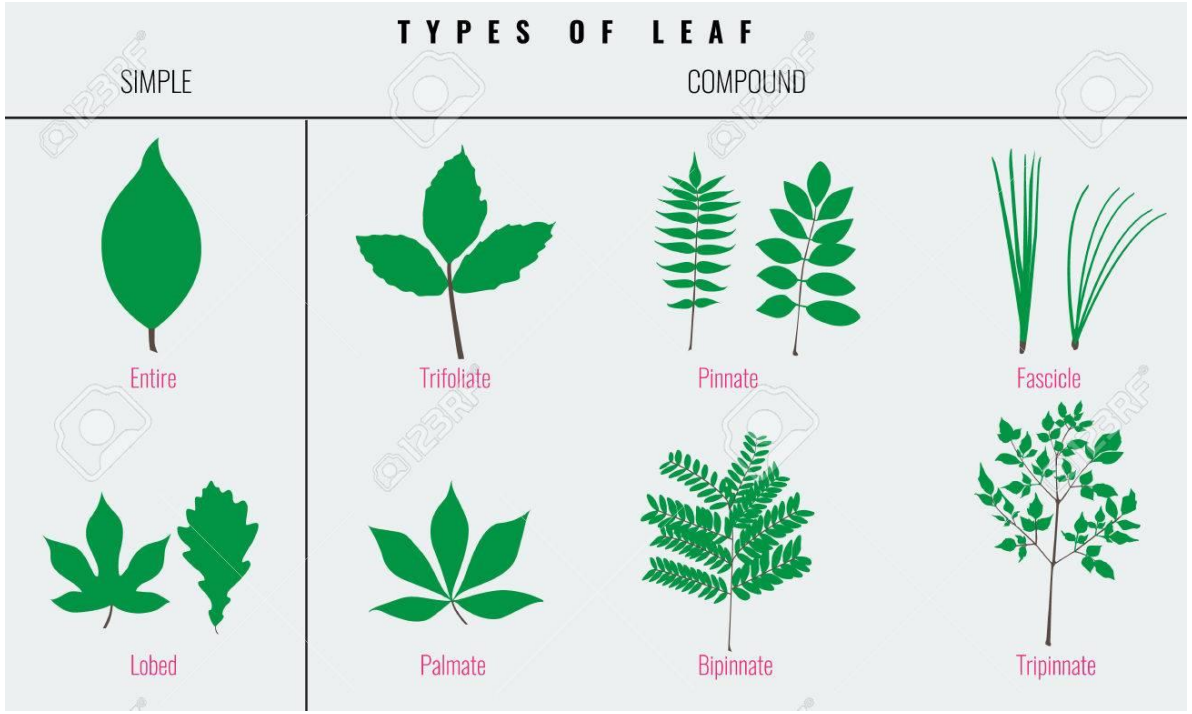
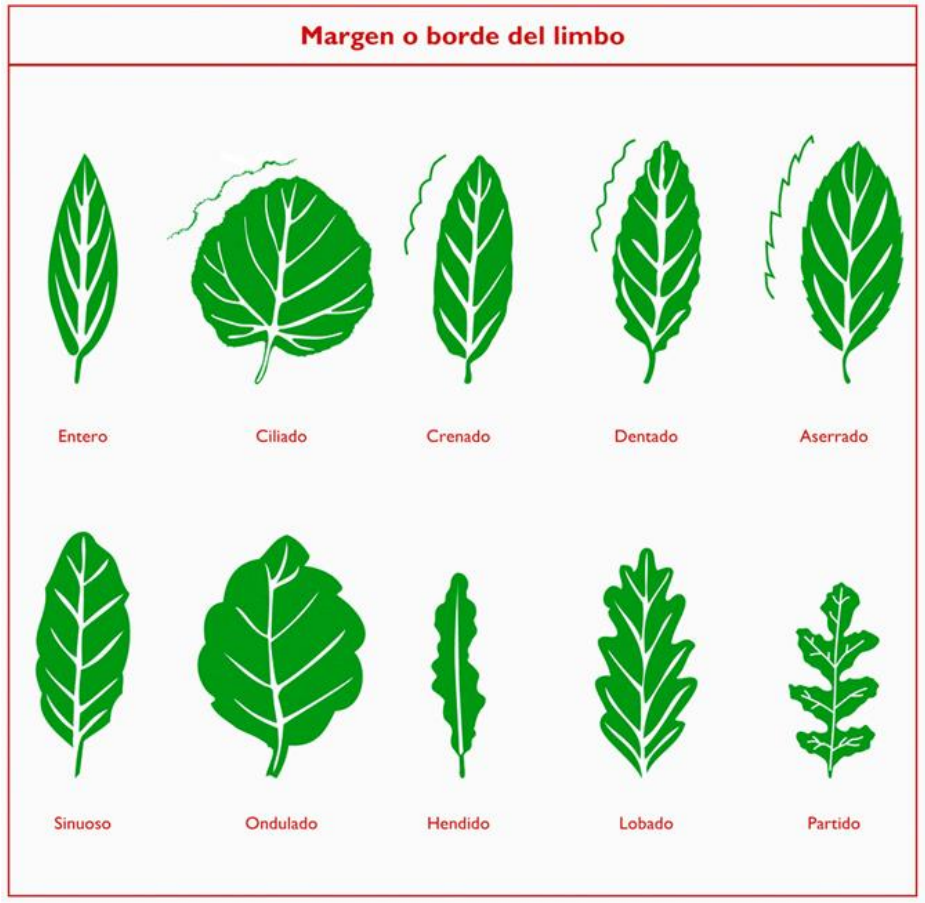
Índices de Similitud

