Matemáticas básicas en la Ciencia de Datos y la Inteligencia Artificial

INDICE

El algebra lineal, la estadística y las probabilidades son herramientas fundamentales en la ciencia de datos.

Ejemplo del sector inmobiliario:

	AREA	NUMERO HABITACIONES	PRECIO
Vivienda 1	75m2	3	200,000 euros
Vivienda 2	100m2	4	300,000 euros

Podemos establecer un sistema de dos ecuaciones lineales para estos datos

$75x + 3y = 200.000$

$100x + 4y = 300.000$

ALGEBRA LINEAL

Este sistema usa el algebra lineal para encontrar el valor de las dos variables x e y.

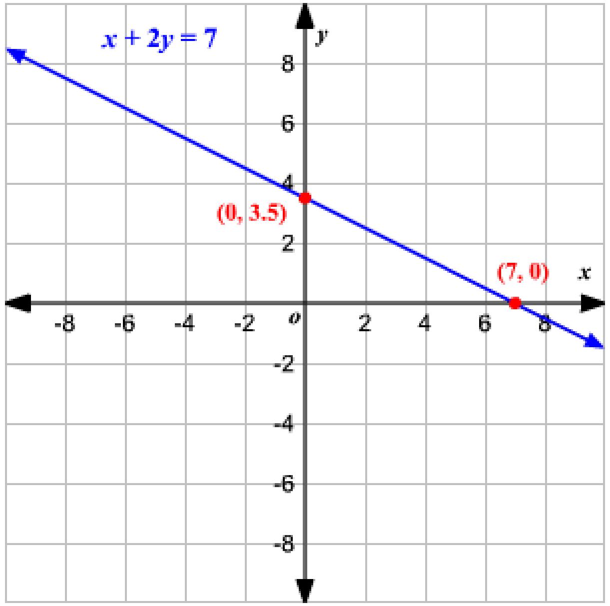
Podemos decir qué en nuestro ejemplo de las dos viviendas con sus metros cuadrados, números de habitaciones y precios podríamos crear la siguiente expresión matemática.

$$\begin{bmatrix} 75 & 3 \\ 100 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 200000 \\ 300000 \end{bmatrix}$$

matriz vector vector

ALGEBRA LINEAL

Crear grafica de una ecuación lineal x + 2y = 7



ALGEBRA LINEAL

- Crear grafica de una ecuación de segundo grado

$f(x)= x^2 + 4$

SISTEMAS DE ECUACIONES

Métodos para resolver sistemas de 2 y 3 ecuaciones

.- Método de sustitución

Ejemplo

$x + y = 6$

$x – y= 4$

Despejamos

$y = 6 – x$

Sustituimos

$x – (6 - x) = 4$ >>>> $x = 5$

La otra incógnita

$x + y = 6$ >>>> $y = 1$

SISTEMA DE ECUACIONES

.- Método de reducción

Ejemplo

$x + 2y = 25$

$2x + 3y = 40$

Multiplicamos la primera ecuación por -2 y la segunda la dejamos igual

$- 2x – 4y = - 50$

$2x + 3y = 40$

Sumamos las dos ecuaciones

$0 – y = -10$ >>>> $y = 10$

Sustituimos en la primera ecuación el valor de y

$x + 20 = 25$ >>> $x = 5$

SISTEMA DE ECUACIONES

.- Método de igualación

Ejemplo

$2x - y = -1$

$3x + y = 11$

Despejamos x o y en ambas ecuaciones

$y = 2x + 1$

$y = 11 - 3x$

Si los dos primeros miembros de las ecuaciones son iguales los segundos también lo son

$2x + 1 = 11 - 3x$ >>>> $x = 2$

Sustituimos en cualquiera de las dos ecuaciones

$2x - y = -1$ >>>> $4 - y = -1$ >>> $-y = -5$ >>> $y = 5$

SISTEMAS DE ECUACIONES

Resolución de sistema de 3 ecuaciones con 3 incógnitas

Método de sustitución: En este método, se despeja una de las incógnitas en una de las ecuaciones y se sustituye en las otras dos ecuaciones, reduciendo así el sistema a un sistema de 2 ecuaciones con 2 incógnitas. Luego, se utilizan técnicas de resolución de sistemas de 2 ecuaciones para encontrar los valores de las dos incógnitas restantes. Finalmente, se sustituyen los valores encontrados en la ecuación original para encontrar el valor de la tercera incógnita.

Ejemplo

$x + y + z = 6 \ggg\gg x = 6 - y - z$

$x - y + 2z = 5 \ggg\gg 6 - y - z - y + 2z = 5 \gg -2y + z = -1$

$x - y - 3z = -10 \ggg 6 - y - z - y - 3z = -10 \gg -2y - 4z = -16$

SISTEMAS DE ECUACIONES

Ya estamos en el caso de dos ecuaciones

$-2y + z = -1$

$-2y - 4z = -16$

Multiplicamos la primera ecuación por – 1

$2y - z = 1$

$-2y - 4z = -16$

Sumamos las dos ecuaciones

$-5z = -15$ >>>> $z = 3$

Sustituimos en la primera ecuación

$-2y + 3 = -1$ >>> $y = 2$

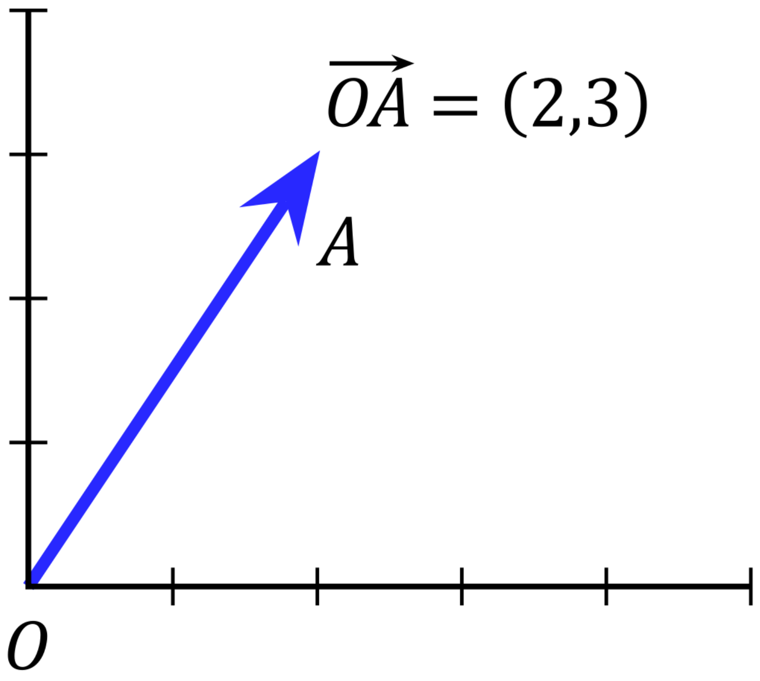
Sustituimos en una ecuacion del sistema de 3 ecuaciones $x + y + z = 6$

$x + 2 + 3 = 6$ >>> $x = 1$

VECTORES

En términos sencillos el vector es como una flecha que parte desde el punto de origen y se extiende en una dirección especifica hasta alcanzar un punto final

Puede ser representado como una línea recta con una longitud específica y una dirección en el espacio. Los vectores se utilizan para describir cantidades físicas como la velocidad, la aceleración, la fuerza y el desplazamiento, entre otros.



VECTORES

Los vectores tienen muchas aplicaciones en matemáticas, física, ingeniería, ciencias de la computación y otras áreas de la ciencia y la tecnología. Son una herramienta fundamental en el estudio del movimiento en el espacio, la programación gráfica en computación, y muchas otras aplicaciones en la ciencia y la ingeniería.

VECTORES Y LENGUAJES DE PROGRAMACIÓN

- Los vectores son una estructura matemática fundamental que se utiliza ampliamente en la programación para representar y manipular datos de una dimensión, es decir, secuencias de valores ordenados. Los lenguajes de programación proporcionan herramientas y operaciones específicas para trabajar con vectores, lo que permite a los programadores realizar operaciones eficientes y convenientes en datos de una dimensión.
- La relación entre vectores y lenguajes de programación se basa en el hecho de que los vectores son una forma natural de representar datos en muchos escenarios del mundo real. Por ejemplo, en procesamiento de señales, análisis de datos, manipulación de imágenes, simulaciones y muchas otras aplicaciones, los datos se pueden representar como vectores para facilitar su procesamiento y análisis.

Vectores en la vida diaria

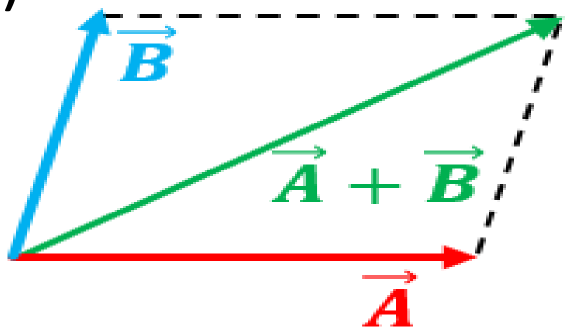
- Un ejemplo cotidiano del uso de vectores es la navegación con mapas en dispositivos de GPS (Sistema de Posicionamiento Global) o aplicaciones de navegación en teléfonos inteligentes.
- Cuando utilizas un dispositivo de GPS o una aplicación de navegación, ingresas una dirección o destino específico, y el dispositivo o la aplicación calcula una ruta para llegar allí. Esta ruta se representa como una serie de direcciones de giro, distancias y coordenadas geográficas. Estos datos se organizan en forma de vectores para representar la dirección y distancia de cada paso de la ruta.

Operaciones con vectores

Para sumar dos vectores $\vec{A}(7, 3)$ y $\vec{B}(2, -4)$.

$\vec{A} + \vec{B}$ = (7+2, 3-4) = (9, -1)

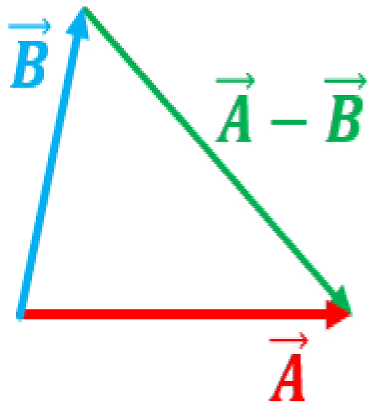
Graficamente



La resta es parecida pero restando

$\vec{A} - \vec{B}$ = (7-2, 3+4) = (5, 7)

Graficamente



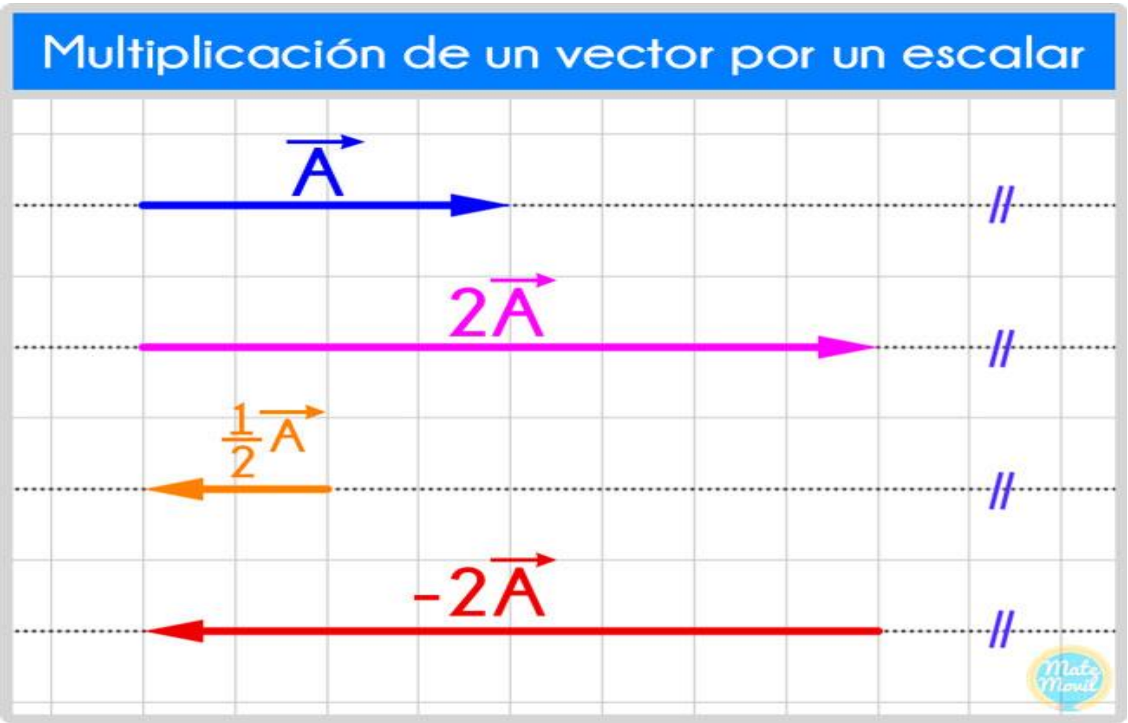
Multiplicación de vectores

- $\vec{A}(7, 3)$

$2\,\vec{A} = 2(7, 3) = (14, 6)$

$-\,2\,\vec{A} = -2(7, 3) = (-14, -6)$

Gráficamente



Matrices

Son un conjunto de números o expresiones ordenados en forma de filas y columnas

$$\underset{2x3}{A} = \begin{bmatrix} 7 & 18 & 3 \\ -6 & 0 & 12 \end{bmatrix} \quad \textbf{2 filas}$$

$$\textbf{3 columnas}$$

Matrices y lenguajes de programación

- La relación entre matrices y lenguajes de programación radica en el hecho de que las matrices son estructuras de datos bidimensionales que se utilizan comúnmente en la programación para representar y manipular datos en forma de tablas o arreglos de valores. Los lenguajes de programación proporcionan herramientas y operaciones para crear, acceder, modificar y procesar matrices, lo que permite a los programadores trabajar con datos en forma organizada y eficiente.
- Por ejemplo, en bases de datos relacionales, los datos se organizan en tablas que se pueden considerar como matrices, donde las filas representan registros y las columnas representan atributos o campos de datos.

Matrices y lenguajes de programación

- Algunos lenguajes de programación populares, como Python, R, MATLAB, Julia y muchos otros, tienen una amplia variedad de bibliotecas y paquetes especializados para trabajar con matrices y realizar operaciones matemáticas y numéricas de manera eficiente.

Matrices en la vida diaria

- Un ejemplo de la vida real donde se utiliza una matriz es en la gestión de inventarios en un almacén.
- Por ejemplo, supongamos que un almacén tiene 5 productos diferentes: A, B, C, D, y E. Se puede crear una matriz de inventario con 5 filas y varias columnas, donde cada fila representa un producto y las columnas representan atributos como el código de producto, la cantidad en stock, el precio unitario, y otros atributos relevantes.

Para sumar y restar matrices deben tener el mismo número de filas y columnas

$$A = \begin{pmatrix} 2 & 0 & 1 \\ 3 & 0 & 0 \\ 5 & 1 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

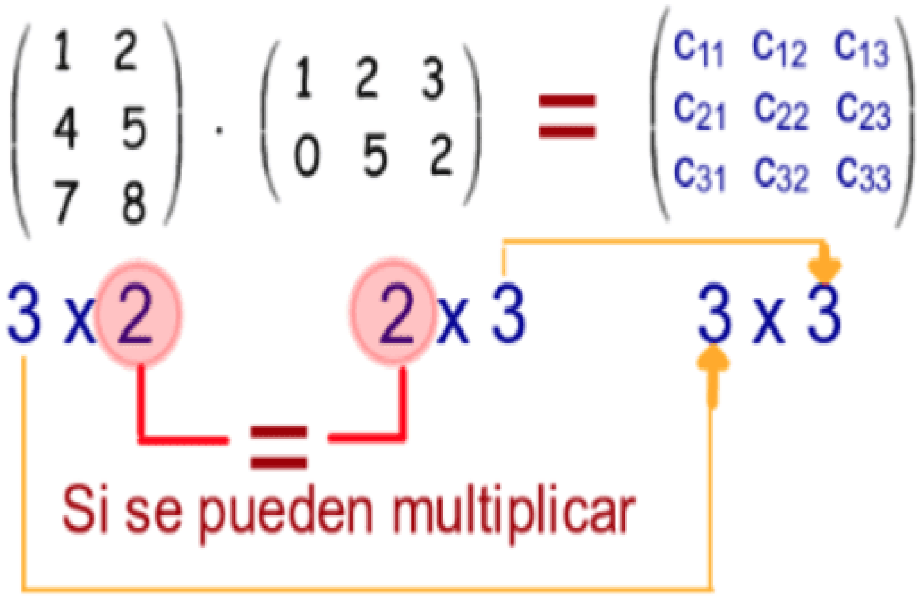
$$A + B = \begin{pmatrix} 2+1 & 0+0 & 1+1 \\ 3+1 & 0+2 & 0+1 \\ 5+1 & 1+1 & 1+0 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 2 \\ 4 & 2 & 1 \\ 6 & 2 & 1 \end{pmatrix}$$

$$A - B = \begin{pmatrix} 2-1 & 0-0 & 1-1 \\ 3-1 & 0-2 & 0-1 \\ 5-1 & 1-1 & 1-0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & -2 & -1 \\ 4 & 0 & 1 \end{pmatrix}$$

Multiplicar matrices

Dos matrices son multiplicables si el número de columnas de A coincide con el número de filas de B

C11 = cada numero de la fila 1 se multiplica por cada número de orden correspondiente de la columna 1
C12 = cada numero de la fila 1 se multiplica por cada número de orden correspondiente de la columna 2
C21 = cada numero de la fila 2 se multiplica por cada número de orden correspondiente de la columna 1
C22 = cada numero de la fila 2 se multiplica por cada número de orden correspondiente de la columna 2
Etc.



Ejemplo de multiplicación de matrices

$$A \cdot B = \begin{pmatrix} 2 & 0 & 1 \\ 3 & 0 & 0 \\ 5 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} 2\cdot 1+0\cdot 1+1\cdot 1 & 2\cdot 0+0\cdot 2+1\cdot 1 & 2\cdot 1+0\cdot 1+1\cdot 0 \\ 3\cdot 1+0\cdot 1+0\cdot 1 & 3\cdot 0+0\cdot 2+0\cdot 1 & 3\cdot 1+0\cdot 1+0\cdot 0 \\ 5\cdot 1+1\cdot 1+1\cdot 1 & 5\cdot 0+1\cdot 2+1\cdot 1 & 5\cdot 1+1\cdot 1+1\cdot 0 \end{pmatrix} =$$

$$= \begin{pmatrix} 3 & 1 & 2 \\ 3 & 0 & 3 \\ 7 & 3 & 6 \end{pmatrix}$$

Estadística básica

Media. Suma de todos los valores y dividido por el número de valores

Moda. El valor que más se repite en una serie de valores

Mediana. Una vez que ordenamos los números de una serie o lista buscamos el que está justo en el centro, si hay dos números en el centro los sumamos y dividimos entre dos.

Vamos a calcular la media, moda y mediana de la siguiente lista de edades de una clase(Puede haber más de una moda)

25, 37, 46, 47, 38, 42, 46, 32, 39, 48, 39, 37, 49, 51, 46, 32

Desviación típica

Desviación típica (σ) = $\sqrt{}(\Sigma((Xi - \mu)^2) / N)$

Donde:

- σ es la desviación típica.
- Xi son los valores individuales en el conjunto de datos.
- μ es la media del conjunto de datos.
- Σ es la suma.
- N es el número total de elementos en el conjunto de datos.
- En esta fórmula, se calcula la diferencia entre cada valor individual (Xi) y la media (μ), se eleva al cuadrado y luego se suman todos los resultados. El resultado se divide por el número total de elementos en el conjunto de datos (N) y luego se toma la raíz cuadrada para obtener la desviación típica. La desviación típica es una medida de dispersión que indica cuánto se alejan los valores individuales del valor promedio (la media) en un conjunto de datos.

Desviación típica

Desviación típica es una medida que se utiliza para cuantificar la variación o la dispersión de una serie de valores.

Si es un numero bajo quiere decir que la mayoría de los valores se agrupan cerca de la media de esos valores

Calcula la desviación típica de la siguiente lista de números

6, 8, 1, 4, 2, 3, 5, 2, 4, 5

1.- Encontrar la media

Media= suma de todos los datos/número de datos

40/10 = 4

2.-Calcular la suma de las diferencias al cuadrado entre cada número y la media

$(6-4)^2 + (8-4)^2 + (1-4)^2$ etc.

4 + 16 + 9 + 0 + 4 + 1 + 1 etc.

3.-Dividir la suma de cuadrados por la cantidad de datos

Varianza = suma de cuadrados/N= 40/10 = 4

4.- Encontrar la raíz cuadrada de la varianza para obtener la desviación típica

Desviación típica = $\sqrt{varianza}$

Desviación típica = 2

PROBABILIDADES

- La teoría de probabilidades es una rama de las matemáticas que estudia los eventos aleatorios, cuantificando la incertidumbre asociada a ellos mediante números entre 0 y 1. Se utiliza en diversos campos como estadística, economía, ciencias de la computación, medicina, ingeniería, ciencias sociales, ciencia de datos, inteligencia artificial, etc.

