AWAKELAB

BASECAMP

Ciencia de Datos

Inferencia Estadística

Objetivo de la jornada

- Explicar los principales conceptos de probabilidad asociados a un evento aleatorio.

Probabilidades y sus definiciones

Probabilidad

La probabilidad se refiere a la mayor o menor posibilidad de que ocurra un suceso. Su noción viene de la necesidad de medir la certeza o duda de que un suceso dado ocurra o no. Esta establece una relación entre el número de sucesos favorables y el número total de sucesos posibles.

Es la posibilidad de que un evento suceda dependiendo de las condiciones dadas para que acontezca (ejemplo: qué probabilidad hay de que llueva). Será medida entre 0 y 1 o expresada en porcentajes, dichos rangos podrán observarse en ejercicios resueltos de probabilidad. Para ello se medirá la relación entre los sucesos favorables y los posibles.

Los sucesos favorables son los válidos según la experiencia del individuo; y los posibles son los que pueden darse si son válidos o no a su experiencia. La probabilidad y estadística están relacionadas al ser el área donde se registran sucesos. La etimología del término viene del latín probabilitas o posibilitatis, relacionadas a “probar” o “comprobar” y tat que se refiere a “cualidad”. El término se relaciona a la cualidad de probar.

Aplicaciones de probabilidad

Pólizas de seguros

Para asignar el valor de una póliza de seguros, ya sea de seguros o de automóviles, se divide a la población en categorías por edades y antecedentes, ya que cada categoría tiene diferente probabilidad de sufrir determinados percances. Por ejemplo, hay grupos con más riesgo de sufrir accidentes por choques de autos, o de tener un ataque cardíaco.

Control de calidad

Interesa saber qué probabilidades hay de encontrar piezas defectuosas en un lote, con la finalidad de tomar medidas para mejorar la calidad del producto y ofrecer la garantía adecuada.

Mercado petrolero

La probabilidad de conflictos que involucran a países productores de determinadas materias primas estratégicas como el petróleo, inciden notablemente en los precios de este y sus derivados, como la gasolina.

Experimento aleatorio y determinista

Experimento Aleatorio: Experimento cuyo resultado no se puede predecir, habiendo un conjunto de resultados posibles.

Experimento determinista: Experimento en que sabemos de antemano lo que va a ocurrir, ejemplos de ello son: El tiempo que demora una piedra en caer desde una misma altura, sacar una galleta de un paquete de criollitas, etc.

Espacio Muestral

El espacio muestral es el conjunto de todos los posibles resultados de un experimento aleatorio y se suele representar como E (o bien como omega, Ω, del alfabeto griego).

Por ejemplo, cuando lanzamos una moneda, ¿cuáles son todos los posibles resultados que podemos obtener? Que salga cara o cruz, ¿verdad? En total son dos posibles resultados, por lo que el espacio muestral tiene 2 elementos.

$$E = \{cara, cruz\}$$

Y si lanzamos un dado, tenemos en total 6 posibles resultados que pueden salir. Por lo tanto el espacio muestral sería de 6 elementos.

$$E = \{1, 2, 3, 4, 5, 6\}.$$

Evento Aleatorio

Es cualquier resultado que ocurre al realizar un experimento aleatorio, se representa mediante un subconjunto del espacio muestral. Los sucesos (o eventos) se mencionan con las primeras letras del alfabeto, escritas con mayúsculas. Por ejemplo, “sacar cara” en el lanzamiento de una moneda, “sacar el número 5” o “sacar un número primo” en el lanzamiento de un dado son sucesos.

Ejemplo Probabilidad

En una bolsa hay 10 bolas numeradas del 11 al 20, idénticas, salvo en el color, pues unas son rojas y las otras verdes.

a) Sacamos, sin mirar, una bola . ¿Cúal es la probabilidad de obtener un número primo?

b) Se sabe que la probabilidad de sacar una bola verde es de 3/5. ¿Cuántas bolas hay de cada color?

Solución:

Veamos cuál sería el espacio muestral en el primer apartado de nuestro ejercicio.

¿Cuáles son todos los posibles resultados? Nos referimos a los números de las bolas, que son los números del 11 al 20.

Nuestro espacio muestral tiene 10 elementos:

$$E = \{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$$

Y el suceso por el que nos preguntan es “obtener un número primo”.

Ahora, ¿cómo calculamos la probabilidad de este suceso?

Cuando todos los sucesos elementales tienen las misma probabilidad de ocurrir, la probabilidad de un suceso cualquiera A se define como el cociente entre el número de casos favorables y el número de casos posibles. Esta es la Ley de Laplace.

$$P(A) = \frac{\text{Nº de casos favorables}}{\text{Nº de casos posibles}}$$

En el ejemplo de lanzar una moneda, los sucesos elementales serían: “Sacar una cruz” o “Sacar una cara”. Si la moneda no está trucada, la probabilidad de que ocurra cada suceso elemental es la misma. Por lo tanto, la probabilidad de que salga cruz es 1/2.

Volviendo a nuestro ejercicio: En una bolsa hay 10 bolas numeradas del 11 al 20, algunas rojas y otras verdes.

Sacamos sin mirar una bola, ¿cuál es la probabilidad de sacar un número primo?

Empezamos calculando el número de casos favorables y el número de casos posibles.

Número de casos favorables = número de primos = 4 son los números primos dentro de los resultados posibles (Los números 11, 13, 17 y 19 son primos)

Número de casos posibles = 10 (Todos los números del 11 al 20)

La probabilidad de sacar un número primo entre las 10 bolas, es de 4/10 que simplificado es 2/5.

Solución:

P (número primo)=2/5

¿Cuántas bolas hay de cada color?

Nos dice que la probabilidad de que salga verde es 3/5.

El número de casos posibles, es decir, el número de bolas que pueden salir, sigue siendo 10.

El número de casos favorables, es decir, el número de bolas de color verde (nuestro suceso) es una de las cosas que queremos calcular.

Sabemos que 3/5 es equivalente a 6/10. Por lo tanto, si aplicamos la Ley de Laplace:

$$P(sacar\ una\ bola\ verde) = \frac{Nº\ de\ casos\ favorables}{Nº\ de\ casos\ posibles} = \frac{6}{10}$$

En total hay 6 bolas verdes en la bolsa. Así que podemos deducir que el resto, 4, son bolas rojas.

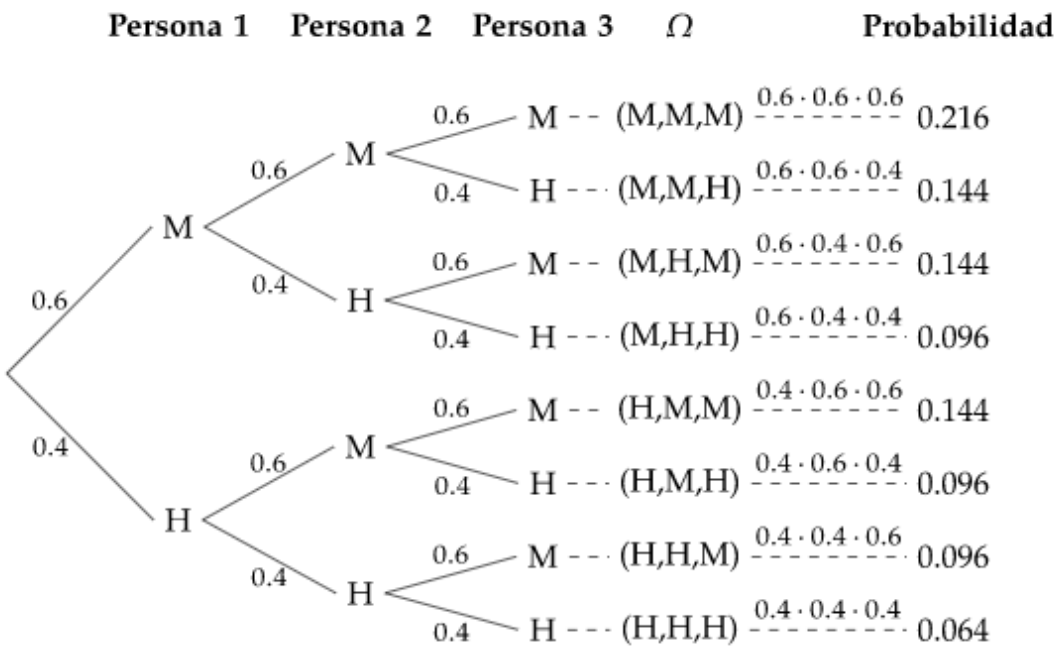
Solución: Hay 6 bolas verdes y 4 bolas rojas

Árbol de probabilidades.

En experimentos donde se mide más de una variable, la determinación del espacio muestral puede resultar compleja. En tales casos es recomendable utilizar un árbol para construir el espacio muestral.

