



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA




# Clasificación Multi-Etiqueta en Entornos Jerárquicos

**Trabajo Final de Máster**  
**Máster en Ingeniería del Software, Métodos Formales  
y Sistemas de Información**

**Alumno: José Luis de la Cruz Garrido**  
**Director: Cèsar Ferri Ramirez**

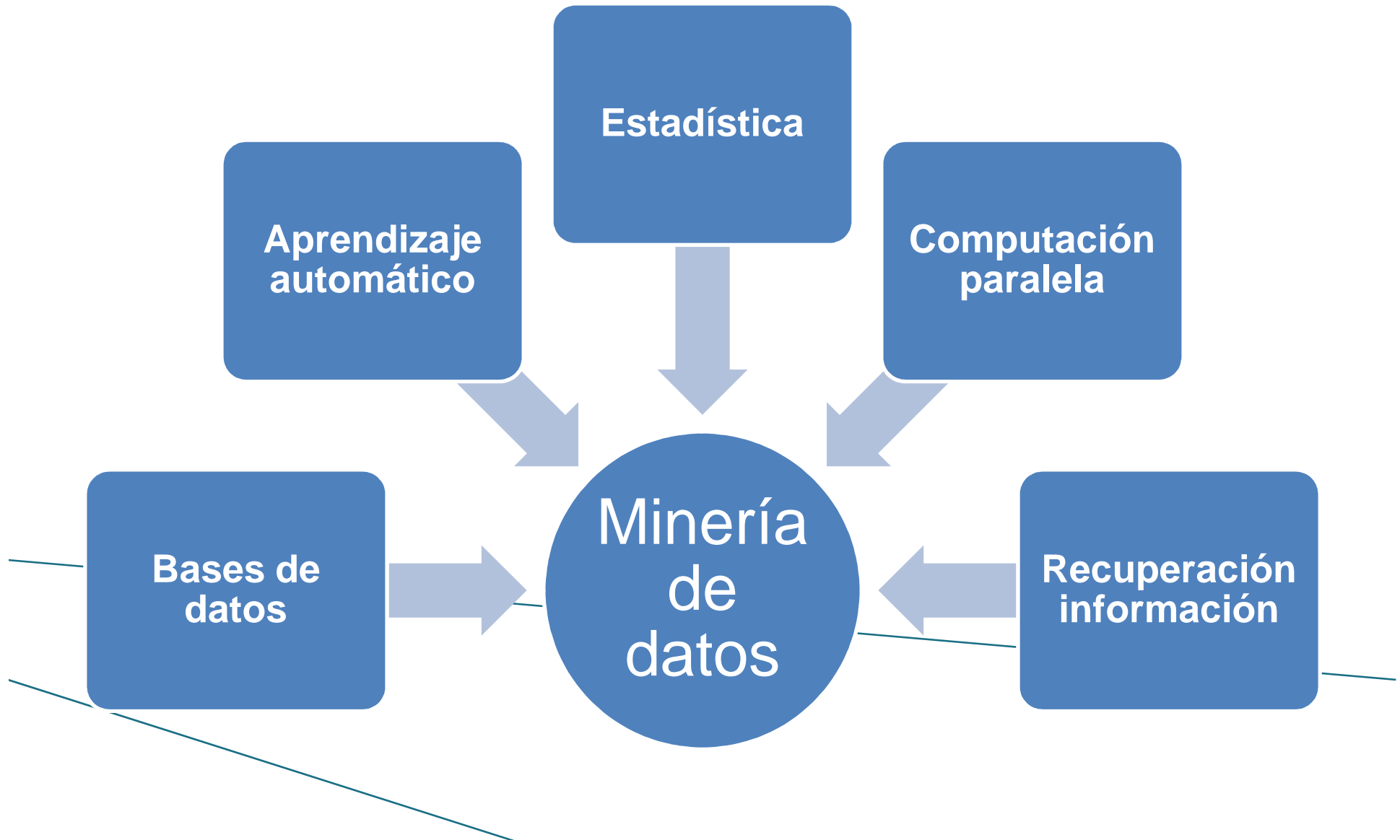
# Índice

- ▶Objetivos
  - ▶Minería de datos
  - ▶Clasificación Multi-Etiqueta
  - ▶Clasificación Jerárquica
  - ▶Wikipedia
  - ▶Nueva Métrica
  - ▶Tecnologías
  - ▶Experimentos y resultados
  - ▶Conclusiones y trabajo futuro
- 

