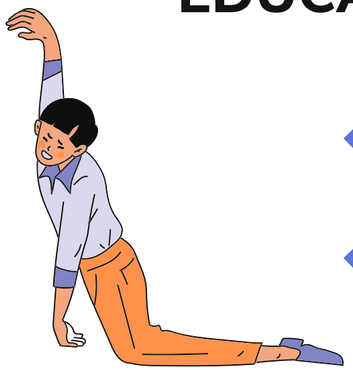
# PROYECTO FINAL

## Modelo de Machine Learning para predecir el éxito en el Saber Pro según el Saber 11

UNIVERSIDAD EAFIT 60 años

# Contexto colombiano

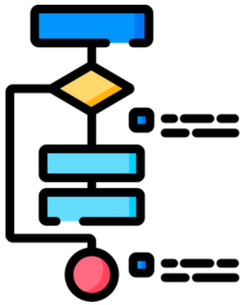


**EDUCACIÓN**

- 42% de deserción
- 2ndo país en latinoamerica con mayor tasa de deserción

(Banco Mundial, 2019)

# Objetivo



Crear un árbol de decisión que identifica qué influencia el éxito de un estudiante en Colombia en el Saber Pro según el Saber 11.

**¿CÓMO MEDIREMOS EL ÉXITO ACADÉMICO?**

La probabilidad que tiene un estudiante de obtener un puntaje total, superior al promedio de su cohorte, en la pruebas Saber Pro.

UNIVERSIDAD EAFIT® 60 años dseafit

# Datos analizados

**¿QUÉ SON?**

Resultados del Saber 11 y Saber Pro del 2014 al 2019.

**VARIABLES**

Socioeconómicas

- Nivel de educación de los padres
- Número de televisores en la casa
- Género
- Horas que pasa en internet
- ...

Resultados de cada área del Saber 11:

- Ciencias sociales
- Matemáticas
- Inglés
- ...

UNIVERSIDAD EAFIT® 60 años dseafit

*Estos datos están disponibles online en la página del ICFES

# ¿Cómo abordar este reto?

## LECTURA Y ALMACENAMIENTO DE DATOS

Conjunto de datos n*m.
Recomendación:

Usar estructura de datos con baja complejidad para acceder.
Matriz, libreria de librerias, ...

## MÉTODOS AUXILIARES

Para tomar decisiones sobre cuando partir los datos.

## ÁRBOL DE DECISIÓN

Investigar sobre los diferentes algoritmos y elegir el que te guste más.

