Introducción a la Inteligencia Artificial

Situación Profesional 1 - Clase 1

Ver en Video:

Título: [IA] Clase 1
link: https://www.youtube.com/watch?v=PH93rtOgYoc (https://www.youtube.com/watch?v=PH93rtOgYoc)
desde: inicio
hasta: final

Out[79]:

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[80]:

SITUACIÓN PROFESIONAL

Tenemos el siguiente problema que nos planteó un médico:

> "Para cierta enfermedad existe un **tratamiento** que si se empieza **tempranamente** dá muy **buenos resultados**, pero tiene un defecto, si se aplica a pacientes que **no** tienen la enfermedad puede **dañarle el hígado**.
> Existe un **estudio** que puede determinar con gran certeza si el paciente tiene o no la enfermedad, pero sus resultados **demoran varios meses**, entonces si espero a tener los resultados de los pacientes pierdo la oportunidad de aplicar tempranamente el tratamiento.
> Quiero saber si es posible hacer rápidamente un diagnóstico presuntivo a los **nuevos pacientes**, para comenzar el tratamiento con aquellos que pronostiquemos que podrían tener la enfermedad antes de que lleguen los resultados del estudio".

Una de las variantes más efectivas de la IA actual, el Aprendizaje Automático o Machine Learning, necesita datos de los cuales "aprender", supongamos que el médico nos provee de los siguientes datos correspondientes a un **análisis de sangre** que hizo a todos los pacientes que ha tenido con anterioridad y a los cuales, luego, les hizo el estudio que le daba certeza sobre su condición. Los datos están en un archivo con extensión csv que es una de las más usados para intercambiar datos.

Nota: Estos datos son ficticios.

```
Tabla 1
```

```
Out[82]:
```

	Sustancia_x1	enfermedad
0	1.0	no
1	13.0	si
2	11.5	si
3	2.0	no
4	6.5	no
5	8.0	no
6	9.0	no
7	12.0	si
8	4.0	no
9	11.0	si
10	2.0	no
11	4.5	no

Veamos los datos.

- Los datos corresponden a 12 pacientes (cada fila de la Tabla 1 corresponde a un paciente) para quienes tenemos la cantidad de cierta sustancia en sangre, la Sustancia_x1, y el diagnóstico obtenido varios meses después sobre si cada uno de los pacientes tenía o no la enfermedad. A cada fila solemos llamarla "observación' o 'caso'.
- La primer columna que contiene los números del 0 al 11 es sólo un **identificador** que automáticamente agregó el software (observe que la numeración comienza en cero, lo cual es típico del lenguaje de programación Python, en nuestro análisis no tendrá ninguna trascendencia) y sólo nos sirve para decir que el paciente identificado con el número 0 tiene 1.0 de la Sustancia_x1 y no tiene la enfermedad, en cambio el paciente identificado con el número 1 tiene 13.0 de la Sustancia_x1 y si tiene la enfermedad.

Este es un típico problema que se puede resolver las técnicas de Machine Learning: tenemos datos de la Sustancia_x1 de casos o pacientes previos de los cuales conocemos el resultado correcto, en este caso obtenidos luego del estudio que demoraba mucho, que indicaban si el paciente tenía no la enfermedad, y necesitamos **pronosticar** si **nuevos** pacientes tienen o no la enfermedad conociendo sólamente el valor de la cantidad de Sustancia_X1.

- Ahora, tómese su tiempo, analice los datos (papel y lápiz nunca están de más) y fíjese si puede establecer algún criterio o descubrir algún **patrón** en estos datos que le permitan ayudar al médico para **determinar cómo se relaciona la cantidad de Sustancia_x1 con el diagnóstico de la enfermedad**.

```
Tabla 2
```

Out[83]:

	Sustancia_x1	enfermedad
0	1.0	no
1	13.0	si
2	11.5	si
3	2.0	no
4	6.5	no
5	8.0	no
6	9.0	no
7	12.0	si
8	4.0	no
9	11.0	si
10	2.0	no
11	4.5	no

¿No está muy claro?

A veces algo tan simple como **ordenar** los datos con cierto criterio puede resultar revelador y nos permite ver algún **patrón** que de otra forma nos costaría ver. Le sugiero que ordene los datos según los valores de la columna **enfermedad**, quizá esto nos permita observar algo ...

Veamos ...

La siguiente Tabla tiene exactamente los mismo datos que la anterior, pero ordenados según el criterio indicado, observe que hemos puesto al principio a todos los pacientes (observaciones o casos dicho en forma general) que no tienen la enfermedad y luego a los que sí la tienen.

```
Tabla 3
```

Out[84]:

	Sustancia_x1	enfermedad
0	1.0	no
3	2.0	no
10	2.0	no
8	4.0	no
11	4.5	no
4	6.5	no
5	8.0	no
6	9.0	no
9	11.0	si
2	11.5	si
7	12.0	si
1	13.0	si

Parece **razonable** decir que:

- si el valor de la Sustancia_x1 es de hasta un valor de 10 aproximadamente el paciente **no** tiene la enfermedad
- si el valor de la Sustancia_x1 es mayor a 10 el paciente **sí** tiene la enfermedad

¿Qué podríamos decir si a un nuevo paciente cuando le hacen el análisis de sangre le dá que tiene 3.1 de la Sustancia_x1?

¿Le resultó más fácil ***aprender a partir de los datos*** cuál era la **regla o patrón** para decidir entre los No y los Si?

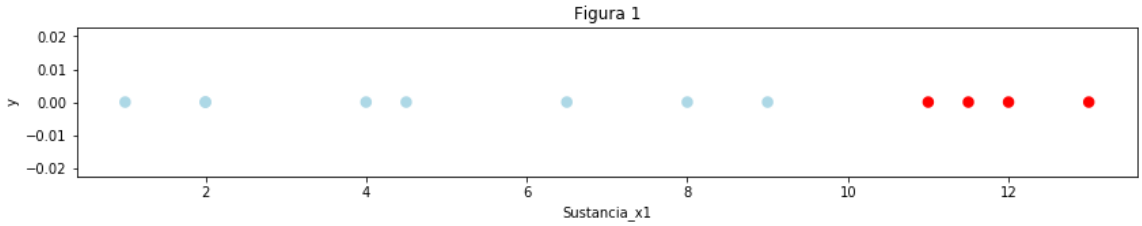
En Ciencia de Datos se dá mucha importancia a la **visualización** de los datos, vemos por qué.

La siguiente figura muestra los mismos datos que las tablas anteriores, en ella hemos dispuesto un eje horizontal con los valores de Sustancia_x1 y sobre él hemos marcado con puntos azules los casos en los cuales el valor de enfermedad es no, y con puntos rojos cuando el valro de enfermedad es sí.

En forma espontánea y sin esfuerzo, podemos observar que los puntos azules dejan de aparecer cuando el valor de la Sustancia_x1 ronda en valores cercanos a 10 y partir de allí comienzan a aparecer los puntos rojos.

