INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

SITUACIÓN PROFESIONAL 2

CLASE 2

Prof. Ricardo Piña

Ver en Video:

Título: [IA] Clase 2
link: https://www.youtube.com/watch?v=CfC6_3cwVIA&t=13s (https://www.youtube.com/watch?v=CfC6_3cwVIA&t=13s)
desde: inicio
hasta 1:08:47

Out[2]:

A esta altura de la carrera Ud. todavía no sabe programar en Python, así que en este archivo hemos ocultado las celdas que contienen código para facilitar su lectura. Si Ud. quiere ver el código u ocultarlo, haga click aquí.

Out[3]:

SITUACIÓN PROFESIONAL

Dado el éxito obtenido al aplicar IA para ayudar a los médicos a efectuar el prediagnóstico de la enfermedad en base a los datos de los análisis de sangre, ha recibido una petición de un banco.

> Los bancos dan créditos y es muy importante para ellos que los clientes devuelvan el dinero que les prestaron, para lo cual cuentan con personal especializado que son los Analistas de Crédito, cuya tarea principal es evaluar si un solicitante devolverá el crédito o no.
> En este caso el banco posee cierta información histórica sobre otros clientes que con anterioridad han recibido créditos del banco, algunos devolvieron los créditos que les dieron y otros no.
> El banco le solicita a Ud que determine si existe algún patrón en los datos históricos que les permita anticipar si futuros clientes serán buenos pagadores o no.

Nota: Los datos provistos para este problema son ficticios.

En la Situación Profesional anterior habíamos visto que podíamos aplicar un sencillo **método gráfico** que nos permitía extraer **reglas** de un conjunto de datos, las cuales luego podíamos plasmar en el método del **Árbol de Decisiones**.

Dos condiciones nos permitieron utilizar el **método gráfico**: primero, que sólo teníamos un par de variables, y segundo que estas variables eran numéricas, lo cual nos permitía asimilarlas a un par de ejes coordenados cartesianos de la misma manera que aprendió a usarlos en el colegio.

En el caso actual, en cambio, no se cumplirá ninguna de estas dos condiciones, así que deberemos aguzar nuestro ingenio para arribar a un árbol de decisiones que ayude al banco.

Primero veamos los datos:

Out[5]:

	Tiene_Deuda	Genero	Trabaja	Propietario	Dar_Credito
0	Si	F	No	Si	No
1	Si	M	No	Si	No
2	No	F	Si	Si	Si
3	No	F	Si	No	Si
4	No	F	No	No	No
5	No	M	Si	No	No
6	No	M	No	Si	No
7	No	M	No	No	No
8	Si	F	No	No	No
9	Si	M	Si	No	No
10	Si	M	No	No	No
11	No	F	No	Si	No
12	No	M	Si	Si	Si

En este caso **Dar_Credito** es lo que deseamos pronosticar, basándonos en los datos que poseemos de las otras variables.
Los datos corresponden a clientes a quienes se ha otorgado préstamos en el pasado, algunos pagaron su crédito y otros no, a los primeros se les cargó el valor Si en la columna Dar_Credito y a los que no pagaron se les cargó el valor No en la columna Dar_Credito.

Como se puede observar tenemos 4 variables o características (features) con las que deberemos pronosticar el valor de Dar_Credito.

- **Tiene_Deuda**: indica si el solicitante tenía alguna deuda al momento de solicitar el crédito.
- **Genero**: femenino o masculino
- **Trabaja**: indica si la persona trabajaba o no al momento de solicitar el crédito.
- **Propietario**: indica si el solicitante poseía una propiedad (casa/ depto, etc) o no al momento de solictar el crédito

Nota: Observe que hemos procurado que los nombres de las variables no contengan espacios, símbolos especiales, acentos, ni ñ, ya que pueden traer problemas en programas hechos en inglés y también hemos evitado los espacios ya que puede traer problemas en algunos programas.

Como comentábamos anteriormente para intentar **pronosticar** Dar_Credito, **no podremos apelar** a un gráfico cartesiano para ubicar los puntos ya que:

- por un lado tenemos 4 variables y nuestra imaginación sólo nos permitiría graficar hasta 3 y,
- por otro lado los valores que asumen estas variables **no son numéricos**, son valores **categóricos**, así por ejemplo los valores de la variable Genero corresponden a dos categorías: F o M, en el caso del resto de las variables las categorías son Si / No.

