# Unidad 3: Información y su codificación

Información y su codificación. Noción de Información. Particiones y árboles. Codificación. Longitud media de código. Concepto de Entropía. Información condicional. Clasificación de códigos. Condición de prefijo y códigos instantáneos. Inecuación de Kraft. Construcción de códigos compactos. Códigos de Shannon, Fano y Huffman. Rendimiento y redundancia de un código. Extensión de fuentes. Propiedad extensiva de la Entropía. Primer Teorema de Shannon.

*Particiones y árboles. Codificación. Longitud media de código. Concepto de Entropía. Condición de prefijo y códigos instantáneos. Inecuación de Kraft.  Código de Huffman. Rendimiento y redundancia de un código.*

## Codificación

La **codificación** se refiere al proceso de transformar información de un formato a otro, con el fin de poder almacenarla, transmitirla o procesarla de manera más eficiente. En el contexto de la teoría de la información, la codificación se utiliza para comprimir la información para su almacenamiento y transmisión.

Un **código unívoco** es un tipo de código que garantiza que cada elemento en un conjunto dado de elementos tenga una representación única en el código. En otras palabras, cada elemento tiene asignado un código único que lo identifica de manera unívoca.

Por ejemplo, en un diccionario, cada palabra tiene una entrada única y específica, y cada entrada se identifica mediante una palabra clave única que la distingue de las demás entradas en el diccionario. De manera similar, en una base de datos, cada registro tiene un identificador único, como un número o una combinación de letras y números, que lo distingue de los demás registros en la base de datos.

## Condición de prefijo

La condición de prefijo es una propiedad importante de algunos tipos de códigos, especialmente de los códigos de longitud variable, que garantiza que ningún código en el conjunto sea un prefijo de otro código en el mismo conjunto. Es decir, ninguna palabra en el código puede ser una concatenación de otras palabras en el mismo código.

Esta propiedad es importante porque garantiza que los códigos de longitud variable se puedan decodificar de manera unívoca y sin ambigüedad. Si un código no cumple con la condición de prefijo, entonces es posible que haya varias maneras diferentes de decodificar una secuencia de códigos, lo que puede llevar a errores y a una pérdida de información.

## Longitud media de código

La longitud media de código es una medida utilizada en la teoría de la información para evaluar la eficiencia de un código de longitud variable. Esta medida se calcula multiplicando la longitud de cada código por la probabilidad de que aparezca ese código, sumando los productos y dividiendo por la probabilidad total de todos los códigos.

* Pi: Probabilidad del símbolo i
* Li: Longitud del símbolo i
* Ejemplo
  + A → 1
  + D → 01
  + C→ 001
  + B → 001
    - Condición prefijo: ACB→1001000
    - Longitud media de código: L=

## Ecuación de Kraft

La igualdad de Kraft es un resultado importante en la teoría de la información que establece una condición necesaria y suficiente para que un conjunto de códigos de longitud variable sea válido.

En particular, la igualdad de Kraft establece que un conjunto de códigos de longitud variable es válido si y solo si la suma de las longitudes de código para todos los símbolos es menor o igual a 1, donde la longitud de un código se mide en términos de la cantidad de bits necesarios para representar el código.

Longitud de código mínima

La longitud de código mínima es el menor número de bits necesarios para codificar una fuente de información con un conjunto de códigos de longitud variable que cumplan ciertas restricciones.