Out[85]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x26cbb3bb1d0>
```



EJERCICIO 1

La situación es la misma del problema anterior, pero los datos que se obtienen de los análisis de sangre son los que se muestran en la siguiente tabla.
Descubra Ud el patrón que se esconde en los datos.

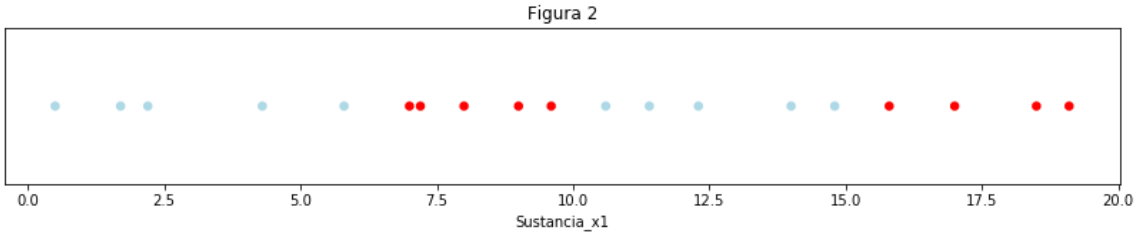
```
Tabla 4
```

Out[86]:

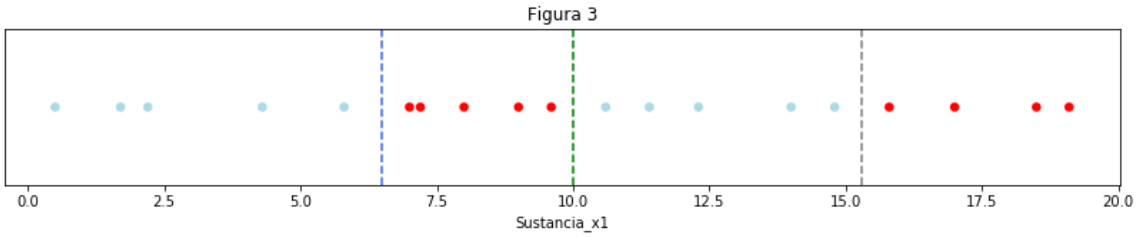
	Sustancia_x1	enfermedad
0	7.2	si
1	12.3	no
2	9.0	si
3	19.1	si
4	2.2	m
5	14.0	no
6	4.3	no
7	11.4	no
8	8.0	si
9	17.0	si
10	1.7	no
11	18.5	si
12	7.0	si
13	10.6	no
14	0.5	no
15	15.8	si
16	9.6	si
17	14.8	no
18	5.8	no

SOLUCIÓN

Probemos haciendo un gráfico, ubiquemos los valores de la Sustancia x1 sobre un eje y marquemos con un punto azul cuando el valor de enfermedad es no y con un punto rojo cuando el valor de enfermedad es si:



Se ven claramente 4 zonas o rangos o intervalos, distintos. Veamos cuáles son los valores que limitan cada una de estas zonas:



A partir de esta información podríamos decir que:

- Si el valor de Sustancia_x1 es menor que 6,5 (podríamos haber elegido otro valor cercano), entonces **no** tiene la enfermedad,
- Si el valor x1 está entre 6,5 y 10, entonces **sí** tiene la enfermedad,
- Si el valor de x1 está entre 10 y 15,3, **no** tiene la enfermedad,
- Si el valor de x1 es mayor que 15,3 entonces, **sí** tiene la enfermedad.

Pregunta:

Qué podríamos **pronosticar** para un paciente que tuviera 13 para el valor de la Sustancia_x1?

Solución de la Situación Profesional, un problema con 2 variables.

En los problemas anteriores hemos tenido una sola variable, la cantidad de Sustancia_x1 en sangre para intentar pronosticar si el paciente tiene o no la enfermedad.
Generalmente nuestros problemas tendrán más de una variable.

En el caso del planteo de la Situación Profesional supondremos que es un problema de dos variables:en el Análisis de Sangre además del valor de la Sustancia_x1, se medía también el de otra que denominaremos Sustancia_x2 (de ahora en más las abreviaremos como x_1 y x_2).

Carguemos los datos, e intentemos ver si podemos descubrir la regla o patrón que nos permitiría determinar cuándo el resultado es Si y cuándo es No.

```
Tabla 5
```

Out[89]:

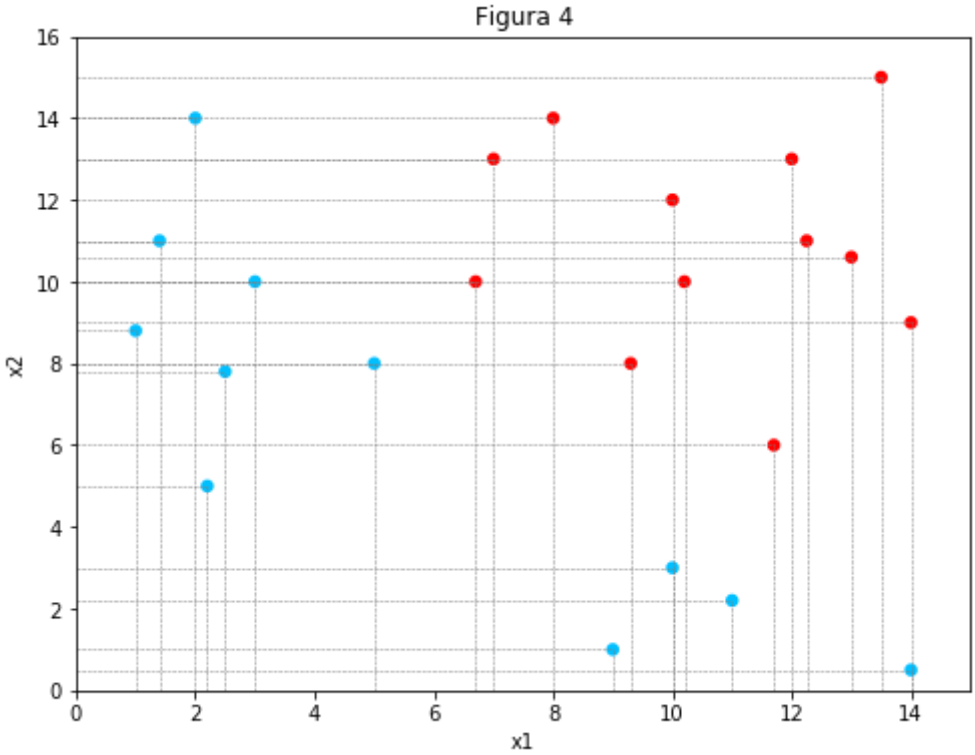
	x1	x2	enfermedad
0	5.00	8.0	No
1	2.20	5.0	No
2	11.70	6.0	Si
3	1.40	11.0	No
4	1.00	8.8	No
5	14.00	9.0	Si
6	9.00	1.0	No
7	11.00	2.2	No
8	10.20	10.0	Si
9	10.00	3.0	No
10	9.30	8.0	Si
11	10.00	12.0	Si
12	13.00	10.6	Si
13	14.00	0.5	No
14	13.50	15.0	Si
15	6.70	10.0	Si
16	2.50	7.8	No
17	12.00	13.0	Si
18	12.25	11.0	Si
19	2.00	14.0	No
20	7.00	13.0	Si
21	8.00	14.0	Si
22	3.00	10.0	No

Veamos cómo podemos analizarlo.

Ahora tenemos dos variables x_1 y x_2 con las que prentedemos pronosticar un resultado, el valor de la columna enfermedad.
Con los ejemplos anteriores hemos aprendido que la **visualización** suele ser de gran ayuda, y en el secundario Ud aprendió a hacer gráficos cartesianos.

Ubiquemos en el plano x_1 x_2 los distintos puntos para las combinaciones de valores de x_1 y x_2 y asignemos un color distinto si para esa combinación el resultado fue de enfermedad o no, en este caso celeste para los no y rojo para los que si tienen la enfermedad.

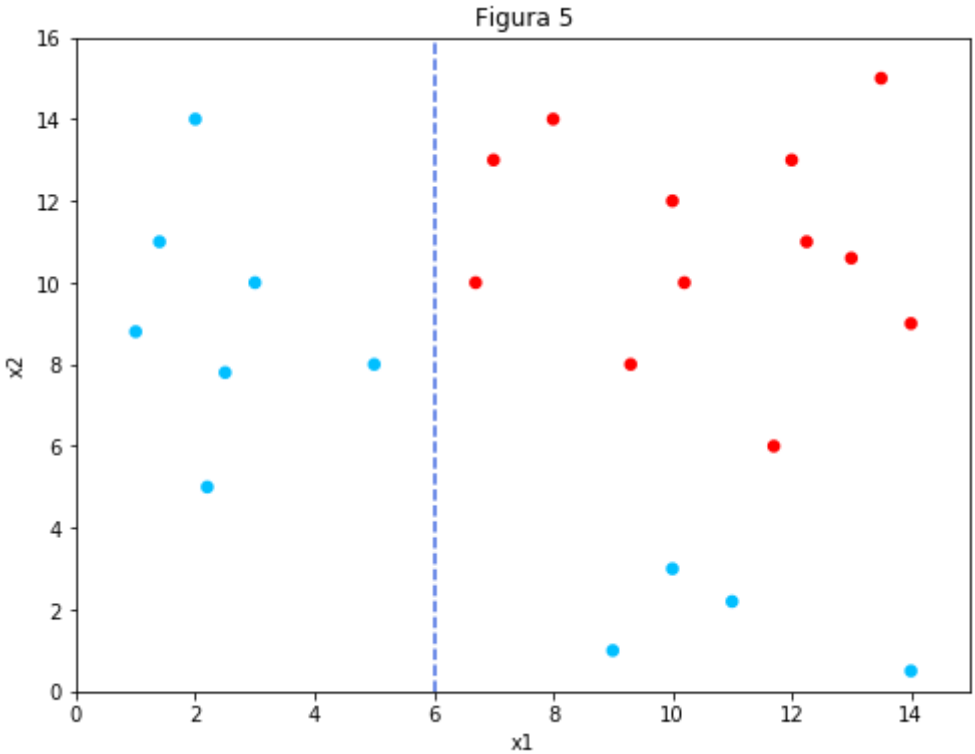
```
<Figure size 432x288 with 0 Axes>
```



En la figura podemos ver que de alguna manera los puntos celestes (no) y los rojos (si) **están "separados"** o son separables.

Podemos pensar en proceder de la siguiente forma.