PROBABILIDADES

Probabilidades y el modelo GPT de CHATGPT

El modelo GPT (Generative Pre-trained Transformer), como ChatGPT, utiliza probabilidades en varias maneras fundamentales para generar texto. A continuación, algunas de las formas en que se usa el concepto de probabilidad:

- ❖Entrenamiento del Modelo: Durante la fase de entrenamiento, el modelo aprende a asignar probabilidades a secuencias de palabras o tokens, de modo que las secuencias que son más probables de aparecer en el conjunto de datos de entrenamiento tengan una mayor probabilidad.

- ❖Predicción de la Palabra Siguiente: GPT está diseñado para predecir la siguiente palabra en una secuencia de palabras dadas. Utiliza las probabilidades que ha aprendido durante el entrenamiento para elegir la palabra que es más probable que siga a una secuencia de palabras dada.

PROBABILIDADES

- ❖Decodificación: En la fase de generación de texto, se utilizan varios algoritmos de decodificación como "greedy", "beam search" o "núcleo de muestreo" para elegir la próxima palabra, basándose en las probabilidades que el modelo ha asignado a las posibles palabras siguientes.

- ❖Manejo de la Incertidumbre: Las probabilidades también ayudan al modelo a manejar la incertidumbre en el lenguaje. Por ejemplo, si se le da una oración incompleta o ambigua, el modelo utiliza las probabilidades para elegir la continuación que cree que es más probable.

- ❖Calidad de Generación de Texto: Al tener en cuenta las probabilidades, el modelo puede generar texto que es gramaticalmente correcto y contextualmente relevante, ya que "aprende" las estructuras lingüísticas y las dependencias contextuales que son más comunes en el idioma.

En resumen, las probabilidades son fundamentales tanto para el entrenamiento como para la generación de texto en modelos como ChatGPT. Permiten que el modelo capture la estructura del lenguaje y genere respuestas coherentes y contextuales.

PROBABILIDADES

Si tiramos un dado con 6 caras y numerados desde el 1 hasta el 6, la probabilidad de que salga cualquiera de los números es de 1/6

- La fórmula básica para calcular la probabilidad de un evento *A* es:

P(A)=Número de casos favorables/Número total de casos posibles

P(tirar dado y que salga un 4) =1/6

¿Cuál es la probabilidad de que tanto en la primera tirada como en la segunda salga un 3?

P(A)= 1/6 x 1/6 = 1/36

Y cuál es la probabilidad de que salga un 3 en la primera tirada, y también en la segunda tirada y también en la tercera tirada

P(A)= 1/6 x 1/6 x 1/6= 1/216

PROBABILIDADES

- La probabilidad de que un evento A NO ocurra es $1 - P(A)$.

¿Cuál es la probabilidad de que no salga un 3 en una tirada del dado?

PROBABILIDADES

- DIAGRAMA DE VENN

En un club deportivo, el 80% de los socios juegan al fútbol y el 40% al baloncesto. Sabiendo que el 30% de los socios practican los dos deportes, calcula la probabilidad de que un socio elegido al azar:

a) Juegue sólo al fútbol

b) Juegue sólo al baloncesto

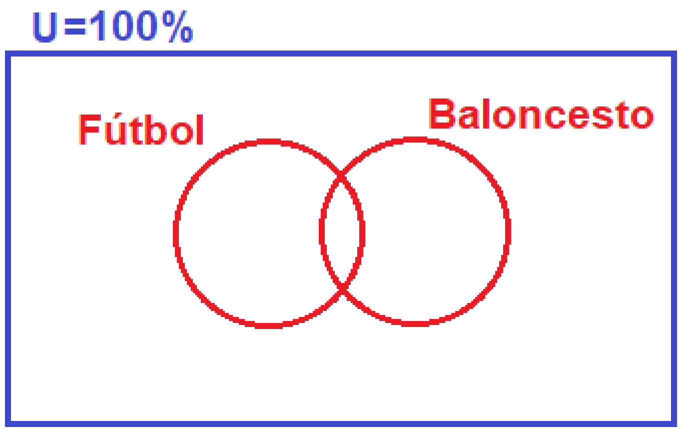
c) Juegue al fútbol o al baloncesto

d) No juegue a ninguno de los dos deportes

Empezamos dibujando el diagrama de Venn, con los dos conjuntos, fútbol y baloncesto, e indicando que el conjunto universal es igual a 100%:

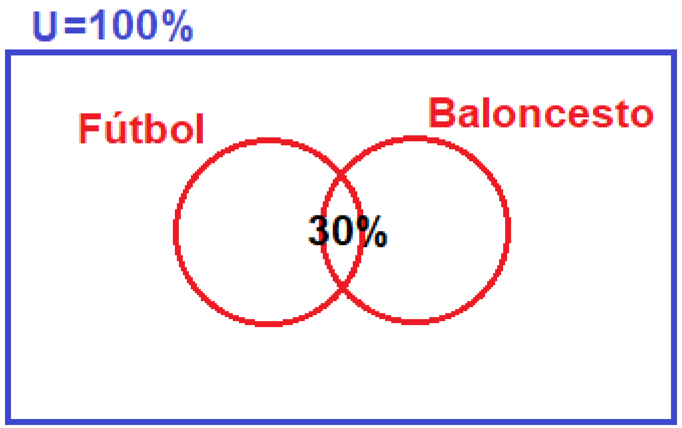
PROBABILIDADES

- Empezamos dibujando el diagrama de Venn, con los dos conjuntos, fútbol y baloncesto, e indicando que el conjunto universal es igual a 100%:



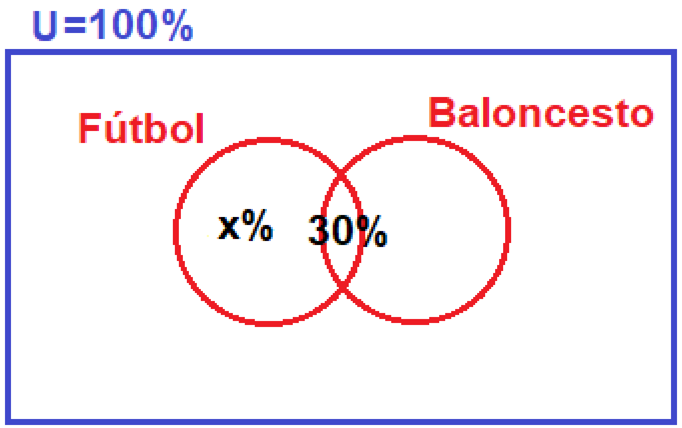
PROBABILIDADES

- El primer dato que ubicamos es el 30% que practican ambos deportes, que corresponde a la zona 2, o lo que es lo mismo, la intersección entre fútbol y baloncesto.



PROBABILIDADES

- Vamos a calcular el porcentaje de las personas que juegan sólo al fútbol, que lo indicamos como x%:



PROBABILIDADES

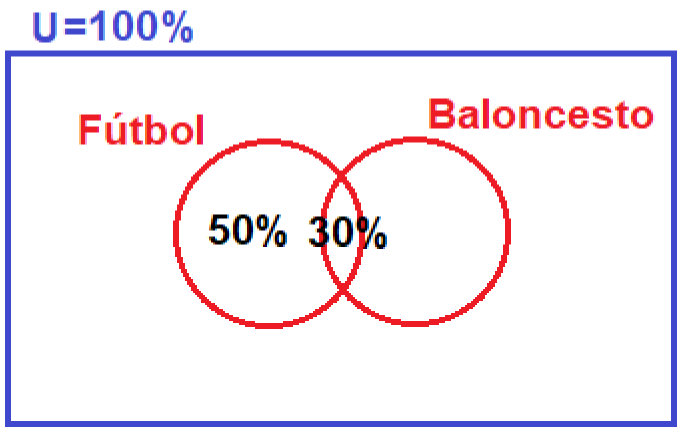
Sabemos que el 80% juega al fútbol y que de ese porcentaje, el 30% juega a los dos deportes y que x% juega sólo al fútbol, luego el 80% es la suma de x% más 30%:

80% = X% + 30%

Despejamos x% y calculamos:

X% = 80% - 30% = 50%

Tenemos que un 50% de los socios juega sólo al fútbol. Colocamos este resultado en el diagrama de Venn:



PROBABILIDADES

Ahora vamos a calcular el porcentaje de las personas que juegan sólo al fútbol, que lo llamamos y%:

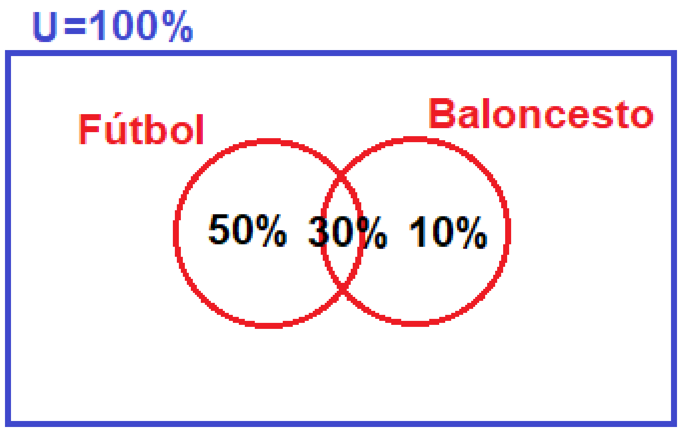
Sabemos que el 40% juega al baloncesto y qué de ese porcentaje, el 30% juega a los dos deportes y qué y% juega sólo al baloncesto, luego el 40% es la suma de y% más 30%:

40%=y%+30%

Despejamos y% y calculamos:

Y%=40%-30%=10%

Por lo que un 10% de los socios juega sólo al baloncesto. Colocamos este resultado en el diagrama de Venn:



PROBABILIDADES

- El porcentaje de socios que juegue al fútbol o al baloncesto es igual al porcentaje de los socios que juegan sólo al fútbol más el porcentaje de los socios que juegan sólo al baloncesto:
- 50% + 10% = 60%
- Es decir, que un 60% de los socios juegan al fútbol o al baloncesto.
- Para calcular el porcentaje de los socios que no juegue a ninguno de los dos deportes, al total de socios (100%) le restamos el 50% que juega sólo al fútbol, el 10% que juega sólo al baloncesto y el 30% que juega a los dos deportes:
- 100% - 50% - 10% - 30% = 10%
- Lo cual resulta que un 10% de los socios no juega a ningún deporte.

FRECUENCIAS

En estadística se pueden distinguir hasta cuatro tipos de frecuencias:

- Frecuencia absoluta

Frecuencia absoluta es el número de veces que se repite un resultado en el conjunto de todos los observados.

- Frecuencia relativa

La Frecuencia relativa es la proporción de cada frecuencia absoluta, es decir, el número de veces que se produce ese resultado (frecuencia absoluta) dividido por el número total de datos observados. (fi), es el cociente entre la frecuencia absoluta y el tamaño de la muestra (N). Es decir,

Se presenta en una tabla o nube de puntos en una distribución de frecuencias. Si multiplicamos la frecuencia relativa por 100 obtendremos el porcentaje o tanto por ciento (pi).

- Frecuencia absoluta acumulada

(Ni), se refiere al total de las frecuencias absolutas para todos los eventos iguales o menores de un cierto valor, en una lista ordenada de eventos.

- Frecuencia relativa acumulada

(Hi), es el cociente entre la frecuencia absoluta acumulada y el total de la muestra.

- Ejemplos de frecuencias

Supongamos que las calificaciones de un estudiante de secundaria fueran las siguientes:

18, 13, 12, 14, 11, 08, 12, 15, 05, 20, 18, 14, 15, 11, 10, 10, 11, 11. Entonces:

La frecuencia absoluta del número 11 es 4, pues 11 aparece 4 veces.

La frecuencia relativa del número 11 es 0.22, porque corresponde a la división 4/18 (4 de las veces que aparece de las 18 notas que aparecen en total).

La frecuencia absoluta acumulada para el valor 11 es 8, porque hay 8 valores menores o iguales a 11.

La frecuencia relativa acumulada para el valor 11 es 0.44, porque corresponde a la división 8/18 (frecuencia absoluta acumulada dividida entre el número total de muestras).

FRECUENCIAS

Usando como ejemplo la tabla de abajo, crea una tabla de frecuencias con los datos de la diapositiva anterior.

TABLAS DE FRECUENCIA

EJEMPLO 1

x_i	f_i	f_r	%	F
0	6	0,24	24	6
1	7	0,28	28	13
2	6	0,24	24	19
3	4	0,16	16	23
4	1	0,04	4	24
5	1	0,04	4	25
	25	1		