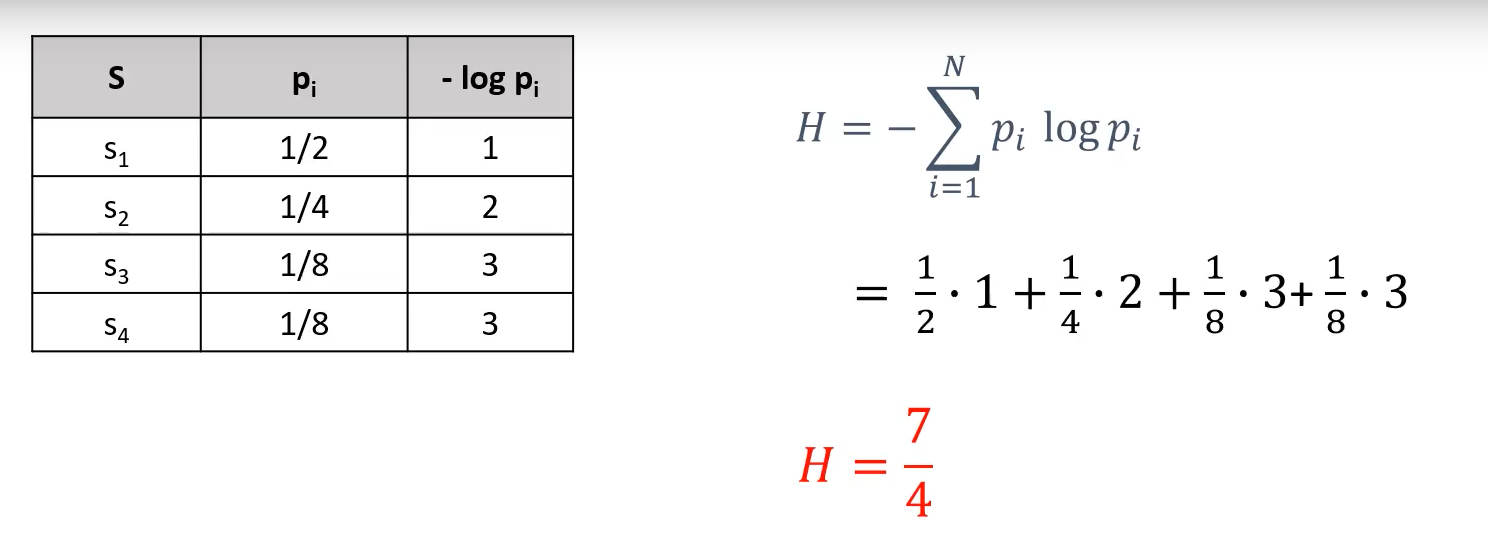
## Entropía

La entropía es una medida de incertidumbre o información promedio contenida en una fuente de datos. Se utiliza en la teoría de la información para cuantificar la cantidad de información contenida en una fuente de datos o para estimar la cantidad de información necesaria para transmitir una fuente de datos a través de un canal de comunicación.

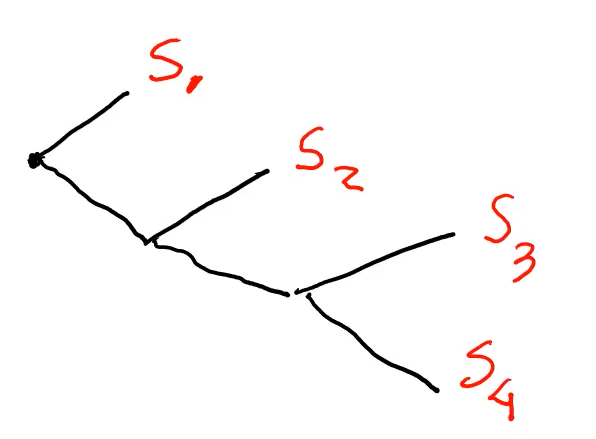
La entropía se mide en bits y se puede calcular a partir de las probabilidades de los símbolos que componen la fuente de datos.

“Mínimo número de preguntas binarias que hay que hacer en promedio para identificar un símbolo de la fuente.”

* Ejemplo:



* Árbol optimo



* Tener en cuenta: En la formula el log esta en base dos. La fórmula para cambiar de base es:

## Código de Huffman

El código de Huffman es un algoritmo de compresión de datos que se utiliza para crear códigos de longitud variable para cada símbolo de una fuente de datos, de tal manera que los símbolos más comunes tengan códigos más cortos y los símbolos menos comunes tengan códigos más largos. Fue desarrollado por David A. Huffman en 1952 mientras trabajaba en su tesis doctoral en el MIT.

El algoritmo de Huffman comienza con una lista de frecuencias de cada símbolo en la fuente de datos. A partir de esta lista, se construye un árbol de códigos binario de longitud variable que se utiliza para codificar los símbolos. El árbol se construye de la siguiente manera:

1. Se crean nodos para cada símbolo y se ordenan en orden ascendente de frecuencia.
2. Se toman los dos nodos con menor frecuencia y se combinan en un nuevo nodo padre cuya frecuencia es la suma de las frecuencias de sus hijos.
3. Se repite el paso 2 hasta que todos los nodos estén combinados en un solo nodo raíz.
4. Se asigna un código binario a cada símbolo en el árbol, de tal manera que los símbolos más comunes tengan códigos más cortos y los símbolos menos comunes tengan códigos más largos. Los códigos se asignan recorriendo el árbol de arriba hacia abajo, asignando un 0 a cada arco izquierdo y un 1 a cada arco derecho.

Una vez que se ha construido el árbol de códigos, se pueden codificar los símbolos de la fuente de datos utilizando los códigos asignados a cada símbolo en el árbol. La compresión se logra al reemplazar cada símbolo en la fuente de datos con su código correspondiente en el árbol de códigos.

El algoritmo de Huffman es óptimo en el sentido de que produce el conjunto de códigos de longitud variable con la longitud media más corta posible, dadas las frecuencias de los símbolos en la fuente de datos.

* Ejemplos:

