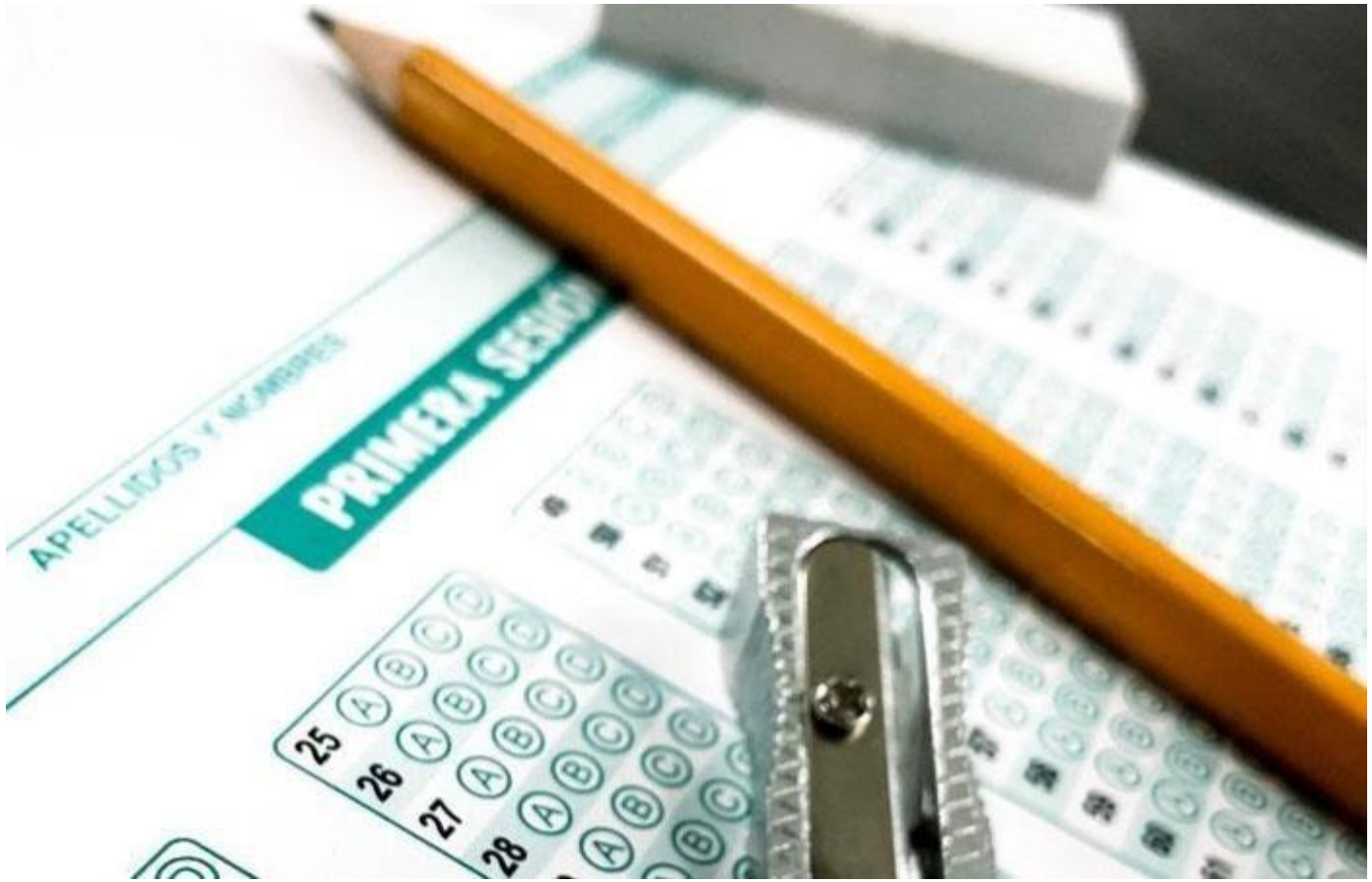
# DISEÑO DE UN ALGORITMO PARA PREDECIR EL ÉXITO EN LAS PRUEBAS SABER PRO




UNIVERSIDAD EAFIT

# Presentación del equipo

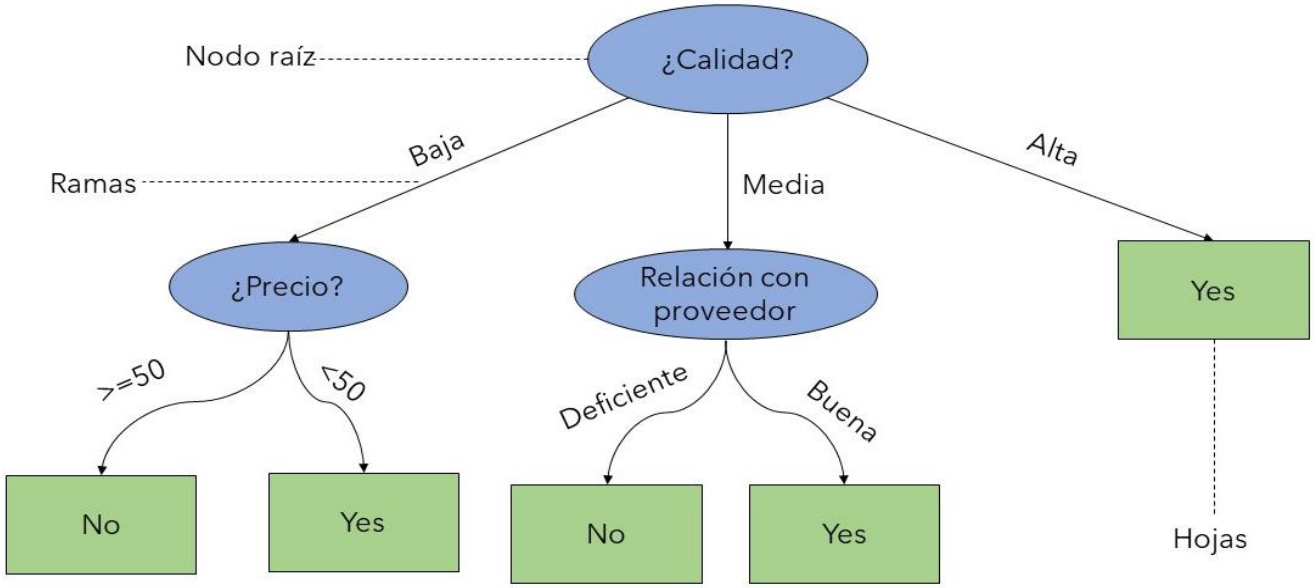
Dennis Castrillon

Sebastian Castaño

Miguel Correa

Mauricio Toro

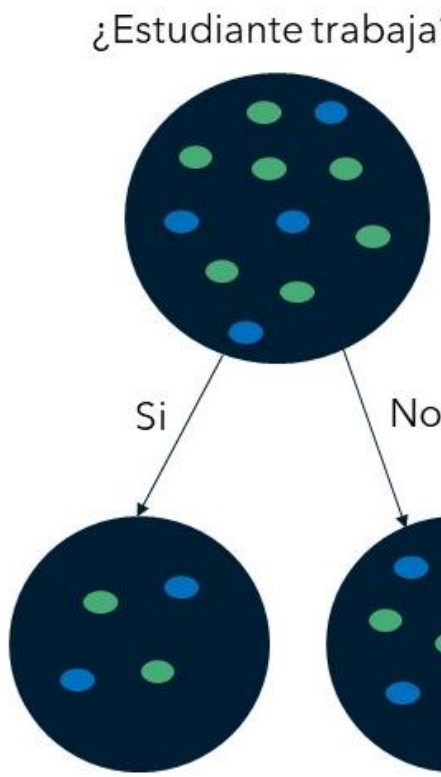
# Diseño del Algoritmo





Algoritmo para construir un árbol binario de decisión usando C4.5. En este ejemplo, mostramos un modelo para predecir si uno debe o no adquirir un material específico en una compañía, dependiendo de la calidad del material, precio y relación con el proveedor.