Ahora sí, tómese su tiempo porque llegó el momento de pensar!: vea los datos, e intente encontrar Ud las reglas por las cuales es posible distinguir cuándo el valor de Dar_Credito es Si o es No.

Todo está permitido: si es necesario tome papel y lápiz y haga los esquemas que crea necesarios, si se siente cómodo usando planillas de cálculo como Excel, puede cargar los datos y pruebe cómo puede hacer hacer para distinguir estos casos.

No siga leyendo hasta no haber probado de diversas maneras cómo resolver el problema, no importa si lo consigue o no, pero es necesario que utilice su razonamiento en ese sentido durante un buen tiempo (una media hora puede ser un tiempo razonable).

Apele a su ingenio!

Deberemos apelar entonces a armar el esquema de **Árbol de Decisión**, pero internamente sabemos que estaremos dividiendo el espacio de 4 dimensiones en zonas de tal manera que en cada zona sólo queden casos donde Dar_Credito sea Si o No.

Pensemos en cómo armar nuestro árbol.

Primero analicemos qué tenemos.

Tenemos 13 observaciones o casos o registros, como le guste llamarlos, de los cuales:

- 3 son Si
- 10 son No

Ahora bien, ¿qué es lo que estamos buscando?

Lo ideal sería encontrar una varible que separara o distinguiera con total "precisión" los casos de No y Si de Dar_Credito, algo como ésto:

ideal	Tiene_Deuda	Genero	Trabaja	Propietario	Dar_Crédito
A	Si	F	No	Si	No
A	Si	F	No	No	No
A	Si	M	Si	No	No
A	Si	M	No	Si	No
A	Si	M	No	No	No
A	No	F	No	Si	No
A	No	F	No	No	No
A	No	M	Si	No	No
A	No	M	No	Si	No
A	No	M	No	No	No
B	No	F	Si	Si	Si
B	No	F	Si	No	Si
B	No	M	Si	Si	Si

Pero, lamentablemente esa variable ideal no existe entre nuestros datos. Las variables que poseemos seguramente no separarán tan bien como la ideal, pero tenemos una idea de lo que estamos buscando.

Veamos qué pasaría si decidimos comenzar a analizar con la variable **Tiene_Deuda**. Ordenaré las observaciones por la variable Tiene_Deuda para facilitar la visualización

Out[7]:

	Tiene_Deuda	Genero	Trabaja	Propietario	Dar_Credito
2	No	F	Si	Si	Si
3	No	F	Si	No	Si
4	No	F	No	No	No
5	No	M	Si	No	No
6	No	M	No	Si	No
7	No	M	No	No	No
11	No	F	No	Si	No
12	No	M	Si	Si	Si
0	Si	F	No	Si	No
1	Si	M	No	Si	No
8	Si	F	No	No	No
9	Si	M	Si	No	No
10	Si	M	No	No	No

Seguramente sólo prestó atención a las columnas Tiene_Deuda y Dar_Credito así que podemos obviar las otras:

Out[8]:

	Tiene_Deuda	Dar_Credito
2	No	Si
3	No	Si
4	No	No
5	No	No
6	No	No
7	No	No
11	No	No
12	No	Si
0	Si	No
1	Si	No
8	Si	No
9	Si	No
10	Si	No

Para facilitar la visualización las mostraré de la siguiente manera con un poco de color:

Tiene_Deuda	Dar_Credito
No	Si
No	Si
No	No
No	No
No	No
No	No
No	No
No	Si
Si	No
Si	No
Si	No
Si	No
Si	No

Es muy interesante lo que ocurre para los casos en que Tiene_Deuda toma el valor Si: en todos esos casos el valor de Dar_Credito es No.

Veamos qué pasa si lo hacemos por la variable Genero

Genero	Dar_Credito
F	No
F	Si
F	Si
F	No
F	No
F	No
M	No
M	No
M	No
M	No
M	No
M	No
M	Si

Si el último caso de Genero = M hubiera dado No en Dar_Credito hubiera sido también un caso muy interesante, no es cierto?