# Objetivos

- ▶Revisión del campo clasificación jerárquica multi-etiqueta
- ▶Construcción de datasets con datos de la Wikipedia
- ▶Realización de experimentos de clasificación jerárquica multi-etiqueta
- ▶Proponer una métrica basada en distancias para medir la precisión de los modelos generados
- ▶Validar experimentalmente los modelos generados y la métrica propuesta

# Minería de datos



# Clasificación Multi-Etiqueta

La clasificación multi-etiqueta es un tipo de clasificación supervisada donde una instancia puede estar asociada a más de un patrón o clase

$L = \{\text{berenjenas, recetas fáciles, plato de la semana}\}$

## PLATO DE LA SEMANA

### Berenjenas Fritas

Tiempo de Preparación: 10 minutos.

Ingredientes:

- 4 Berenjenas.
- Sal.
- Pimienta.
- 4 Cucharadas de harina y aceite.

Realización:

1. Preparar los ingredientes:
  - Lavar y cortar las berenjenas en rodajas.
  - Dejar que suelten el agua durante 30 minutos.
2. Cocción:
  - Enharinar y freír las berenjenas durante 5 minutos.
  - Depositarlas sobre papel absorbente..

# Métodos de Clasificación Multi-Etiqueta

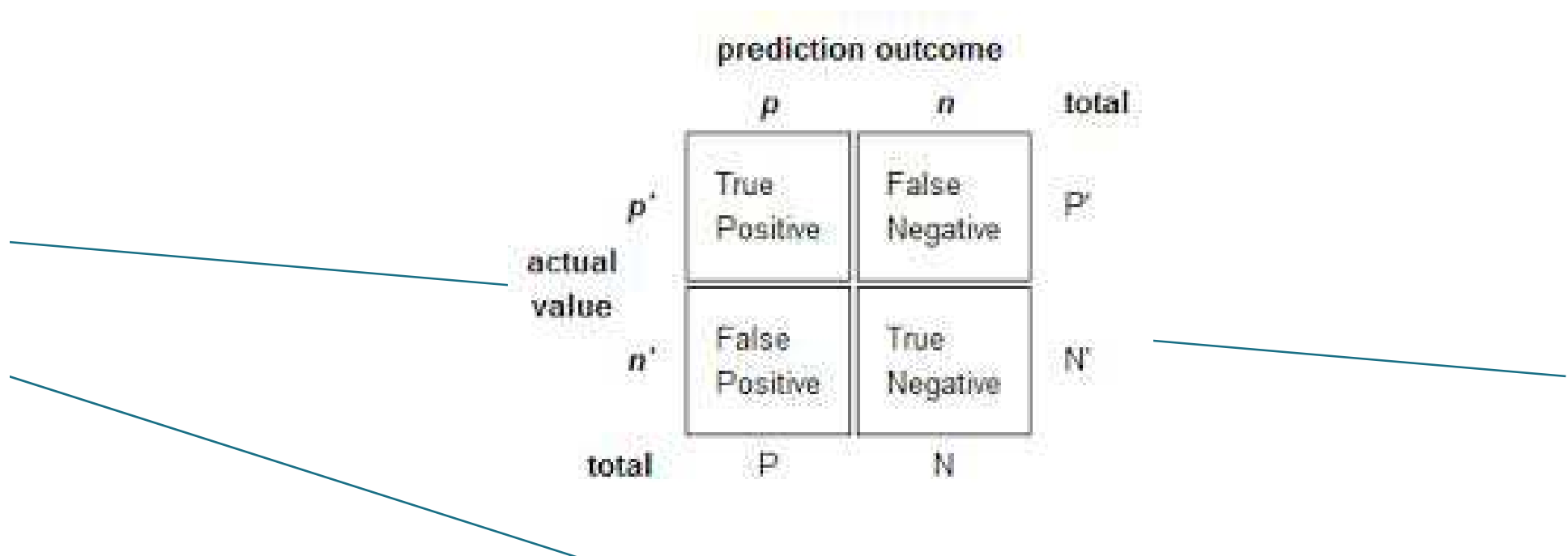
## Métodos de Clasificación Multi-Etiqueta

- Métodos de adaptación de algoritmos: Emplean una técnica de clasificación adaptada para trabajar directamente con datos multi-etiqueta.
- Métodos de transformación de problemas: Convierten un conjunto de datos multi-etiqueta en uno o varios conjuntos de datos de una sola etiqueta.