```
<Figure size 432x288 with 0 Axes>
```



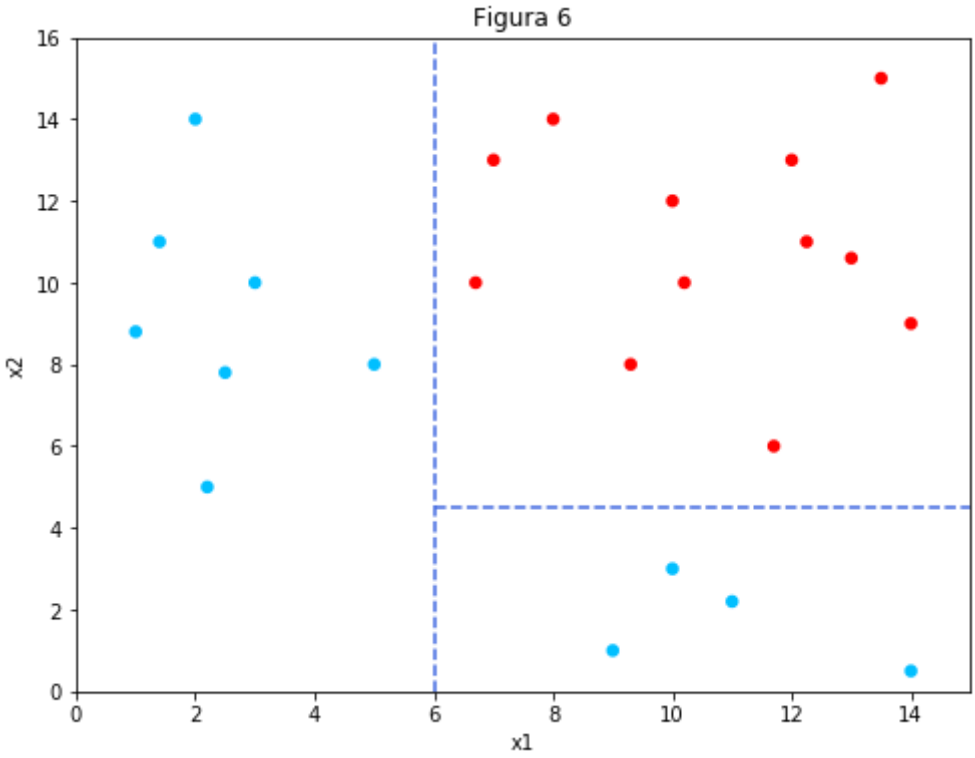
Hemos trazado un línea vertical cuando x1 vale aproximadamente 6, hemos tomado al 6 con cierta arbitrariedad, podría haber sido un valor algo más a la izquierda o algo más a la derecha. Observemos que para todos los puntos del plano que se encuentran a la izquierda de la recta se cumple que su componente x_1 es menor que 6, x_1 <= 6 y podemos ver que todos los casos que se encuentra a la izquierda de esta recta, corresponden a pacientes que no tienen la enfermedad.

> **REGLA 1:** Podemos decir que **si un paciente tiene cantidad de sustancia x_1 <= 6, entonces no tiene la enfermedad** y hemos encontrado una regla o patrón que nos permite resolver el problema para unos 7 puntos del total.

Qué pasa a la derecha de la recta, es decir cuando x_1 > 6?

Por este lado no está tan claro qué pasa porque tenemos tanto puntos rojos como puntos celestes; sin embargo si miramos bien, podemos pensar algo como lo siguiente:

```
<Figure size 432x288 with 0 Axes>
```



Ahora hemos trazado una recta horizontal a la altura de aproximadamente x_2 = 4,5 la cual divide al sector derecho en exactamente dos grupos de puntos, los celestes por debajo y los rojos por arriba y hemos conseguido dividir el plano x_1 x_2 en tres zonas, en cada una de las cuales sólo hay puntos de un color o lo que es lo mismo valores de la misma categoría o clase para la enfermedad.

Ejemplo:

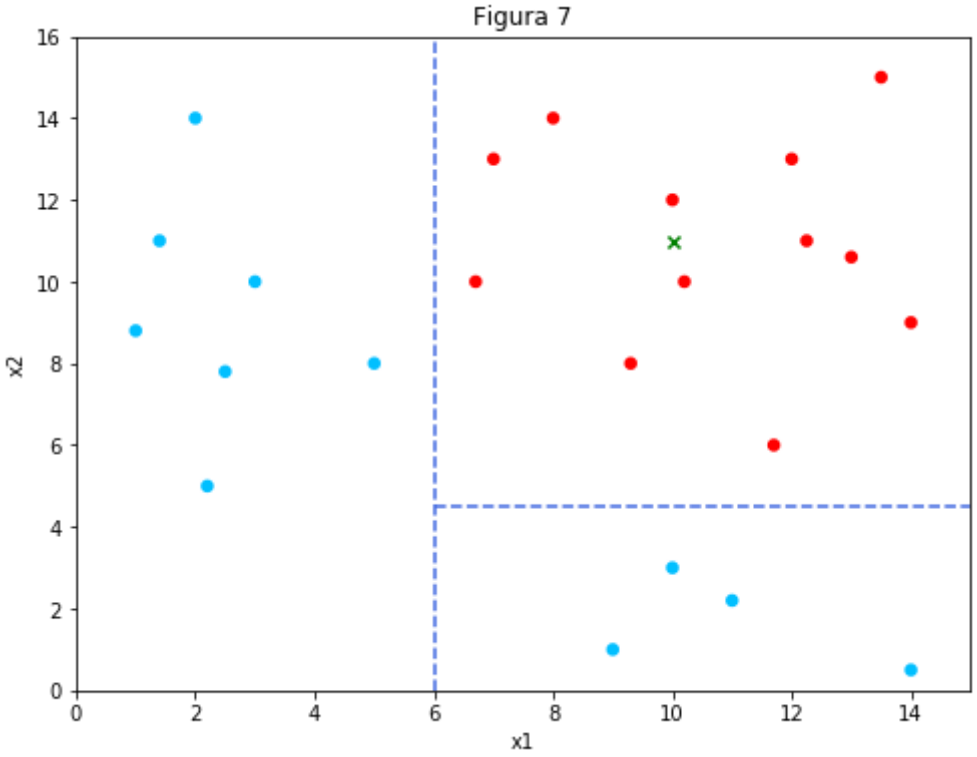
Qué podríamos decir de un nuevo paciente a quien los resultados del análisis de sangre le dieron:

- $x_1 = 10$
- $x_2 = 11$

Respuesta: que tiene la enfermedad

Para resolverlo basta con ubicar el punto formado por (x_1, x_2) y observar que en zona está:

```
<Figure size 432x288 with 0 Axes>
```



Hemos marcado con una **x** de color verde al punto correspondiente a los valores del nuevo paciente y observamos que está en la zona de puntos rojos, por lo tanto, tiene la enfermedad.

Ejercicio:

Qué podríamos decir de un nuevo paciente a quien los resultados del análisis de sangre le dieron:

- $x_1 = 11$
- $x_2 = 1,5$

Respuesta: que no tiene la enfermedad

Ahora podemos decir que después de aplicar la regla anterior, es decir para los puntos con x1 > 6, hemos establecido una segunda regla:

> **REGLA 2:** Si la **cantidad de sustancia x_2 <= 4,5 entonces el paciente no tiene la enfermedad y si x_2 > 4,5; la tiene**.

Si se fija bien en nuestra figura luego de aplicar estas dos reglas no nos quedaría ningún punto sin analizar, es decir que hemos resuelto todo el problema.

PATRÓN DESCUBIERTO EN LOS DATOS

A partir del simple gráfico hemos podido descubrir el **patrón** (pattern) que existía en el conjunto de datos, ahora resumamos nuestro análsis y explicitemos las **reglas** que conforman el **patrón** que hemos descubierto:

> 1) **si un paciente tiene cantidad de sustancia x_1 <= 6, entonces no tiene la enfermedad**

Si el paciente tiene x_1 > 6, entonces pueden darse dos situaciones:

> 2) Si la **cantidad de sustancia x_2 <= 4,5 entonces el paciente no tiene la enfermedad y si x_2 > 4,5; la tiene**

Fue **muy fácil** descubrir el **patrón** escondido en los datos de la Tabla 5, no es cierto?

Observe que hemos aplicado conceptos básicos de matemática que aprendió en el colegio secundario para resolver un problema que sin ellos hubiera requerido una inteligencia notable, en nuestro caso, lo inteligente estuvo en el "modelo" matemático que utlizamos para resolverlo. Esto es algo habitual en IA, hay muchos modelos matemáticos que no son muy complicados de entender pero que aplicados correctamente resuelven problemas de gran complejidad.

ÁRBOLES DE DECISIÓN

Recordemos las reglas que describen el patrón que hemos descubierto en el caso anterior:

> 1) **si un paciente tiene cantidad de sustancia x_1 <= 6, entonces no tiene la enfermedad**

Si el paciente tiene x_1 > 6, entonces:

> 2) Si la **cantidad de sustancia x_2 <= 4,5 entonces el paciente no tiene la enfermedad y si x_2 > 4,5; la tiene**

Existe una manera muy clara y sencilla de resumir estas reglas y que adquiere más trascendencia cuando el patrón consta de muchas reglas y es la representación en forma de **Árbol de Decisión**, uno de los modelos básicos y más claros que utilizaremos en IA.

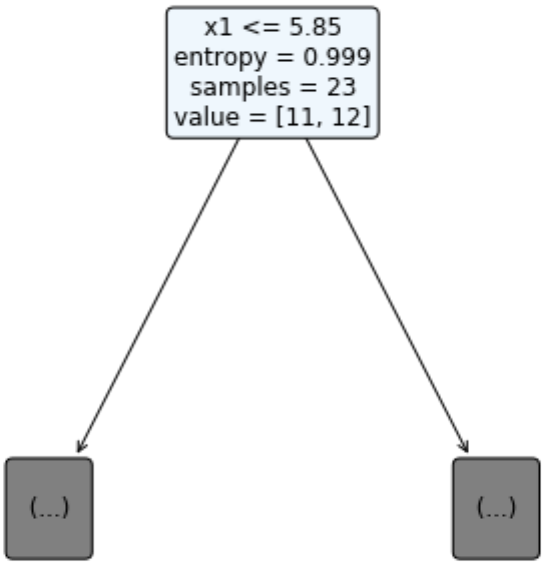
Vamos a construir el Árbol de Decisión para nuestro caso:

Veamos la primer Regla:

> Si x_1 <= 6 , esta premisa tiene dos posibilidades: que se cumpla, es decir que el valor de x_1 sea efectivamente menor o igual que 6 o que no se cumpla, en el caso que x_1 sea mayor que 6.

Esto lo representaremos de la siguiente manera:

Figura 8



Este árbol de decisión fue realizado programando sobre Python, generalmente las salidas son profusas en datos pero bastante escuetas en su descripción, así que hay que ir aprendiendo cómo se lee este tipo de diagrama.

El rectángulo se denomina **nodo** (el primer nodo del árbol se denomina **raíz**) y el software que hemos usado incluye mucha información:

- **samples** = 23, es la cantidad de **casos**, observaciones o filas que tienen nuestros datos, en este caso son 23.
- **value** = [11,12], nos indica que los valores de la columna que queremos pronosticar (enfermedad) están formados por 11 no y 12 con valor si (los coloca en orden alfabético, por eso el primer número corresponde a los no y el segundo número corresponde a los si)
- **entropy (entropía)** = 0,99. Conceptualmente indica el grado de **incertidumbre** que tenemos. El valor 0,999 es alto, porque a esta altura del análisis sólo sabemos que hay 11 observaciones con valores de no y 12 con resultado si: si nos dieran un nuevo caso, qué valor de enfermedad le pronosticaríamos? Están tan parejos los no y los si que tendríamos una gran incertidumbre para pronosticar el resultado. Distinto sería si tuviéramos 100 casos que tuvieran valor no y sólo uno que tuviera valor si, en esa situación si tuviéramos que pronosticar qué resultado arrojaría una nueva observación diríamos que casi seguro es un no. **Cuando los resultados se encuentran en la misma cantidad, la entropía es máxima**. Más adelante en esta materia veremos cómo se calcula el valor numérico de la entropía, pero lo más importante será recordar el concepto.

Ahora sí lo más importante:

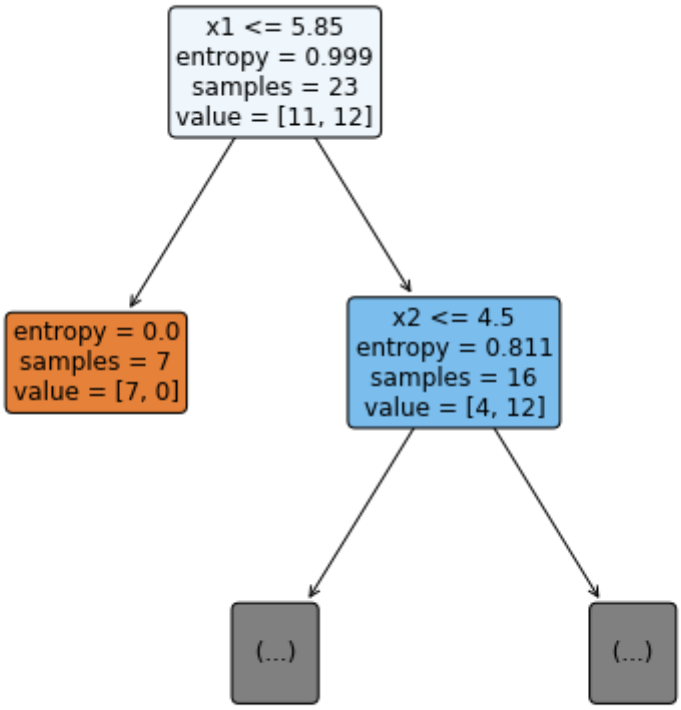
- **x_1 <= 5,85**: nos indica el valor donde dividió; nosotros "a ojo" habíamos tomado 6 para comparar, pero el algoritmo encontró que el más conveniente es 5,85, ya veremos más adelante por qué.
- Puede interpretar **x_1 <= 5,85** como una pregunta: Se cumple que **x_1 <= 5,85**?

Luego vemos dos flechas en la parte inferior denominadas **ramas** (por la analogía con **árbol**) y a no ser que se indique lo contrario:

> la flecha de la **izquierda** corresponde a la respuesta **afirmativa** es decir cuando se cumple la condición **x_1 <= 5,85**
>
> la flecha de la **derecha** corresponde a la respuesta **negativa**, es decir cuando **no se cumple** que **x_1 <= 5,85**, o lo que es lo mismo cuando **x_1 > 5,85**.

Expandamos un nivel más de profundidad a nuestro árbol:

Figura 9



Veamos primero la flecha de la izquierda, la salida afirmativa, es decir para aquellos casos en que x_1 <= 5,85.
Nuestra Regla 1 decía que para todos estos casos correspondían al valor **no** para enfermedad.

Observe en el nodo de la izquierda que nos indica:

- entropía: 0. En esta nodo no hay incertidumbre porque **todos** los casos tienen el mismo valor: no.
- casos: 7. Son los 7 casos que correspondían a x_1 <= 5,85
- value: nos dice que hay 7 casos no, y, 0 casos si.

No se desprenden flechas de este nodo, porque todos los casos han sido clasificados (los 7 casos eran no).

Este tipo de nodo "terminal" se denomina **hoja**.

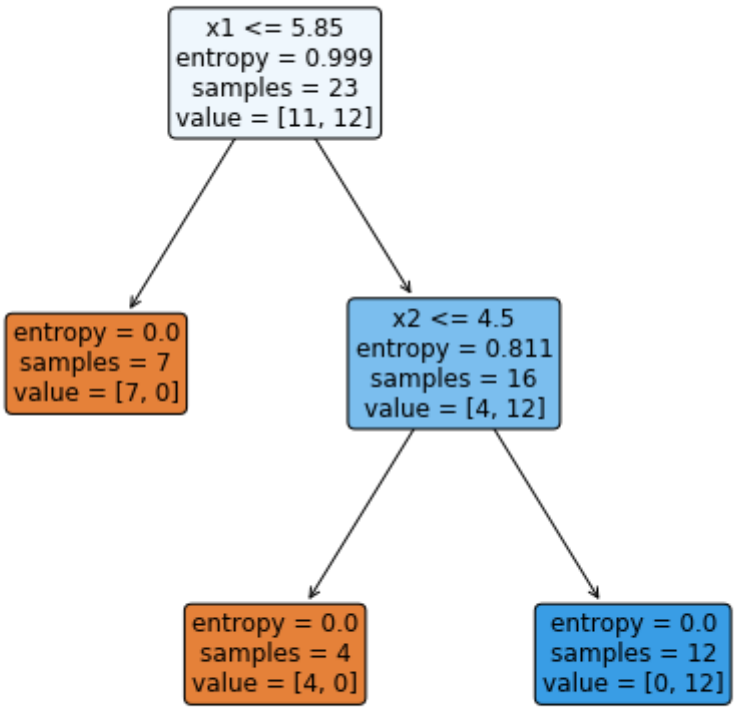
Veamos que pasa en la rama derecha, es decir cuando no se cumple que x_1 <= 5,85 o lo que es lo mismo que x_1 > 0,85.

- samples: 16. Son las 16 observaciones que quedaron del lado derecho.
- value: [4,12] nos indica que de las 16 observaciones, 4 son no y 12 son si.
- entropía: 0,811. Fíjese que es menor que en el nodo raíz, pero no es 0 como en la hoja de la izquierda, estos es así porque en este nodo tenemos algunos casos no y otros casos si lo cual nos indica que hay un cierto nivel de incertidumbre pero menos que al principio, esto es así porque las cantidades de no y si no están tan parejas como al principio.

Finalmente, aunque figura al principio, como tenemos observaciones con ambos valores, tendremos que hacer una nueva división, esta vez con el valor de x_2, y el planteo es si x_2 <= 4.5, de aquí partirán dos ramas, la de izquierda cuando se cumple y la derecha cuando no se cumple.

Expandamos este nodo y veamos cómo queda:

Figura 10



En la rama izquierda vemos que de los 4 casos que cumplen la condición x_2 <= 4,5, todos dan no, por lo cual, no hay incertidumbre, la entropía es 0 y ese nodo es una hoja porque no hay nada más que analizar.

En la rama derecha, nos indica que quedaron 12 casos, todos que corresponden al valor sí, por lo cual tampoco hay incertidumbre, lo que equivale a decir que la entropía es cero y por lo tanto también es un nodo terminal u hoja.

Una vez que se llega a todos nodos terminales u hojas, se acabó el árbol el cual representa las reglas que nos permiten determinar a cuál de los valores de enfermedad corresponden las distintas combinaciones posibles de valores de x_1 y x_2 y estamos en condiciones de pronosticar si un nuevo paciente tiene o no la enfermedad aplicando el árbol.

EJEMPLO

Apliquemos el árbol para determinar si un paciente con valores de $x_1 = 7$ y $x_2 = 9$ tiene o no la enfermedad.

Comenzamos en el primer nodo, el nodo raíz; en él se nos plantea si el valor de x_1 es menor o igual a 5,85. En nuestro caso x_1=7, así que la respuesta es negativa, y por lo tanto deberemos tomar el camino de la rama derecha.

En el siguiente nodo en nuestro camino se nos pregunta ahora por el valor de x_2, es x_2 <= 4,5? En nuestro caso la respuesta es negativa ya que nuestro valor de x_2 es 9, por lo tanto debemos seguir por la rama de la derecha, y llegamos a un nodo terminal u hoja, en la cual nos dice que todos los casos en ella son sí; por lo tanto nuestro pronóstico es que si tiene la enfermedad.

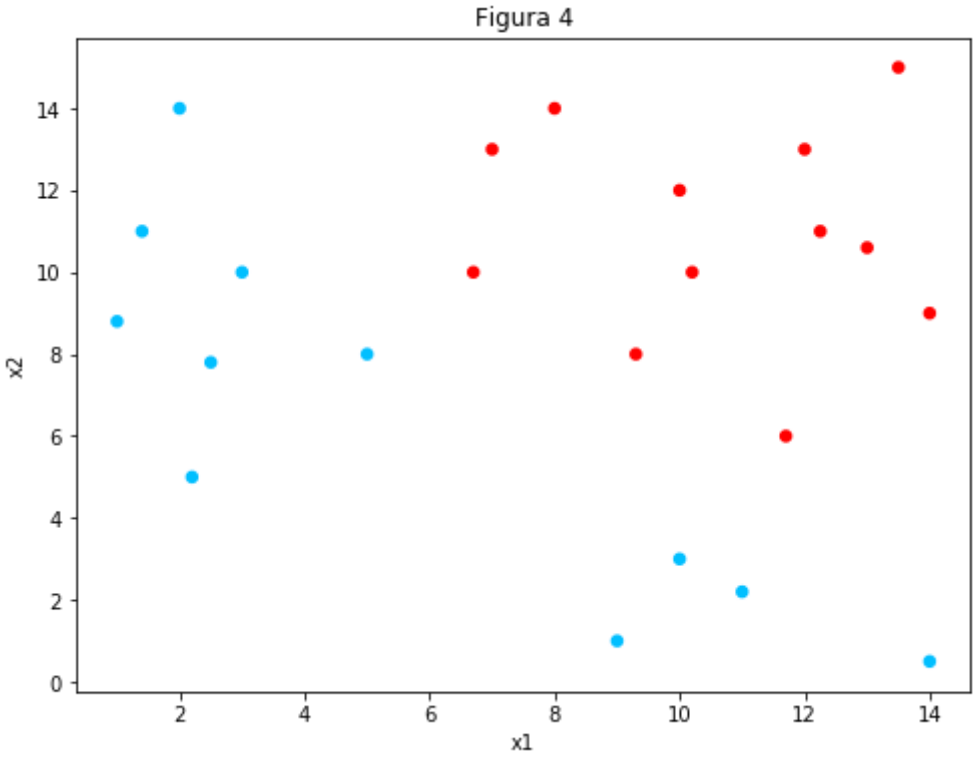
La idea de subdividir el plano convenientemente con **rectas paralelas a los ejes coordenados** y el esquema de árbol de decisión son modelos equivalentes, con una excepción: podemos utilizar el modelo de árbol de decisión para resolver problemas que tengan muchas variables, pero no podemos extender nuestra intuición geométrica de subdividir el plano o el espacio a más de 3 dimensiones o variables. Así que cuando tengamos 10, 20 o mil variables el modelo de árbol seguirá siendo totalmente funcional y sería bueno que Ud mantenga la intuición geométrica de lo que ocurre en el plano, aunque sea imposible de imaginar!

Por dónde dividir?

Finalmente vamos a destacar un detalle que no mencionamos con anterioridad para no apartarnos del centro de la explicación, cuando comenzamos a resolver nuestro problema gráficamente, primero comenzamos a ubicar los puntos en el plano de las variables x_1 x_2 como se muestra en la figura 4 que repetimos a continuación:

```
Figura 11
```

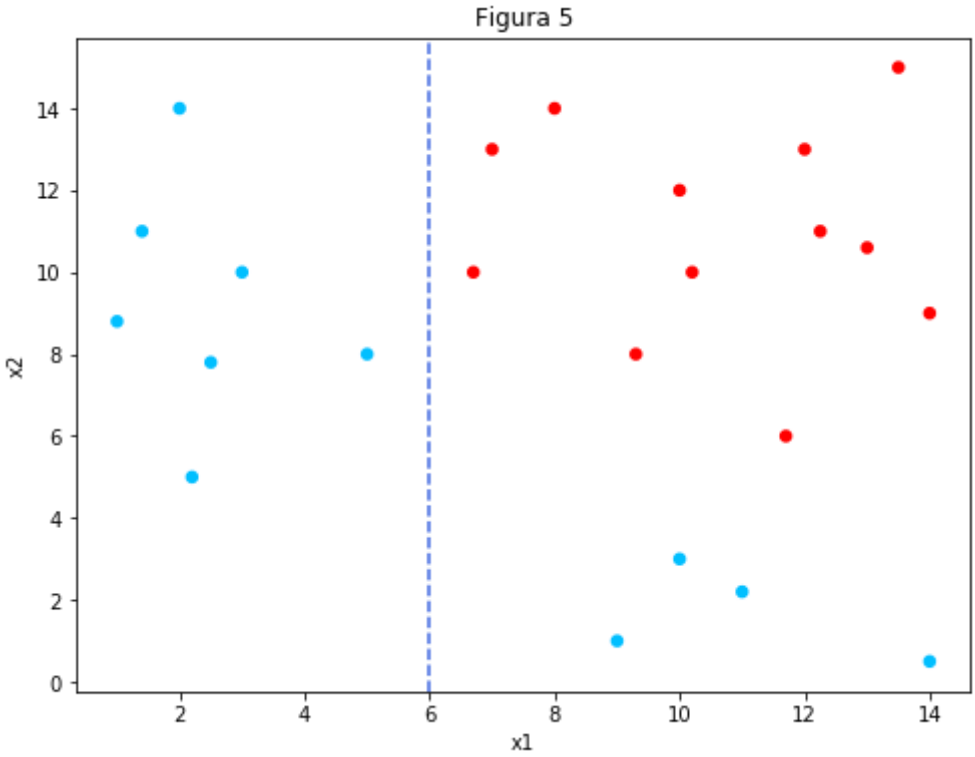
```
<Figure size 432x288 with 0 Axes>
```



La primer división la hicimos con una recta vertical en el valor de $x_1 = 6$, de la siguiente manera:

```
Figura 12
```

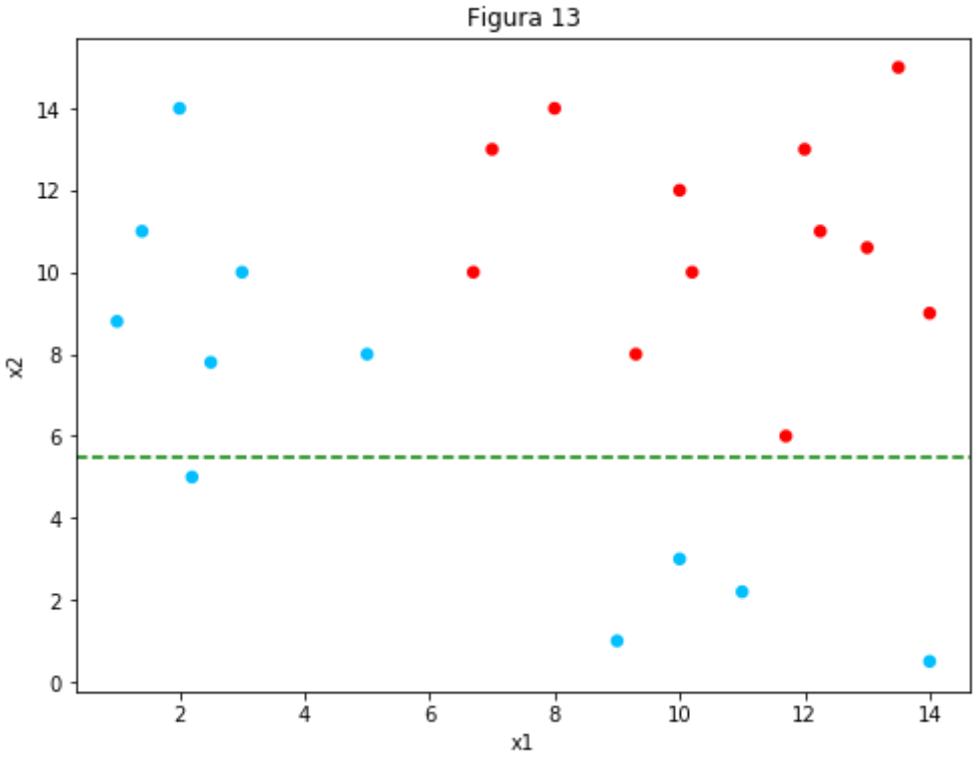
```
<Figure size 432x288 with 0 Axes>
```



Sin embargo, podríamos haber comenzado con una división como la siguiente:

```
Figura 13
```

```
<Figure size 432x288 with 0 Axes>
```



La línea verde muestra que podríamos haber comenzado a subdividir con una recta horizontal a la altura aproximada de $x_2 = 5{,}5$.

La pregunta es por qué resultaba más conveniente comenzar a subdividir el plano con la recta vertical en vez de la horizontal?

La respuesta está en la cantidad de observaciones o puntos que separamos en cada caso:

- al subdividir verticalmente en $x_1 = 6$ dejamos 7 puntos (todos azules, incertidumbre = 0) a la izquierda y a la derecha mezclados los rojos y azules, es decir con un cierto grado de incertidumbre "remanente";
- si hubiéramos comenzado a subdividir con la recta horizontal en $x_2 = 5{,}5$, hubiéramos dejado los 5 puntos de la parte inferior sin incertidumbre, y todos los de arriba mezclados. Como en este segundo caso, sólo hubiéramos "resuelto el problema" para 5 de los puntos en vez de los 7 que conseguíamos con la recta vertical, preferimos la vertical ya que nos dejaría menos incertidumbre para el paso siguiente.

EJERCICIO

Ahora le toca a Ud realizar un análisis completo, con la subdivisión del plano y la generación del diagrama de árbol equivalente.
En este momento deberá resolver el problema con "papel y lápiz", para obtener la comprensión de lo que estamos haciendo; más adelante en esta misma materia utilizará un software para resolver el problema más rápidamente y podrá aplicarlo a situaciones con muchas variables y muchas observaciones.

A continuación le mostramos los datos que el médico ha conseguido recopilar de pacientes anteriores.