Ahora veamos cómo quedaría si lo hacemos con Trabaja

Trabaja	Dar_Credito
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
Si	Si
Si	Si
Si	No
Si	No
Si	Si

Es muy interesante lo que ocurre para los las observaciones donde Trabaja vale No: en todas ellas Dar_Credito es No.

Por último con Propietario

Propietario	Dar_Credito
No	Si
No	No
No	No
No	No
No	No
No	No
No	No
Si	No
Si	No
Si	Si
Si	No
Si	No
Si	Si

Veamos todas las posibilidades juntas para poder comparar.
¿Cuál de las variables le parece mejor para comenzar a armar nuestro árbol de decisiones?

Tiene_Deuda	Dar_Credito
No	Si
No	Si
No	No
No	No
No	No
No	No
No	No
No	Si
Si	No
Si	No
Si	No
Si	No
Si	No

Genero	Dar_Credito
F	No
F	Si
F	Si
F	No
F	No
F	No
M	No
M	No
M	No
M	No
M	No
M	No
M	Si

Trabaja	Dar_Credito
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
Si	Si
Si	Si
Si	No
Si	No
Si	Si

Propietario	Dar_Credito
No	Si
No	No
No	No
No	No
No	No
No	No
No	No
Si	No
Si	No
Si	Si
Si	No
Si	No
Si	Si

Después de pensarlo un rato quizá se haya quedado con dos candidatos, Tiene_Deuda y Trabaja.

Analicemos Tiene_Deuda:

- Observemos que en todos los casos (5 casos) en que Tiene_Deuda = Si, el valor de Dar_Credito = No. Lo cual significa que podríamos crear una regla que dijera:

  > **si el cliente tiene deuda, entonces, no dar el crédito.** Y clasificamos bien 5 casos de los 13 que había al principio.

- Después de aplicar la regla anterior nos quedarían 8 casos con **incertidumbre**, cuando el cliente no tiene deuda, ya que a veces conviene dar el crédito y otras veces no.

Ahora analicemos Trabaja:

- En todos los casos (8 casos) en que Trabaja = No, el valor de Dar_Credito = No. Podríamos establecer una regla que dijera:

  > **si el cliente no trabaja, entonces, no dar el crédito**; y habríamos clasificado bien 8 casos de los 13 que había al principio.

- Después de aplicar la regla anterior nos quedarían sólo 5 casos con **incertidumbre**, cuando el cliente trabaja, ya que en algunos casos conviene dar el crédito y en otros no.

  > Así que en definitiva nos convendría comenzar el Árbol de Decisión separando los casos por la variable Trabaja, de esa manera ya habríamos resuelto el problema para 8 casos y sólo tendríamos **incertidumbre** con respecto a los 5 casos restantes, que luego analizaríamos usando otra de las variables disponibles.

Entropía e Incertidumbre

Pensemos en el caso anterior, iniciamos el problema con una situación donde teníamos que generar algún criterior para clasificar la variable Dar_Credito correctamente en las 13 observaciones, de las cuales 10 eran No y 3 eran Si.
En esta situación inicial diríamos que tenemos un **nivel medianamente alto de incertidumbre** (decimos sólo *medianamente alto* porque sabemos que el máximo valor de incertidumbre se daría cuando los Si y los No se encontraran en la misma cantidad, como habíamos mencionado con anterioridad).

Dar_Credito
No
No
Si
Si
No
No
No
No
No
No
No
No
Si

Luego, descubrimos una regla, correspondiente al uso de la variable Trabaja:

Trabaja	Dar_Credito
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
Si	Si
Si	Si
Si	No
Si	No
Si	Si

Esto mejoró nuestra situación porque al aplicar la regla aprendida **si el cliente no trabaja, entonces, no dar el crédito**, estamos reduciendo nuestro problema a la siguiente situación:

Trabaja	Dar_Credito
Si	Si
Si	Si
Si	No
Si	No
Si	Si

Ahora sólo tenemos **incertidumbre** sobre los 5 casos que nos quedaron.