# Métricas de Clasificación Multi-Etiqueta

Habitualmente, se han utilizado métricas basadas en la matriz de confusión para medir la precisión de los modelos de clasificación.

¿Es mejorable la precisión de estas métricas?



# Métricas de Clasificación Multi-Etiqueta

Etiquetas	Refresco	Bebida Isotónica	Lejía	Coca Cola	Fanta	Gasolina	Bebida sin alcohol
Etiquetas reales	X			X			X
Clasificador 1		X			X		X
Clasificador 2	X		X			X	

Cualquier métrica basada en la matriz de confusión tendrá los mismos valores para el clasificador 1 y el clasificador 2. ¿Son igual de precisos estos dos clasificadores?



# Métricas de Clasificación Multi-Etiqueta

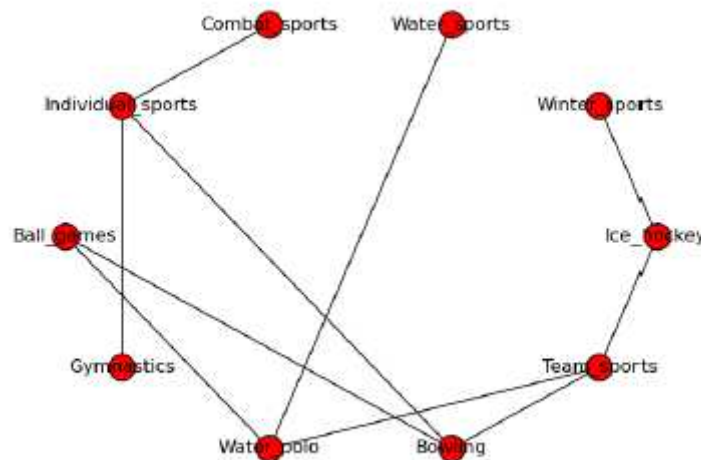
**Problema:** Las métricas basadas en la matriz de confusión solo tienen en cuenta aciertos y errores, penalizando por igual todo los errores.

Cuando se tiene información sobre las relaciones entre clases, tiene sentido definir métricas que no penalicen todos los errores de la misma manera.

En el caso de entornos jerárquicos, una buena manera de medir la precisión es definir métricas basadas en distancias en la jerarquía.

# Clasificación Jerárquica

- ▶ Las clases están organizadas en jerarquías, normalmente un árbol o un DAG
- ▶ La jerarquía ayuda a la hora de definir métricas, implementar algoritmos y entrenar modelos.