En un diagrama de árbol cada variable se representa en un nivel del árbol y cada posible valor de la variable como una rama.

Ejemplo: El siguiente diagrama de árbol representa el espacio muestral de un experimento aleatorio en el que se mide el sexo y el grupo sanguineo de un individuo al azar.

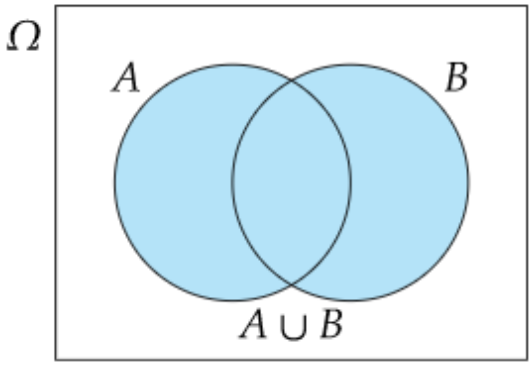


Unión e intersección de eventos

Unión de sucesos

Dados dos sucesos $A,B \subseteq \Omega$, se llama *suceso unión* de A y B, y se denota $A \cup B$, al suceso formado por los elementos de A junto a los elementos de B, es decir,

$$A \cup B = \{x \mid x \in A \text{ o } x \in B\}.$$

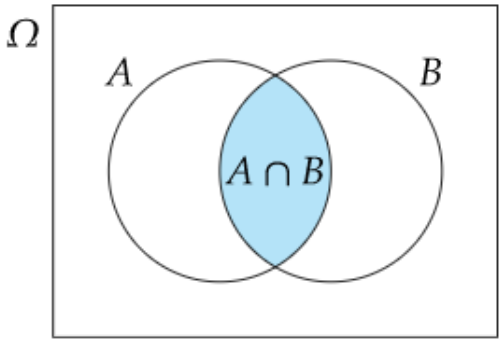


El suceso unión $A \cup B$ ocurre siempre que ocurre A o B.

Intersección de sucesos.

Dados dos sucesos $A,B \subseteq \Omega$, se llama *suceso intersección* de A y B, y se denota $A \cap B$, al suceso formado por los elementos comunes de A y B, es decir,

$$A \cap B = \{x \mid x \in A \text{ y } x \in B\}.$$



El suceso intersección $A \cap B$ ocurre siempre que ocurren A y B.

Diremos que dos sucesos son incompatibles si su intersección es vacía.

Probabilidad Condicional.

Experimentos condicionados

En algunas ocasiones, es posible que tengamos alguna información sobre el experimento antes de su realización. Habitualmente esa información se da en forma de un suceso B del mismo espacio muestral que sabemos que es cierto antes de realizar el experimento.

En tal caso se dice que el suceso es un suceso condicionante, y la probabilidad de otro suceso se conoce como y se expresa.

$$P(A|B).$$

Esto debe leerse como probabilidad de A dado B o probabilidad de A bajo la condición de B.

Los condicionantes suelen cambiar el espacio muestral del experimento y por tanto las probabilidades de sus sucesos.

Ejemplo. Supongamos que tenemos una muestra de 100 hombres y 100 mujeres con las siguientes frecuencias.

	No fumadores	Fumadores
Mujeres	80	20
Hombres	60	40

Entonces, usando la definición frecuentista de probabilidad, la probabilidad de que una persona elegida al azar sea fumadora es

$$P(\text{Fumadora}|\text{Mujer}) = \frac{20}{100} = 0.2.$$

Probabilidad condicionada

Dado un espacio muestral Ω de un experimento aleatorio, y dos sucesos A,B⊆Ω, la probabilidad de A *condicionada* por B es

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

siempre y cuando, P(B)≠0.

Esta definición permite calcular probabilidades sin tener que alterar el espacio muestral original del experimento.

Ejemplo. En el ejemplo anterior

$$P(\text{Fumadora}|\text{Mujer}) = \frac{P(\text{Fumadora} \cap \text{Mujer})}{P(\text{Mujer})} = \frac{20/200}{100/200} = \frac{20}{100} = 0.2.$$

Probabilidad conjunta y marginal.

Conjunta

La probabilidad conjunta es simplemente la probabilidad de que sucedan dos eventos al mismo tiempo. Es la probabilidad de que evento X se produce al mismo tiempo como evento Y . Suena fácil, ¿verdad? Bueno, hay un par de condiciones. Una es que los eventos X e Y deben ocurrir al mismo tiempo. Lanzar dos dados sería un ejemplo de eso. La otra es que los eventos X e Y deben ser independientes entre sí. Eso significa que el resultado del evento X no influye en el resultado del evento Y. Nuestra tirada de dados es de nuevo un buen ejemplo de eventos independientes, ya que el resultado de tirar un dado no influye en el resultado de tirar el otro. Si su primer dado sale un 1, la probabilidad del segundo dado sigue siendo una probabilidad de 1/6 para cada número entre uno y seis.

Entonces, ¿cuál es un ejemplo de dos eventos que no son independientes? Bueno, ¿qué pasa con el evento X es la probabilidad de que haya nubes en el cielo y el evento Y es la probabilidad de que llueva? Incluso Wally the Wacky Weatherman (¡que se equivoca mucho!) Sabe que la lluvia proviene de las nubes. Entonces, la lluvia sólo puede caer cuando hay nubes en el cielo. Eso significa que la presencia de nubes influirá en las posibilidades de lluvia, ¡y eso significa que estos dos eventos no son independientes!

Andrés está jugando un juego de mesa. Es su turno y quiere sacar exactamente un doce para alcanzar su objetivo. La única forma de conseguir esos doce es lanzar un seis en cada dado. Como ya sabemos que lanzar dos dados son eventos independientes, podemos usar la fórmula de probabilidad conjunta para calcular sus posibilidades de éxito. Aquí está la fórmula:

$$P(X,Y) = P(X) * P(Y)$$

Si la probabilidad de sacar un seis en un dado es P (X) y la probabilidad de sacar un seis en el segundo dado P (Y), podemos usar la fórmula P (X,

Y) = P (X) * P (Y). Dado que los dados tienen seis lados y la probabilidad de que salga cualquier lado es igual, P (X) y P (Y) son iguales a 1/6. Por lo tanto, la fórmula se parece a la que aparece en su pantalla en este momento y, finalmente, da como resultado una probabilidad del 2.8%.

$$P(X,Y) = P(X) * P(Y) = 1/6 * 1/6 = 1/36 = 0.277 = 2.8\%$$

Marginal o incondicional

La probabilidad incondicional también se llama probabilidad marginal y mide la probabilidad de ocurrencia sin ignorar la información obtenida de eventos anteriores o externos. Debido a que esta probabilidad ignora nueva información, permanece constante.

La probabilidad incondicional de un evento se puede determinar sumando los resultados del evento y dividiendo por el número total de resultados posibles.

Ejemplo:

Como ejemplo hipotético de las finanzas, examinemos un grupo de acciones y sus rendimientos. Una acción puede ser un ganador, que obtiene un resultado positivo, o un perdedor, que tiene un resultado negativo. Suponga que las acciones A y B son ganadoras de cinco acciones, mientras que las acciones C, D y E se pierden. ¿Cuál es, entonces, la probabilidad incondicional de que elija una acción ganadora? Dado que un ganador dará dos de los cinco resultados posibles, la probabilidad incondicional es de 2 éxitos divididos por 5 resultados totales (2/5 = 0,4), o el 40%.

Independencia entre eventos aleatorios.

Experimentos aleatorios

El estudio de una característica en una población se realiza a través de experimentos aleatorios.

Un experimento aleatorio es un experimento que cumple dos condiciones:

1. El conjunto de posibles resultados es conocido.
2. No se puede predecir con absoluta certeza el resultado del experimento.

Ejemplo. Un ejemplo típico de experimentos aleatorios son los juegos de azar. El lanzamiento de un dado, por ejemplo, es un experimento aleatorio ya que:

- Se conoce el conjunto posibles de resultados {1,2,3,4,5,6}.
- Antes de lanzar el dado, es imposible predecir con absoluta certeza el valor que saldrá.

Otro ejemplo de experimento aleatorio sería la selección de un individuo de una población al azar y la determinación de su grupo sanguíneo.

En general, la obtención de cualquier muestra mediante procedimientos aleatorios será un experimento aleatorio.

Árboles de decisión.

El árbol de decisiones es uno de esos ejemplos de herramientas que facilitan la toma de decisiones. En base a un diagrama de flujo se visualiza el proceso de toma de decisiones mediante el mapeo de diferentes cursos de acción, así como sus posibles resultados.

Por muy diferentes que sean los propósitos que motivan la creación del árbol de decisiones, ejemplos hay muchos, tanto en el ámbito profesional

como en el personal; este tipo de diagramas se componen de tres elementos diferentes:

1. **Nodo Raíz.** Este nodo de nivel superior representa el objetivo final o la gran decisión que se está intentando tomar.