Lo que hemos estado haciendo es luchar contra la **incertidumbre**, tratando de que ésta **disminuya** cuando elegimos separar con alguna de las variables.

Para armar el Árbol de Decisiones procedermos como lo hicimos con este ejemplo:

> **Probaremos con todas las variables posibles para ver cómo clasifican a las observaciones y elegiremos comenzar con la variable que más disminuya la incertidumbre inicial.**

Sin embargo debemos reconocer que el concepto de **incertidumbre no ha quedado definido** y la idea de cuándo hay más o menos incertidumbre sería mejor resolverla con un **valor numérico**.

Fíjese en estos dos casos donde tenemos dos variables x1 y x2, cuyos valores no nos interesan, por eso los dejamos en blanco; y una variable que deseamos pronosticar.
¿En cuál de los casos le parece que habrá **más incertidumbre** en determinar el valor de esa variable?

Caso 1

x1	x2	variable a pronosticar
		B
		A
		B
		B
		B
		B
		B
		B
		B
		B

Caso 2

x1	x2	variable a pronosticar
		A
		B
		A
		A
		B
		A
		B
		B
		A
		B

Si pensó que el Caso 2 es más incierto, coincidimos en el análisis.

- En el Caso 1 tenemos 10 observaciones una sola es A y las 9 restantes son B; si hubiera una nueva observación casi seguro que sería una B.
- En el Caso 2, también tenemos 10 observaciones, pero 5 son A y las otras 5 son B; ¡si hubiera una nueva observación la verdad es que no sabría por cuál de las dos apostar! Hay más incertidumbre en el Caso 2.

Afortunadamente **la incertidumbre se puede medir con un valor numérico**, mediante la fórmula de la **Entropía de Shannon** y como esta fórmula nos dará un valor numérico; será fácil de comparar.
En esta materia aceptaremos como valedera esta fórmula, más adelante la profundizaremos.

Procedimiento para crear el Árbol de Decisión:

1. Calcularemos la Entropía (incertidumbre) de la variable que deseamos pronosticar con la fórmula que veremos luego.
2. Probaremos **una por una** las variables como primer nodo, veremos cómo clasifica, particiona o divide las observaciones y **calcularemos la entropía para cada una de estas clasificaciones**.
3. **Calcularemos la entropía luego de la partición.**
4. La variable que nos dé una **mayor disminución de la entropía** (incertidumbre) será la elegida como primer nodo.
5. Crearemos las ramas de este primer nodo. Si tuvimos mucha suerte las dos ramas terminarían en hojas, pero seguramente por lo menos una de las ramas irá hacia un nodo donde todavía hay incertidumbre (casos con ambos valores).
6. Repetiremos los pasos anteriores para cada uno de los nodos que se creen, excepto que sea una hoja.

Sólo nos falta la fórmula para calcular la Entropía

Fórmula de la Entropía de Shannon

Supongamos que queremos calcular la **entropía** o incertidumbre en el Caso 1 anterior; observamos que:

- la variable a pronosticar puede tomar 2 valores, A o B.
- en total son 10 observaciones
- el valor A aparece 1 vez
- el valor B aparece 9 veces

Una de las cosas que deberemos calcular es la **proporción** en que aparece cada uno de los valores (A y B en este caso):

- $p_A = 1/10$
- $p_B = 9/10$

Ahora estamos en condiciones de presentar la Fórmula de la Entropía de Shannon:

$$\mathbf{S = - [p_A \log_2(p_A) + p_B \log_2(p_B)]}$$

Nota: *la expresión log_2 significa logaritmo en base 2. Generalmente en las calculadoras científicas no figura el logaritmo en base 2, en su lugar suelen estar log() cuya base es 10 o Ln() cuya base es el número e. Afortunadamente es posible calcular el logaritmo en cualquier base conociendo cualquiera de los anteriores, usando la fórmula de cambio de base de los logaritmos, de la siguiente manera:*

$\log_a(x) = \log(x) / \log(a)$
o también:
$\log_a(x) = Ln(x) / Ln(a)$

Aplicándola a nuestro caso nos daría:

$S1 = - [1/10 \log_2(1/10) + 9/10 \log_2(9/10)]$

```
La Entropía del Caso 1 es:  0.47
```

Ahora calcule la entropía del Caso 2, a ver si nuestra intuición se verifica:

$S2 = - [5/10 \log_2(5/10) + 5/10 \log_2(5/10)]$

```
La Entropía del Caso 2 es:  1.0
```

Tal como esperábamos había más incertidumbre (entropía) en el Caso 2 que en el Caso 1.