UNIVERSIDAD EAFIT® 60 años dseafit

# Entradas y salidas

## TRAINING (ENTRENAR)


## TESTING (PROBAR)


UNIVERSIDAD EAFIT® 60 años dseafit

# Árbol de decisión

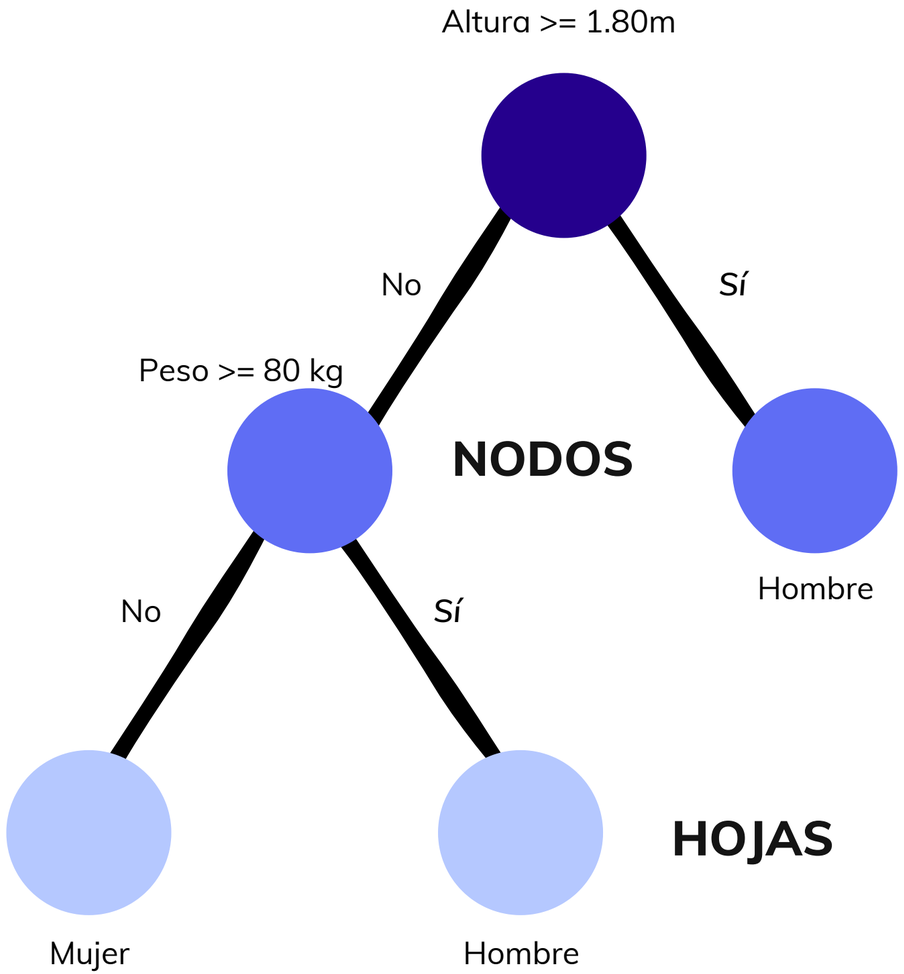
**TRAINING (ENTRENAMIENTO)**

Este paso depende del algoritmo que uses.

En general, en cada nodo se determina cuál variable nos reduce la incertidumbre.

**TESTING (CLASIFICACIÓN DE NUEVOS DATOS)**

Se pasa un set de datos nuevo y se divide según cada nodo hasta llegar a las hojas.