2. **Las ramas**, que provienen de la raíz, representan diferentes opciones o cursos de acción que están disponibles al tomar una decisión en particular. Generalmente se indican con una línea de flecha y a menudo incluyen los costes asociados, así como la probabilidad de que ocurra.

3. **Nodo de hoja**. Los nodos foliares, que están unidos al final de las ramas, representan posibles resultados para cada acción. Por lo general, hay dos tipos: los nodos de hoja cuadrada, que indican otra decisión a tomar, y nodos de hoja circular, que representan un evento fortuito o un resultado desconocido.

Ahora que sabe exactamente qué es un árbol de decisiones, los ejemplos de sus principales beneficios te ayudarán a encontrar la utilidad a esta herramienta. Un diagrama de este tipo tiene importantes ventajas como:

- Los árboles de decisión no son lineales, lo que significa que hay mucha más flexibilidad para explorar, planificar y predecir varios resultados posibles para cada curso de acción.

- Comunicación efectiva. Como demuestran visualmente las relaciones de causa y efecto, los árboles de decisiones proporcionan una vista simplificada, incluso de procesos potencialmente complicados.

- Objetividad. Los árboles de decisión se centran en la probabilidad y los datos, no en las emociones

Como crear un árbol de decisiones

a) Comenzar estableciendo el objetivo general en la parte superior (raíz). Representa la decisión que se está intentando tomar.

b) Dibujar las flechas para cada curso de acción posible. Estas flechas salen de la raíz y deben hacer referencia a los costes asociados con cada acción, así como la probabilidad de éxito.

c) Incluir nodos de hoja al final de las ramas. ¿Cuáles son los resultados de cada curso de acción? Si se debe tomar otra decisión, se dibuja un nodo de hoja cuadrada. Si el resultado es incierto, se dibuja un nodo de hoja circular.

d) Determinar las probabilidades de éxito de cada punto de decisión. Al crear un árbol de decisión, es importante investigar, para poder predecir con precisión la probabilidad de éxito.

e) Evaluar riesgo vs recompensa. Calcular el valor esperado de cada decisión en el árbol ayuda a minimizar el riesgo y aumentar la probabilidad de alcanzar un resultado favorable.

Ejemplo:

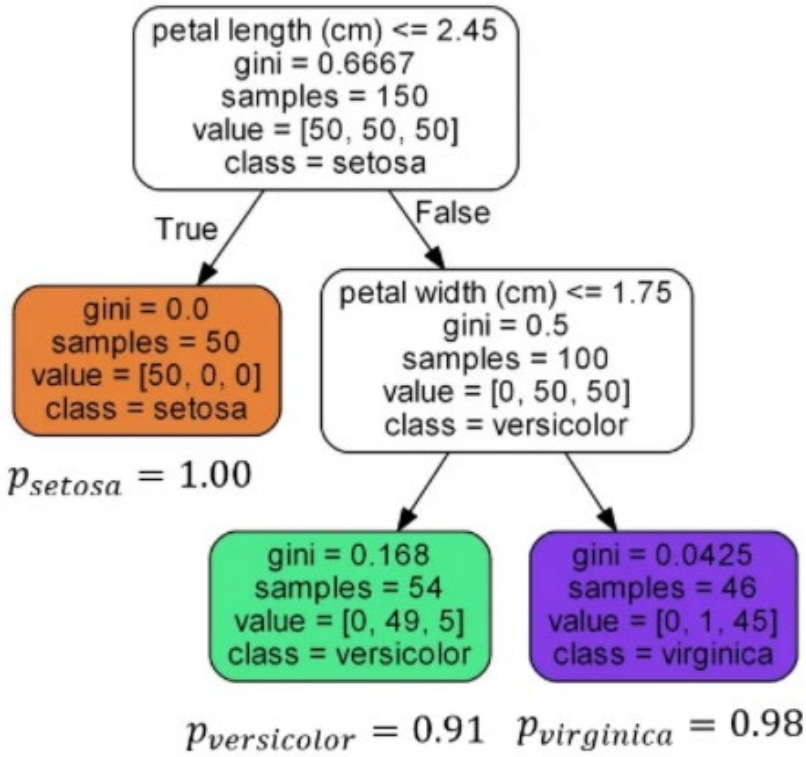
Si le damos las 150 flores del conjunto de datos Iris a un árbol de decisión para que lo clasifique, nos quedaría un árbol como el que se muestra a continuación. Vamos a aprender a leerlo:

- cada color representa a una clase. El marrón para setosa, el verde para versicolor y el lila para virginica.
- el color es más intenso cuanto más seguros estamos que la clasificación es correcta
- los nodos blancos, por tanto, evidencia la falta de certeza
- Hay dos tipos de nodo:

- *Nodos de decisión*: tienen una condición al principio y tienen más nodos debajo de ellos
- *Nodos de predicción*: no tienen ninguna condición ni nodos debajo de ellos. También se denominan «nodos hijo»

La información de cada nodo es la siguiente:

- condición: si es un nodo donde se toma alguna decisión
- ini: es una medida de impureza. A continuación veremos cómo se calcula
- samples: número de muestras que satisfacen las condiciones necesarias para llegar a este nodo
- value: cuántas muestras de cada clase llegan a este nodo
- class: qué clase se le asigna a las muestras que llegan a este nodo



La interpretación del árbol de este árbol de decisión sería: si la longitud del pétalo es menos de 2.45 centímetros, entonces la flor iris pertenece a la variedad setosa. Si por el contrario, la longitud del pétalo es mayor que 2.45 centímetros, habría que mirar al ancho del pétalo. Cuando el ancho del pétalo es menor o igual a 1.75 centímetros, pertenece a la variedad versicolor con un 91% de probabilidad. Si no, parece que sería virginica con un 98% de probabilidad.

Ley de Bayes.

Los sucesos de un sistema completo de sucesos A1,⋯,An también pueden verse como las distintas hipótesis ante un determinado hecho B.

En estas condiciones resulta útil poder calcular las probabilidades a posteriori P(Ai|B) de cada una de las hipótesis.

Definición: Dado un sistema completo de sucesos A1,…,An y un suceso B de un espacio muestral Ω y otro suceso B del mismo espacio muestral, la probabilidad de cada suceso Ai i=1,…,n condicionada por B puede calcularse con la siguiente fórmula:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}.$$

Ejemplo. En el ejemplo anterior, una pregunta más interesante es qué diagnosticar a una persona que presenta el síntoma.

En este caso se puede interpretar E y E— como las dos posibles hipótesis para el síntoma S. Las probabilidades a priori para ellas son P(E)=0.2 y P(E—)=0.8. Esto quiere decir que si no se dispone de información sobre el síntoma, el diagnóstico será que la persona no tiene la enfermedad.

Sin embargo, si al reconocer a la persona se observa que presenta el síntoma, dicha información condiciona a las hipótesis, y para decidir entre ellas es necesario calcular sus probabilidades a posteriori, es decir, P(E|S) y P(\overline{E}|S)$.

Para calcular las probabilidades a posteriori se puede utilizar el teorema de Bayes:

$$P(E|S) = \frac{P(E)P(S|E)}{P(E)P(S|E) + P(\overline{E})P(S|\overline{E})} = \frac{0.2 \cdot 0.9}{0.2 \cdot 0.9 + 0.8 \cdot 0.4} = \frac{0.18}{0.5} = 0.3$$

$$P(\overline{E}|S) = \frac{P(\overline{E})P(S|\overline{E})}{P(E)P(S|E) + P(\overline{E})P(S|\overline{E})} = \frac{0.8 \cdot 0.4}{0.2 \cdot 0.9 + 0.8 \cdot 0.4} = \frac{0.32}{0.5} = 0.6$$

Como se puede ver la probabilidad de tener la enfermedad ha aumentado. No obstante, la probabilidad de no tener la enfermedad sigue siendo mayor que la de tenerla, y por esta razón el diagnóstico seguirá siendo que no tiene la enfermedad.

En este caso se dice que el síntoma S no es determinante a la hora de diagnosticar la enfermedad.

Referencias

[1] Probabilidad y estadística

https://www.smartick.es/blog/matematicas/probabilidad-y-estadistica/probabilidad-que-es/

[2] Experimento Aleatorio y Determinísticos

https://www.youtube.com/watch?v=ttf8QxwaXxw&t=160s

[3] Probabilidad Conjunta, marginal y condicional

https://programmerclick.com/article/51501800829/

[4] Árboles de decisiones

https://blog.hubspot.es/sales/arbol-decisiones

[5] teorema de Bayes

https://www.profesor10demates.com/2013/09/probabilidad-7-teorema-de-bayes.html


AWAKELAB