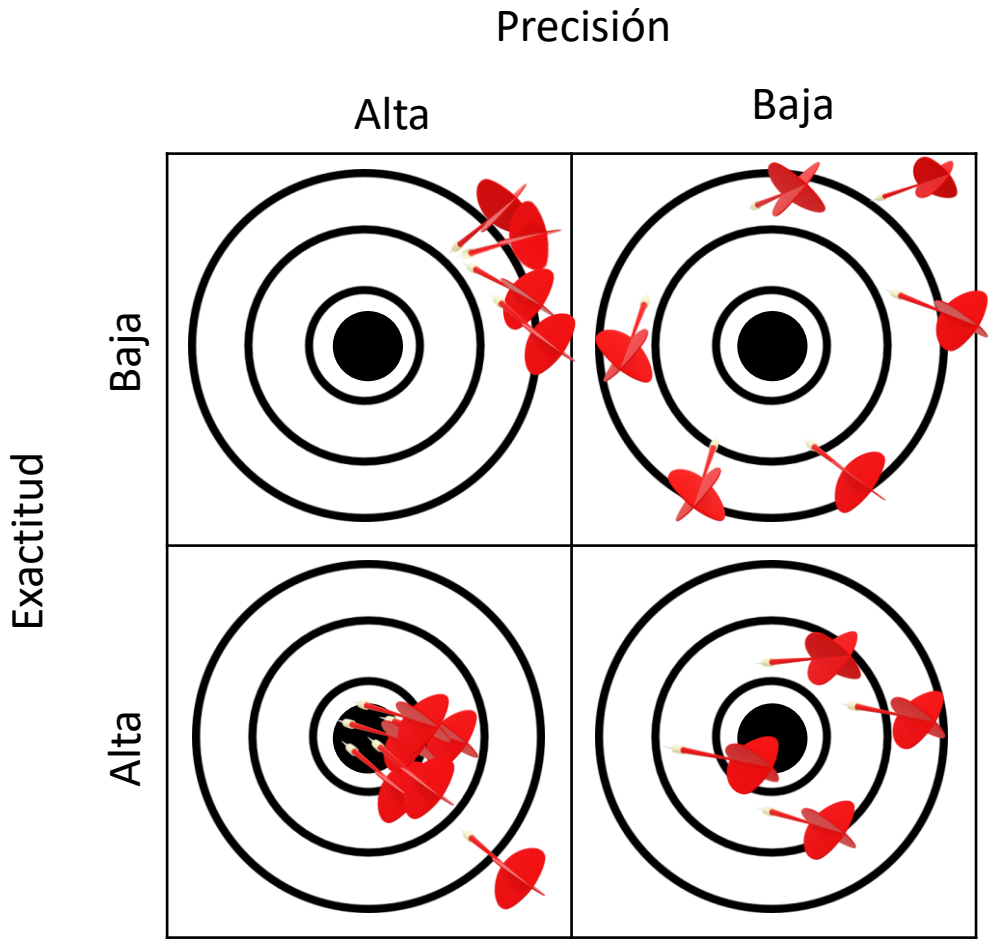
$$\textbf{\textit{Exactitud}} = \frac{total\ de\ predicciones\ acertadas}{total\ de\ los\ datos} \qquad \textbf{\textit{Precisión}} = \frac{predicción\ éxito\ acertada}{total\ de\ exitos}$$

UNIVERSIDAD EAFIT

Métricas de Evaluación

	Conjunto de entrenamiento	Conjunto de validación
Exactitud	0,78	0,77
Precisión	0,79	0,78
Sensibilidad	0,5	0,5

Métricas de evaluación obtenidas con el conjunto de datos de entrenamiento de 135,000 estudiantes y el conjunto de datos de validación de 45,000 estudiantes.

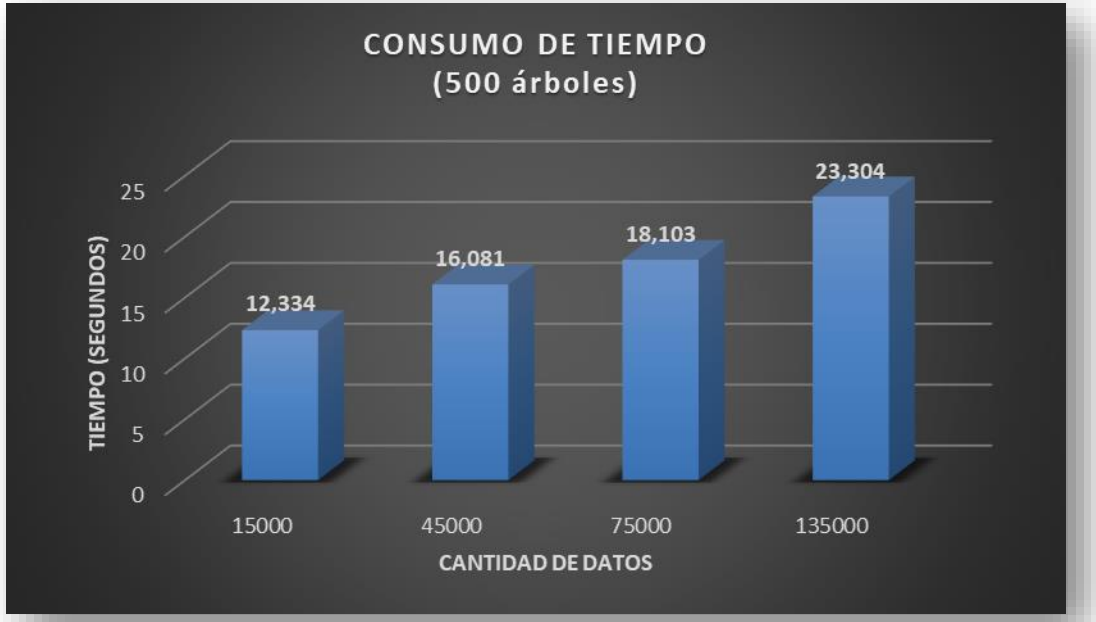


Complejidad del Algoritmo

	Complejidad en tiempo	Complejidad en memoria
Entrenamiento del modelo	O(nlogn*m*p)	O(m*n)
Probar el modelo	O(n*p)	O(m*n)

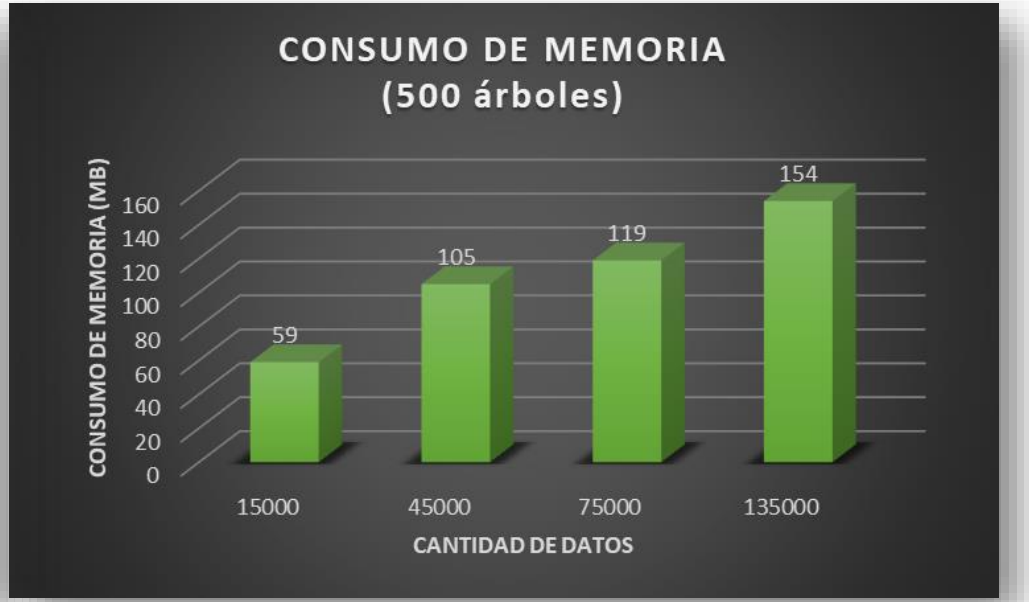
Complejidad en tiempo y memoria del algoritmo que implementa bosques aleatorios basado en árboles CART. Donde la variable **n** representa el numero de estudiantes, **m** el número de variables a tener en cuenta y **p** el número de árboles que conforman el bosque.

Consumo de tiempo y memoria



Consumo de tiempo

Consumo de tiempo y memoria



Consumo de memoria

¡GRACIAS!