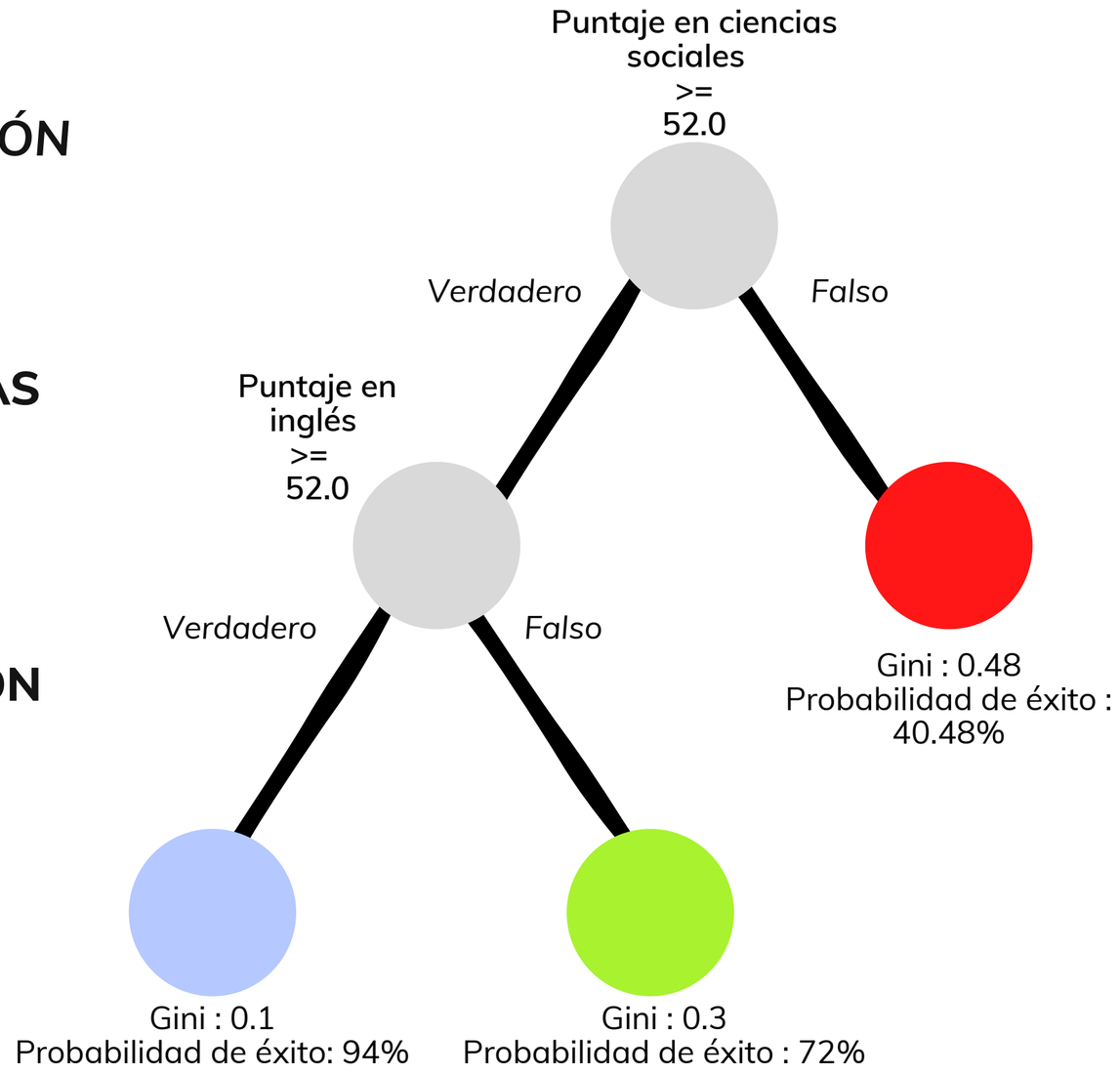



UNIVERSIDAD EAFIT® 60 años

# Algoritmo CART

***ÁRBOL DE CLASIFICACIÓN Y REGRESIÓN***

- **ÁRBOL BINARIO CON NODOS DE DECISIÓN Y HOJAS**
- **LA IMPUREZA DE GINI COMO CRITERIO DE DECISIÓN**




UNIVERSIDAD EAFIT® 60 años

***EJEMPLO DE ALGORITMO PARA CREAR EL ÁRBOL***

# Impureza de gini

Es una medida de qué tanto un elemento escogido aleatoriamente del conjunto de datos se clasificará con la etiqueta incorrecta.

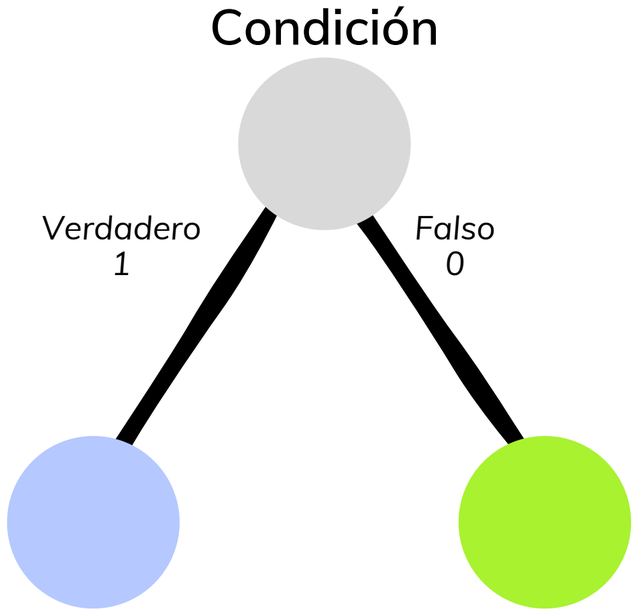
**PROPORCIÓN VERDADERO** $p1 = n1 / (n0 + n1)$