Introducción

Problema


TYPES OF LEAF
SIMPLE
COMPOUND
Entire
Trifoliate
Pinnate
Fascicle
Lobed
Palmate
Bipinnate
Tripinnate


Margen o borde del limbo
Entero
Ciliado
Crenado
Dentado
Aserrado
Sinuoso
Ondulado
Hendido
Lobado
Partido

Problema


Roble
Arce
Nogal
Abedúl
Fresno
Castaño
Sauce Llorón
Olmo
Palmera
Álamo
Pino


HORMIGA CORTADORA DE HOJAS
Cut Ant
Atta
HORMIGA COSECHADORA
Harvester ant
Pogonomyrmex
HORMIGA DE CAMPO
Field Ant
Formica
HORMIGA DE FUEGO
Fire Ant
Solenopsis
HORMIGA DEL PAVIMENTO
Pavement Ant
Tetramorium Caespitum
HORMIGA DOMÉSTICA OLOROSA
Odorous House Ant
Tapinoma sessile
HORMIGA FALSA MIELERA
Small Honey, False Honey Ant
Prenolepsis Imparis
HORMIGA FARAONA
Pharaoh Ant
Monomorium Pharaonis
HORMIGA GRANDE AMARILLA
Yellow or Moisture Ants
HORMIGA LADRONA
Thief Ant
Solenopsis Molesta

Índices de similitud

- Un índice de similitud es una forma de medir el parecido que existe entre dos datos o muestras de datos.

- La propuesta de los primeros índices de similitud fueron propuestas por biólogos que se enfrentaron a problemas de clasificación.

- Para poder aplicar un índice de similitud se requiere tener datos, y saber el tipo de datos que se esta procesando.

Organización de la información

Muestra	6 patas	8 patas	Alas	veneno	colonia	antenas	…	…	…	…
Insecto 1	0	1	1	0	1	1				
Insecto 2	1	0	1	0	1	1				
Insecto 3	1	0	0	1	0	1				
Insecto 4	1	0	1	1	1	1				

Se requiere hacer comparación entre cadenas de 1s y 0s

0	1	0	1	0	1	0	1	1	0
0	1	0	1	0	1	0	1	0	1

Índices de Similitud

- Para datos binarios los índices de similitud generalmente se basan en la matriz de confusión.

- Esta matriz es un conteo de las características comunes y las no comunes entre dos muestras o datos.

		MUESTRA 1	
		SI	**NO**
MUESTRA 2	**SI**	**a**	**b**
	NO	**c**	**d**

Índice de similitud

- Emparejamiento simple: $\delta_{ij} = \frac{a+d}{a+b+c+d}$

0	1	0	1	0	1	0	1	1	0
0	1	0	1	0	1	0	1	0	1

	SI	NO
SI	4	1
NO	1	4

$$\delta_{ij} = \frac{4+4}{4+1+1+4} = 0.8$$

¿Y esto para que me sirve si soy financiero? Me interesa el dinero

Netflix

Índices de Similitud

- En Netflix es común que nos pidan calificar las películas que vemos y en base a eso nos pueden sugerir nuevas películas.

- La recomendación se hace en función de los usuarios que son más parecidos a mi, y me recomienda películas que los otros usuarios hayan visto y yo no.

	Pelicula 1	Pelicula 2	Pelicula 3	Pelicula 4	Pelicula 5	Pelicula 6	Pelicula 7	Pelicula 8	Pelicula 9	Pelicula 10
Usuario 1	1	1	0	1	1	1	0	0	1	0
Usuario 2	1	0	0	1	1	1	0	0	1	1
Usuario 3	1	0	0	0	0	0	0	0	0	1

Consideraciones

Índices de Similitud

- Hay que tener en cuenta que los índices de similitud son formas de medir que tan cercanos o similares son dos datos.
- Existen varios índices de similitud, y dependiendo de la información que se quiera saber es el índice que debemos de usar.

Índices de Similitud

- Volvamos con Netflix

	Terror 1	Terror 2	Terror 3	Infantil 1	Infantil 2	infantil 3	Romántica 1	Romántica 2	Romántica 3	Romántica 4
Usuario 1	1	0	0	0	0	0	0	0	0	0
Usuario 2	1	0	0	0	0	1	0	0	1	1
Usuario 3	0	0	0	0	0	0	0	0	0	1

El índice de similitud de emparejamiento simple nos diría que el usuario 1 y el usuario 2 tienen los mismos gustos, esto implica que al usuario que le gusta el terror le recomendaría películas románticas y viceversa.

Índice de similitud

- Jaccard: $\delta_{ij} = \frac{a}{a+b+c}$

	Terror 1	Terror 2	Terror 3	Infantil 1	Infantil 2	infantil 3	Romántica 1	Romántica 2	Romántica 3	Romántica 4
Usuario 1	1	0	0	0	0	0	0	0	0	0
Usuario 2	1	0	0	0	0	1	0	0	1	1
Usuario 3	0	0	0	0	0	0	0	0	0	1

	SI	NO
SI	0	1
NO	1	8

$$\delta_{ij} = \frac{0}{0+1+1} = 0$$

Índice de similitud

	Terror 1	Terror 2	Terror 3	Infantil 1	Infantil 2	infantil 3	Romántica 1	Romántica 2	Romántica 3	Romántica 4
Usuario 1	1	0	0	0	0	0	0	0	0	0
Usuario 2	1	0	0	0	0	1	0	0	1	1
Usuario 3	0	0	0	0	0	0	0	0	0	1

	SI	NO
SI	**1**	**0**
NO	**3**	**6**

Emparejamiento simple

$$\delta_{ij} = \frac{1+6}{1+0+3+6} = 0.7$$

Jaccard

$$\delta_{ij} = \frac{1}{1+0+3} = 0.25$$