Nota: si bien el log(0) no existe, en este contexto se considerará que el producto de **0 log(0) = 0**

Ahora que ya sabemos calcular la entropía como una medida de la incertidumbre podemos aplicarlo para crear el Árbol de Decisión de una forma más robusta que antes.

ARBOL DE DECISIÓN POR GANANCIA DE INFORMACIÓN

Los conceptos de entropía o incertidumbre son opuestos al concepto de información, es decir que la disminución de la incertidumbre implica también una ganancia en la información, de allí el nombre del método. Este método es usado en varios algoritmos de creación de árboles de decisión, por ejemplo en el muy conocido **ID3**.

Apliquémoslo al problema del crédito.

Recordemos los datos del problema:

Out[11]:

	Tiene_Deuda	Genero	Trabaja	Propietario	Dar_Credito
0	Si	F	No	Si	No
1	Si	M	No	Si	No
2	No	F	Si	Si	Si
3	No	F	Si	No	Si
4	No	F	No	No	No
5	No	M	Si	No	No
6	No	M	No	Si	No
7	No	M	No	No	No
8	Si	F	No	No	No
9	Si	M	Si	No	No
10	Si	M	No	No	No
11	No	F	No	Si	No
12	No	M	Si	Si	Si

Cálculo de la Entropía Inicial

Vamos a calcular la **entropía inicial** para la variable Dar_Credito que es la que queremos pronosticar:

Tenemos:

- En total son 13 observaciones
- 10 son No
- 3 son Si

Calculemos las proporciones de No y de Si:

p_{No} = 10/13
p_{Si} = 3/13

Calculamos la Entropía Inicial con la fórmula de Shannon:

$S_{inicial}$ = - [10/13 $\log_2$(10/13) + 3/13 $\log_2$(3/13)]

```
La Entropía Inicial es:  0.779
```

Ahora tenemos que calcular la entropía que nos quedaría al utilizar como primer nodo para clasificar a cada una de las variables o características de las que disponemos que en esta caso son:

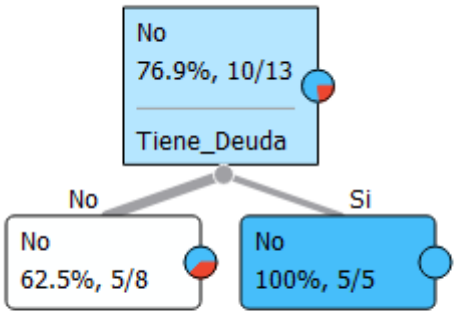
- Tiene_Deuda
- Genero
- Trabaja
- Propietario

Cálculo de las Entropías de los nodos que se forman al particionar con la variable Tiene_Deuda:

Variable Tiene_Deuda

En forma de Tabla **Gráfico de Árbol**

Tiene_Deuda	Dar_Credito
No	Si
No	Si
No	No
No	No
No	No
No	No
No	No
No	Si
Si	No
Si	No
Si	No
Si	No
Si	No



> Calculamos la Entropía de cada rama y luego las sumamos **multipicadas por la proporción de casos con respecto al nodo superior**:

Rama izquierda (**Tiene_Deuda= No**):

- cantidad de casos:8
- **proporción de casos con respcto al nodo superior: 8/13**
- cantidad de No: 5
- cantidad de Si: 3

Su entropía será:

$S1_{tiene\_deuda\_izq}$=-[5/8 $\log_2$(5/8) + 3/8 $\log_2$(3/8)]

Out[13]:

0.954

Rama derecha (**Tiene_Deuda=Si**):

- cantidad de casos: 5
- **proporción de casos con respecto al nodo superior: 5/13**
- cantidad de No: 5
- Cantidad de Si: 0

Su entropía será:

$S1_{tiene\_deuda\_der}$ = -[5/5 $\log_2(5/5)$ + 0 $\log_2(0)$]

Out[14]:

```
-0.0
```

¿Por qué esta rama dio entropía = 0?