```
Tabla 6
```

Out[101]:

	x1	x2	enfermedad
0	11.5	8.0	No
1	12.0	4.0	No
2	1.0	2.0	Si
3	3.0	6.0	No
4	2.5	3.5	Si
5	7.0	7.0	No
6	3.0	4.5	Si
7	11.0	3.0	No
8	4.0	2.5	Si
9	8.5	7.5	No
10	10.0	5.5	No
11	5.0	1.0	Si
12	6.0	4.5	Si
13	14.0	3.0	No
14	7.0	2.5	Si
15	13.5	8.0	No
16	9.0	3.0	Si
17	0.5	8.0	No
18	10.5	6.5	No
19	2.1	7.5	No
20	4.0	4.0	Si
21	4.0	6.5	No
22	8.0	5.5	No
23	11.0	1.0	No
24	8.0	2.0	Si
25	12.0	6.0	No
26	12.0	2.0	No
27	2.0	1.0	Si
28	13.0	7.0	No
29	13.0	3.5	No

El médico necesita pronosticar si los siguientes pacientes tienen o no tienen la enfermedad conociendo sus valores de x_1 y x_2:

paciente_1: x_1 = 12,33; x_2 = 3

paciente_2: x_1 = 6 , x_2 = 2,5

paciente_3: x_1 = 8 , x_2 = 8

Usted debe:

- Primero obtener una solución gráfica
- Luego, obtener el Árbol de Decisión basado en los datos para aplicarlo a estos tres pacientes y a cualquier otro que pudiera presentarse.

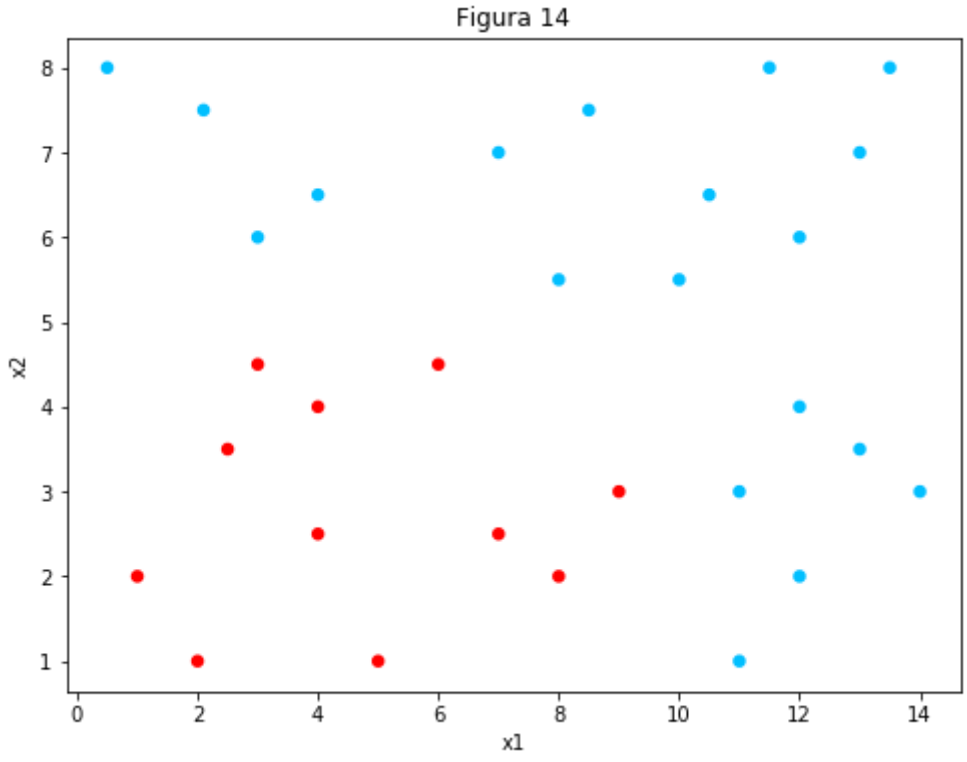
Nota: Cuando arme el árbol de decisión no podrá calcular la entropía ya que aún no le hemos indicado como se calcula, así que ignórela, pero sí podrá indicar el resto de los valores que hemos visto en los árboles anteriores.

SOLUCIÓN

Subdividamos

Comencemos graficando los pares de valores de (x_1,x_2) en el plano, y asignemos un color (rojo) al valor de enfermedad = Si y otro color (celeste) al valor de enfermedad = No

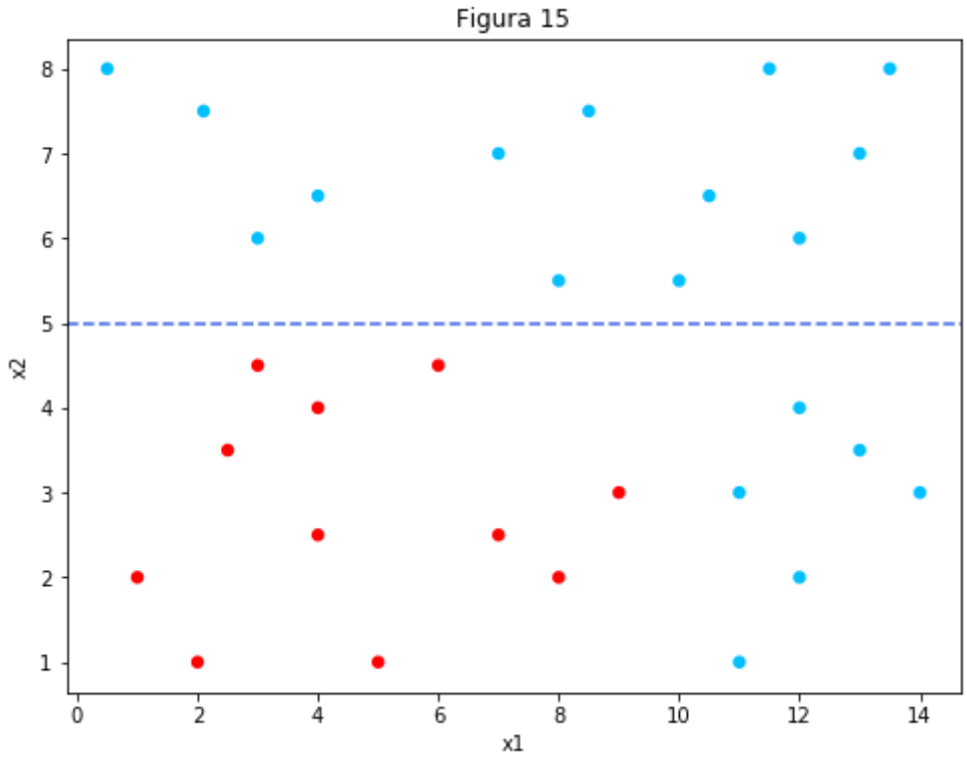
```
<Figure size 432x288 with 0 Axes>
```



Ahora procedamos a subdividir el plano, ya sea con una línea vertical o con una línea horizontal, procurando que nuestro "corte" separe la mayor cantidad de casos del mismo color.

Yo no he encontrado ninguna línea vertical que sirva, así que pienso comenzar con una división horizontal a la altura de $x_2 = 5$, con lo cual nuestra primer subdivisión se vería de la siguiente manera:

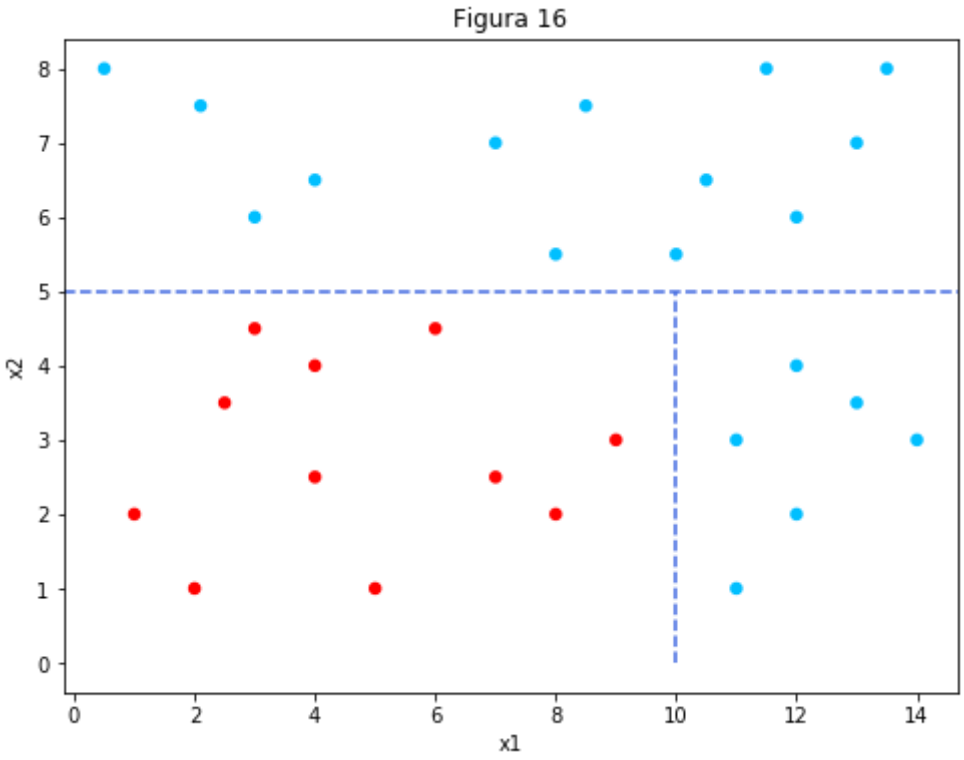
```
<Figure size 432x288 with 0 Axes>
```



Con este "corte " hemos separado los 13 casos No de la parte superior de los 17 casos inferiores donde hay tanto celestes como rojos.

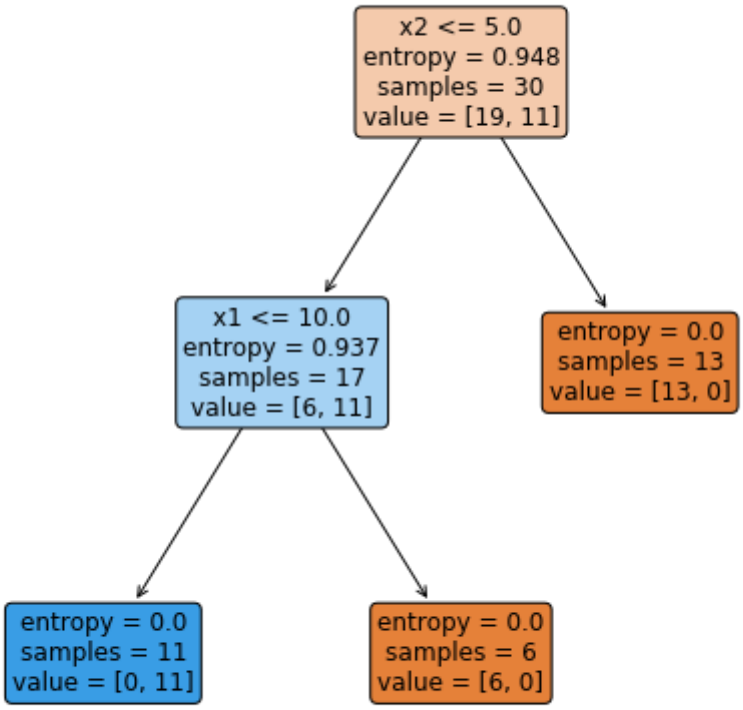
Ahora parece ser que haciendo un corte vertical a la altura de $x_1 = 10$ podría quedar solucionado nuestro problema.

```
<Figure size 432x288 with 0 Axes>
```



Creemos el Árbol de Decisión

Figura 17



Pronóstico para los pacientes nuevos:

paciente_1: x_1 = 12,33; x_2 = 3 >>> enfermedad: No

paciente_2: x_1 = 6 , x_2 = 2,5 >>> enfermedad: Si

paciente_3: x_1 = 8 , x_2 = 8 >>> enfermedad: No

Felicitaciones, hemos hecho Inteligencia Artificial "a mano"!

Sólo nos restaría imprimir el diagrama de árbol y explicarle a nuestro médico amigo cómo usarlo!