# División de un nodo




$$I_{GP} = \frac{(0.5 * 4) + (0.41 * 7)}{11} = 0.44$$

$$I_{G(Izq)} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

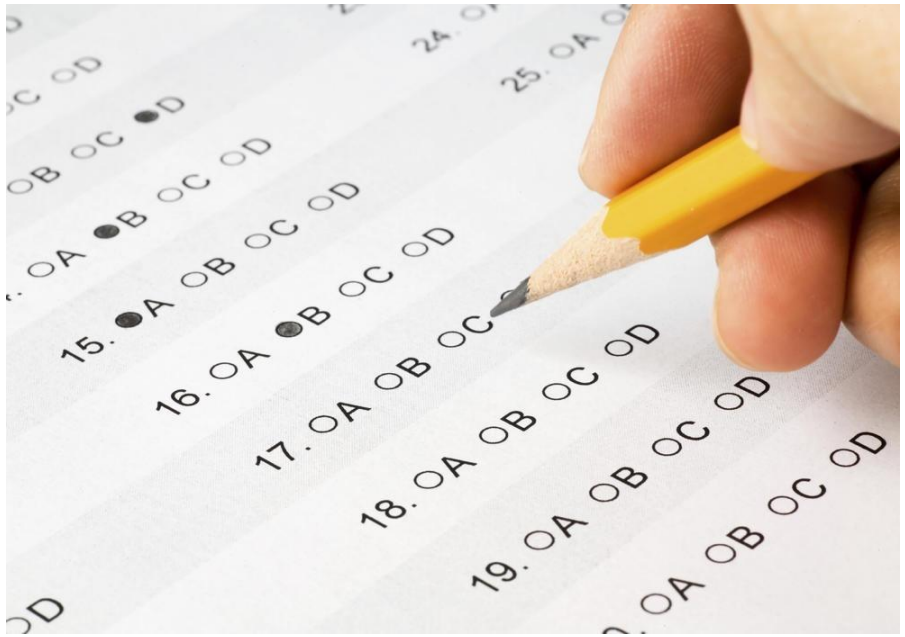
$$I_{G(Der)} = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{5}{7}\right)^2 = 0.41$$

Esta división está basada en la condición “Estudiante trabaja?”. Para ester caso la impureza de Gini del nodo de la izquierda es de 0.5 y para el nodo de la derecha es 0.41. Finalmente, la impureza de Gini ponderada es de 0.44

# Complejidad del Algoritmo

| Algoritmo (operación) | La complejidad del tiempo |
|---|---|
| **Insertar** | O(n*m) |
| **Buscar** | O(n*m) |
| **Borrar** | O(n) |

| | Conjunto 1 | | Conjunto 2 | | Conjunto 3 | |
|---|---|---|---|---|---|---|
| **Tipo** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** |
| **Consumo en MB** | 21,52 | 62,46 | 62,46 | 180,64 | 180,64 | 587,02 |



Complejidad en tiempo y memoria del algoritmo, en este caso n representa las filas del arreglo de arreglos en el que se almacenan los datos y m las columnas.

# Modelo de Árbol de Decisión

```
Is punt_matematicas >= 49.0?
--> True:
  Is punt_fisica >= 45.0?
  --> True:
    Is punt_ciencias_sociales >= 48.0?
    --> True:
      Is cole_jornada == COMPLETA?
      --> True:
        Is cole_caracter == TÉCNICO/ACADÉMICO?
        --> True:
          Is cole_depto_ubicacion == CESAR?
          --> True:
            Predict {'1': 1}
          --> False:
            Predict {'0': 5}
        --> False:
          Is fami_educacionpadre.1 == Primaria completa?
          --> True:
            Predict {'0': 2}
          --> False:
            Is cole_nombre_sede == INST TEC COLOMBO SUECO?
            --> True:
              Predict {'0': 1}
            --> False:
              Predict {'1': 9}
```