Porque en esta rama todos los valores son No, por lo tanto no hay incertidumbre!

Finalmente debemos sumar las entropías de cada rama **multipicadas por la proporción de casos con respecto al nodo superior**:

$S1_{tiene\_deuda}$ = -8/13 x $S1_{tiene\_deuda\_izq}$ - 5/13 x $S1_{tiene\_deuda\_der}$

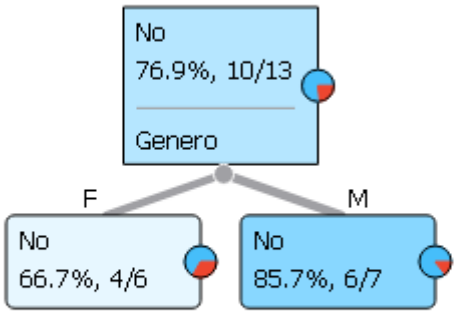
```
La Entropía al usar Tiene_Deuda para particionar el conjunto de datos es:
S1tiene_deuda =  0.587
```

Repitamos el proceso para la variable Genero

Variable Genero

En forma de Tabla **Gráfico de Árbol**

Genero	Dar_Credito
F	No
F	Si
F	Si
F	No
F	No
F	No
M	No
M	No
M	No
M	No
M	No
M	No
M	Si



Calculamos la Entropía de cada rama y luego las sumamos **multipicadas por la proporción de casos con respecto al nodo superior**:

Rama izquierda (**Genero= F**):

- cantidad de casos:6
- **proporción de casos con respcto al nodo superior: 6/13**
- cantidad de No: 4
- cantidad de Si: 2

Su entropía será:

$$\mathbf{S1_{Genero\_izq} = -[\ 4/6\ \log_2(4/6) + 2/6\ \log_2(2/6)\ ]}$$

Out[16]:

```
0.918
```

Rama derecha (**Genero=M**):

- cantidad de casos: 7
- **proporción de casos con respcto al nodo superior: 7/13**
- cantidad de No: 6
- Cantidad de Si: 1

Su entropía será:

$$\mathbf{S1_{Genero\_der} = -[\ 6/7\ \log_2(6/7) + 1/7\ \log_2(1/7)\ ]}$$

Out[17]:

```
0.592
```

Finalmente debemos sumar las entropías de cada rama **multipicadas por la proporción de casos con respecto al nodo superior**:

$$\mathbf{S1_{Genero} = -\ 6/13\ S1_{Genero\_izq} - 7/13\ S1_{Genero\_der}}$$

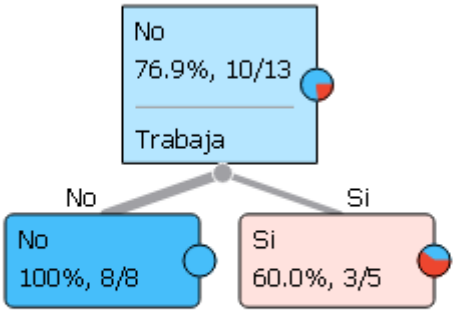
```
La Entropía al usar Genero para particionar el conjunto de datos es: S1Gen
ero  0.742
```

Repitamos el proceso para la variable Trabaja

Variable Trabaja

En forma de Tabla **Gráfico de Árbol**

Trabaja	Dar_Credito
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
Si	Si
Si	Si
Si	No
Si	No
Si	Si



Calculamos la Entropía de cada rama y luego las sumamos **multipicadas por la proporción de casos con respecto al nodo superior**:

Rama izquierda (**Trabaja= No**):

- cantidad de casos:8
- **proporción de casos con respcto al nodo superior: 8/13**
- cantidad de No: 8
- cantidad de Si: 0

Su entropía será:

$S1_{Trabaja\_izq}$ = -[8/8 $\log_2(8/8)$ + 0 $\log_2(0)$]

Out[19]:

```
-0.0
```

Rama derecha (**Trabaja=Si**):

- cantidad de casos: 5
- **proporción de casos con respcto al nodo superior: 5/13**
- cantidad de No: 2
- Cantidad de Si: 3

Su entropía será:

$S1_{Trabaja\_der} = - [2/5 \log_2(2/5) + 3/5 \log_2(3/5)]$

Out[20]:

```
0.971
```

Finalmente debemos sumar las entropías de cada rama **multipicadas por la proporción de casos con respecto al nodo su

$S1_{Trabaja} = - 8/13 S1_{Trabaja\_izq} - 5/13 S1_{Trabaja\_der}$

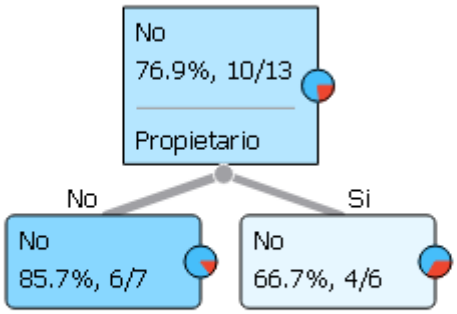
```
La Entropía al usar Trabaja para particionar el conjunto de datos es: S1Tr
abaja  0.373
```

Repitamos el proceso para la variable Propietario

Variable Propietario

En forma de Tabla **Gráfico de Árbol**

Propietario	Dar_Credito
No	Si
No	No
No	No
No	No
No	No
No	No
No	No
Si	No
Si	No
Si	Si
Si	No
Si	No
Si	Si



Calculamos la Entropía de cada rama y luego las sumamos **multipicadas por la proporción de casos con respecto al nodo superior**:

Rama izquierda (**Propietario= No**):

- cantidad de casos:7
- **proporción de casos con respcto al nodo superior: 7/13**
- cantidad de No: 6
- cantidad de Si: 1

Su entropía será:

$S1_{Propietario\_izq}$ = - [6/7 $\log_2$(6/7) + 1/7 $\log_2$(1/7)]

Out[22]:

```
0.592
```

Rama derecha (**Propietario=Si**):

- cantidad de casos: 6
- **proporción de casos con respcto al nodo superior: 6/13**
- cantidad de No: 4
- Cantidad de Si: 2

Su entropía será:

$S1_{Propietario\_der}$ = - [4/6 $\log_2$(4/6) + 2/6 $\log_2$(2/6)]

Out[23]:

```
0.918
```

Finalmente debemos sumar las entropías de cada rama **multipicadas por la proporción de casos con respecto al nodo superior**

$S1_{Propietario}$ = - 7/13 $S1_{Propietario\_izq}$ - 6/13 $S1_{Propietario\_der}$

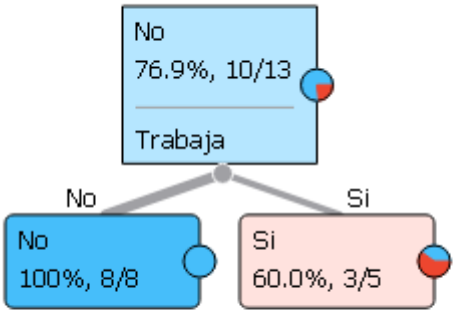
```
La Entropía al usar Propietario para particionar el conjunto de datos es:
S1Propietario=  0.742
```

Ahora que hemos calculado las cuatro posibles maneras de particionar el árbol, veamos en cuál de ellas la entropía o incertidumbre disminuyó más:

```
Sinicial=  0.779
-----------------------
S1_Tiene_Deuda=  0.587
S1_Genero=  0.742
S1_Trabaja=  0.373
S1_Propietario=  0.742
```

Como podemos observar **la mayor caída en la incertidumbre o el mayor aumento en la información** se produce cuando particionamos el árbol con la variable Trabaja; y por lo tanto con ella comenzaremos a armar nuestro árbol.