**PROPORCIÓN FALSO** $p0 = n0 / (n0 + n1)$

**IMPUREZA DERECHA (ID)** $ID = 1 - (p0^2 + p1^2)$

**IMPUREZA IZQUIERDA (II)** $II = 1 - (p0^2 + p1^2)$

**IMPUREZA PONDERADA (IP)** $IP = (n1 * II + n0 * ID) / (n1 + n0)$




UNIVERSIDAD EAFIT® 60 años

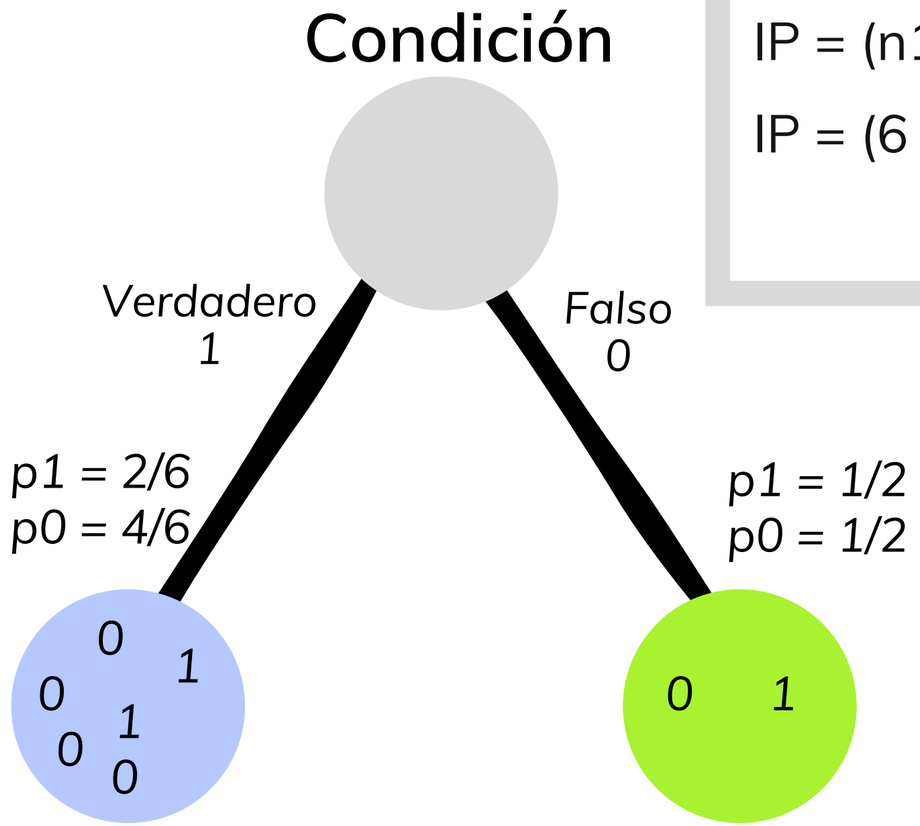
# Impureza de gini

**IMPUREZA PONDERADA (IP)**

IP = (n1 * II + n0*ID) / (n1 + n0)

IP = (6 * 0.44 + 2*0.5) / (2+6)

IP = 0.455




**IMPUREZA IZQUIERDA (II)**

II= 1 – (p0^2+ p1^2)

II= 1 – ((4/6)^2+ (2/6)^2)

II= 0.44

**IMPUREZA DERECHA (ID)**

II= 1 – (p0^2+ p1^2)

II= 1 – ((1/2)^2+ (1/2)^2)

II= 0.55

UNIVERSIDAD EAFIT® 60 años

# Posibles aplicaciones

Usando el algoritmo en situaciones de la vida actual.

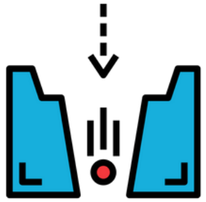
**OTORGAR BECAS**

Quitando los factores socioeconomicos del dataset.

**CRITERIA PARA ENTRAR A LA UNIVERSIDAD**

Quitando los factores socioeconomicos del dataset.

**DECIDIR SECTORES DE INVERSIÓN**



**CERRAR BRECHAS SOCIALES QUE AFECTAN EL ÉXITO**

UNIVERSIDAD EAFIT® 60 años

# Entregas

## 14 DE SEPTIEMBRE

Mandar los equipos en un pdf (equipos de 3)
Enviar las dudas generales del proyecto
Entrada de datos

## 1 DE OCTUBRE

Diseño preliminar del código (qué algoritmo usarán) en el escrito
Carga de datos (si hubo correcciones)
Adelantar métodos auxiliares

## 2 DE NOVIEMBRE

Enviar el escrito
Código (lo que lleven, ojalá terminado)

## 16 DE NOVIEMBRE

Semana de presentaciones
Última versión del código
Última versión del escrito

UNIVERSIDAD EAFIT® 60 años

# Puntos extra

**INGLÉS**

Entregar el escrito, el código y presentar en inglés.

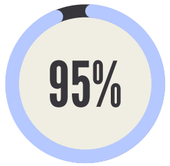
**UN PASO MÁS ALLÁ**

Crear un bosque aleatorio (investigar sobre el tema).
Esto mejora la certeza del algoritmo.

# Incluido en la entrega final

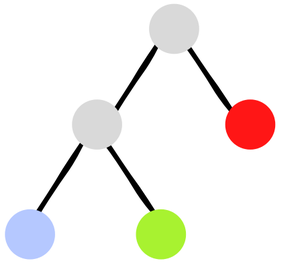
## ASISTIR A MONITORÍAS (5%)

Al menos una vez hayan asistido las tres personas del equipo.



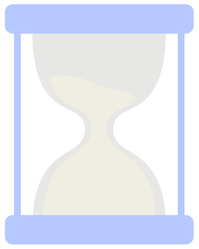

## DETERMINAR CUÁL ES LA CERTEZA DEL ALGORITMO

Incluir matriz de confusión.
Decir qué factores afectan esta certeza.



## MOSTRAR EL ÁRBOL QUE SE CREA

Imprimir una cadena y meterla en un software para visualizarlo.
Probar el árbol con diferentes profundidades de recursión.



## PERFIL DE TIEMPO Y DE COMPLEJIDAD

Según las diferentes profundidades de recursión (probar al menos 4).