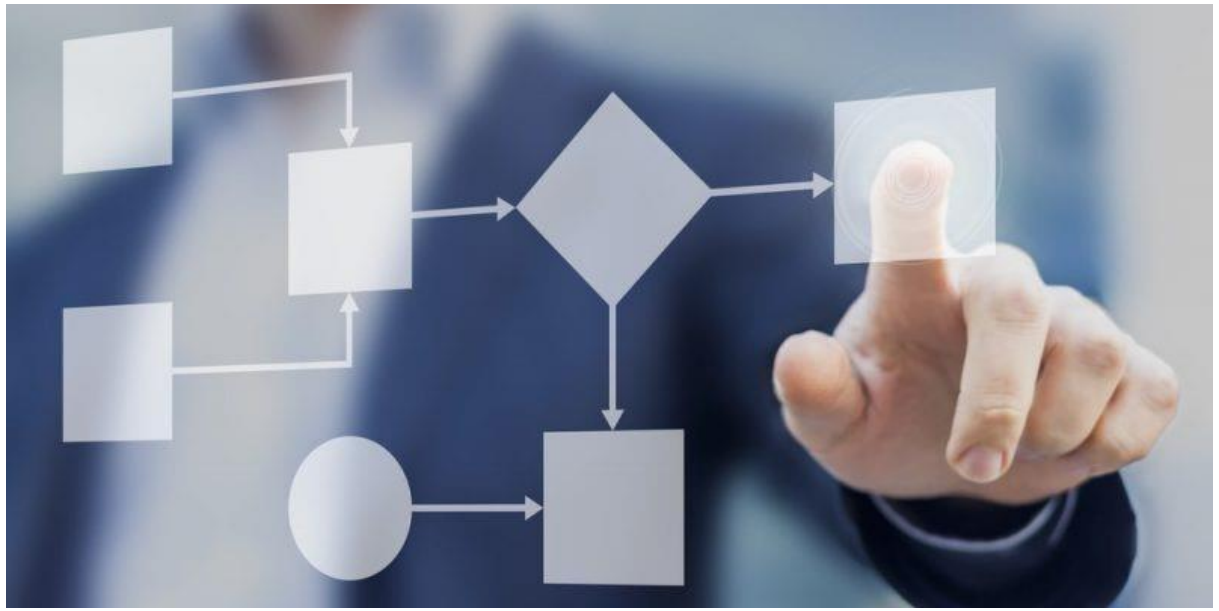
Un árbol de decisión para predecir el resultado del Saber Pro usando los resultados del Saber 11.

## Características Más Relevantes

- Matemáticas
- Inglés
- Ciencias sociales

# Funciones

- ✓ Read Data: Lee el archivo CSV, almacena los datos en un arreglo de arreglos y depura el archivo.
- ✓ Division: Separa nodos falsos y verdaderos de acuerdo a la condición.
- ✓ Gini: Calcula la impureza para un nodo específico
- ✓ Info_gain: Calcula la impureza ponderada.
- ✓ Best_división: Selecciona la condición que mejor divide el conjunto, a través de la ganancia.
- ✓ Accuracy: Calcula la exactitud
- ✓ Tree: Construye el árbol
- ✓ Print tree: Dibuja el árbol