Primer nivel del árbol



La rama correspondiente a Trabaja = No, desemboca en una situación no hay nada más para indagar (los 8 casos son No); allí finaliza esa rama y se denomina **hoja**, pero la rama correspondiente a Trabaja=Si, aún deberá ser particionada con otra variable para seguir avanzando.
Repita el proceso anterior con la fórmula de la entropía para decidir cuál de las variables consigue una mayor disminución en la entropía o una mayor ganancia de información. Y luego siga hasta terminar el árbol.

Segundo nivel

Ahora estamos ubicados en la rama derecha de la imagen anterior, es decir la rama para Trabaja = Si

Trabaja	Dar_Credito
Si	Si
Si	Si
Si	No
Si	No
Si	Si

Como podemos ver nuestra situación mejoró muchísimo con respecto a la situación inicial, ahora tenemos que resolver para sólo 5 observaciones.

Cálculo de la entropía del nodo inicial del nivel 2

En este caso:

- En total son 5 observaciones
- 2 son No
- 3 son Si

Calculemos las proporciones de No y de Si:

p_{No} = 2/5
p_{Si} = 3/5

Calculamos la Entropía Inicial del nivel 2 con la fórmula de Shannon:

S_2 = - [2/5 $\log_2$(2/5) + 3/5 $\log_2$(3/5)]

```
La Entropía Inicial de nivel 2 es:  0.779
```

Nivel 2: Partición con la variable Tiene_Deuda:

id	Tiene_Deuda	Dar_Credito
2	No	Si
3	No	Si
5	No	No
12	No	Si
9	Si	No

Hacer el árbol para tiene_deuda etc.

> Calculamos la Entropía de cada rama y luego las sumamos **multipicadas por la proporción de casos con respecto al nodo superior**:

Rama izquierda (**Tiene_Deuda= No**):

- cantidad de casos:8
- **proporción de casos con respcto al nodo superior: 8/13**
- cantidad de No: 5
- cantidad de Si: 3

Su entropía será:

$S1_{tiene\_deuda\_izq}$=-[5/8 $\log_2$(5/8) + 3/8 $\log_2$(3/8)]

Out[27]:

```
0.954
```

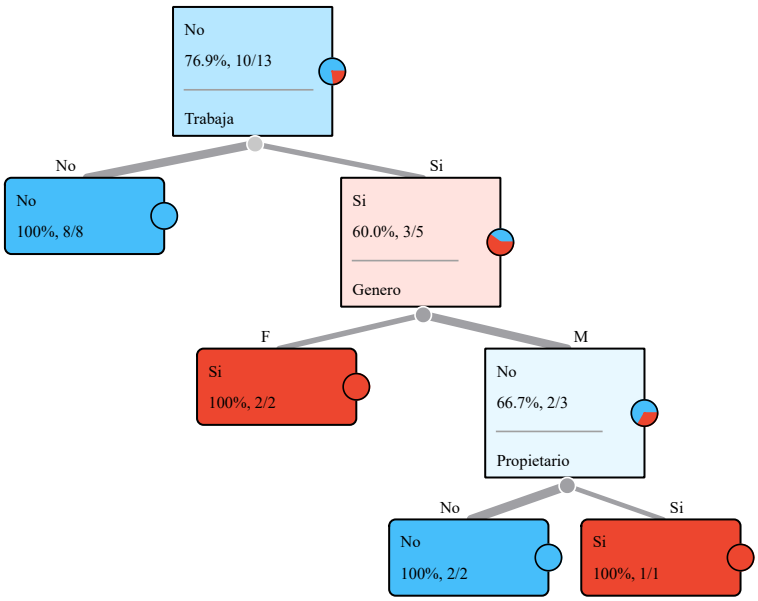
id	Genero	Dar_Credito
2	F	Si
3	F	Si
5	M	No
9	M	No
12	M	Si

id	Propietario	Dar_Credito
3	No	Si
5	No	No
9	No	No
2	Si	Si
12	Si	Si

EJERCICIO

Repetir el procedimiento y terminar de hacer el segudo nodo.

Finalmente el árbol terminado debería verse como el siguiente:


No
76.9%, 10/13
Trabaja
No
Si
No
100%, 8/8
Si
60.0%, 3/5
Genero
F
M
Si
100%, 2/2
No
66.7%, 2/3
Propietario
No
Si
No
100%, 2/2
Si
100%, 1/1