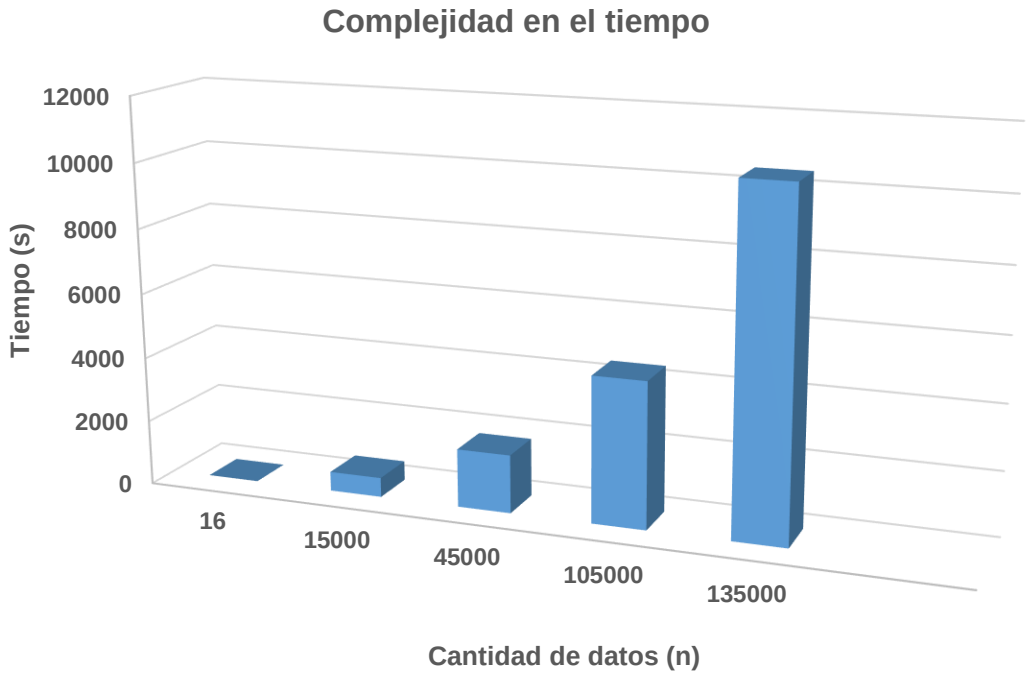
# Métricas de Evaluación

| | Conjunto 1 | | Conjunto 2 | | Conjunto 3 | |
|---|---|---|---|---|---|---|
| | Test | Train | Test | Train | Test | Train |
| **Cantidad datos** | 5000 | 15000 | 15000 | 45000 | 45000 | 135000 |

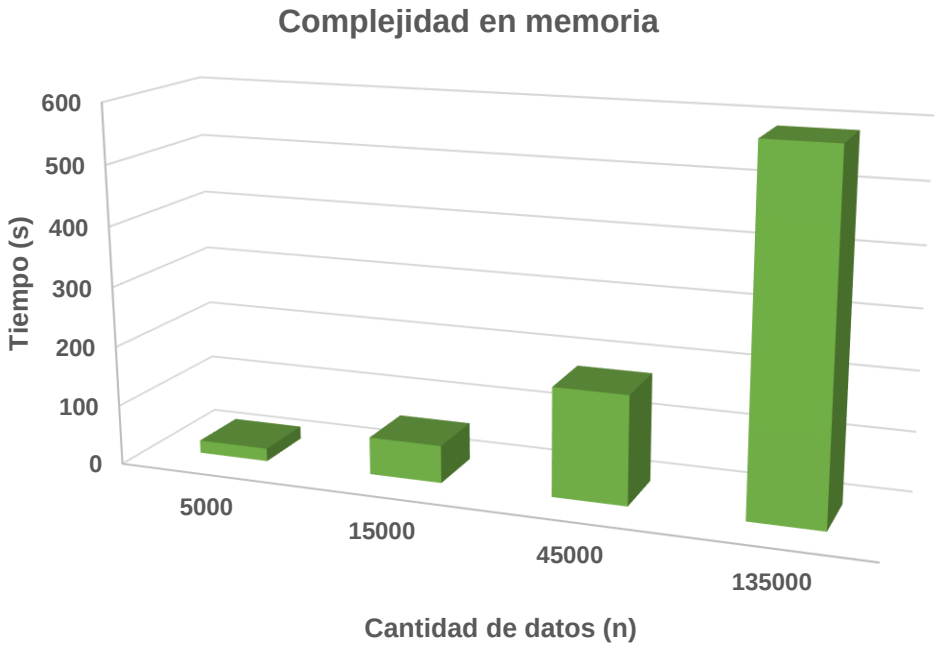
| | Conjunto de datos 1 | Conjunto de datos 2 | Conjunto de datos 3 |
|---|---|---|---|
| **Exactitud** | 0.87 | 0.82 | 0.78 |

Métricas de evaluación obtenidas para los conjuntos de datos propuestos.

# Consumo de tiempo y memoria




Consumo de tiempo




Consumo de memoria

¡GRACIAS!