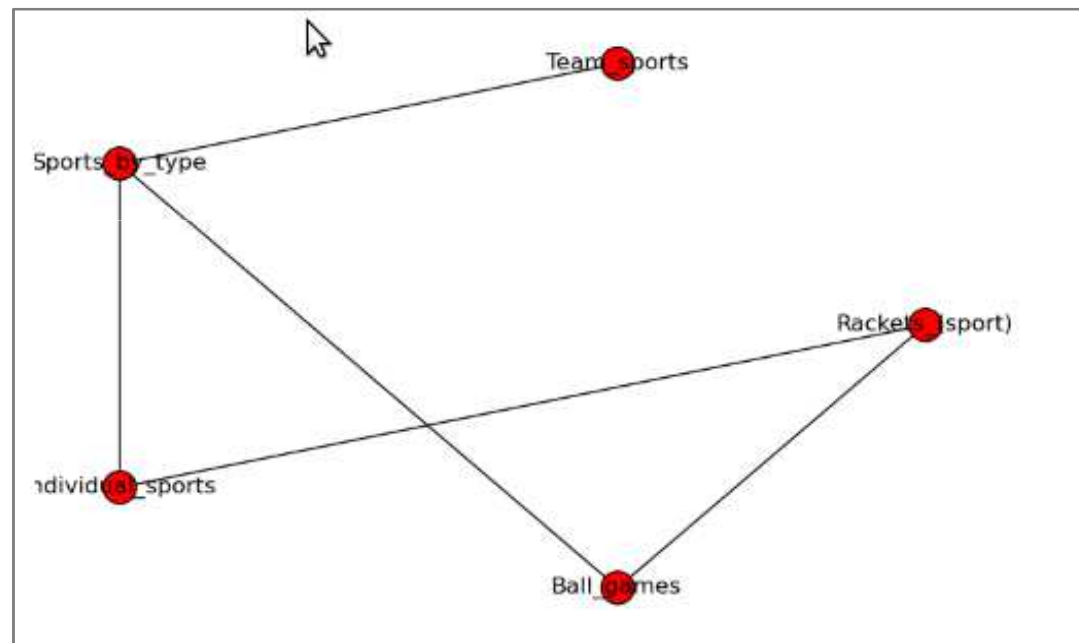
# Wikipedia

## ►Wikipedia

- Los artículos de la Wikipedia están etiquetados por categorías
- Las categorías están organizadas jerárquicamente
- Sports
  - ┆ Team Sports
  - ┆ Water Sports
- Las categorías forman un DAG (Grafo Dirigido Acíclico).
- Asignación automática de las etiquetas:
  - ┆ Clasificación Jerárquica Multi-etiqueta

# Wikipedia

- Las categorías forman grafos en lugar de árboles



# Nueva Métrica

- ▶ Dado un DAG con  $E$  aristas
- ▶  $P$  etiquetas predichas,  $R$  etiquetas reales
- ▶  $\text{Dist}(P-R)$ , Distancia en el grafo entre las etiquetas reales y las predichas
- ▶  $\text{Dist}(R-P)$ , Distancia en el grafo entre las etiquetas predichas y las reales
- ▶  $D = [\text{Dist}(P-R) / (E * |R|) + \text{Dist}(R-P) / (E * |P|)] / 2$

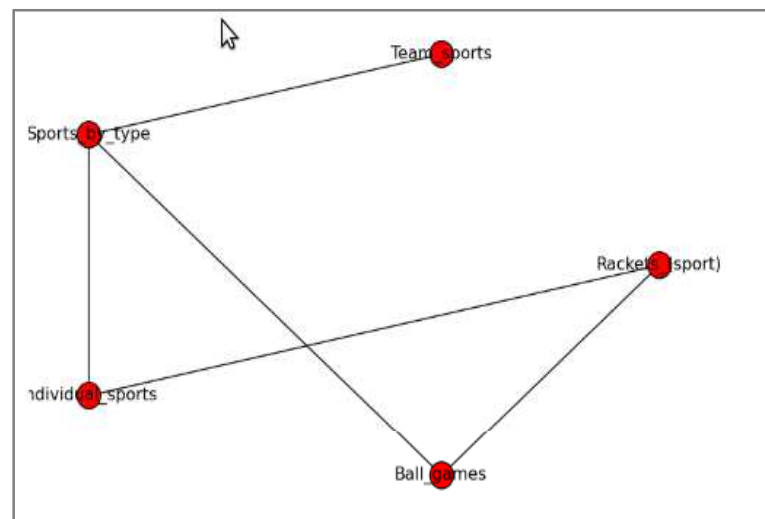
# Nueva Métrica: Ejemplo

Ejemplo	Sports_by_type	Ball_games	Individual_sports	Teams_ports	Rackets_sports
Football		✓		✓	
Chess			✓		
Golf		✓	✓		

Tabla 4.1: Tabla etiquetas reales del conjunto de datos

Ejemplo	Sports_by_type	Ball_games	Individual_sports	Teams_ports	Racket_sports
Football		✓			✓
Chess			✓		
Golf			✓	✓	

Tabla 4.2: Tabla etiquetas predichas por el clasificador

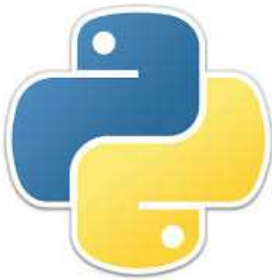


	Football	Chess	Golf
$D(P - R)$	2	0	2
$D(R - P)$	1	0	2
$D = [Dist(P-R)/(E* R ) + Dist(R-P)/(E* P )]/2$	3/20	0	4/20

# Tecnologías



NETWORKX



NLTK



# Experimentos

## ► Datasets empleados

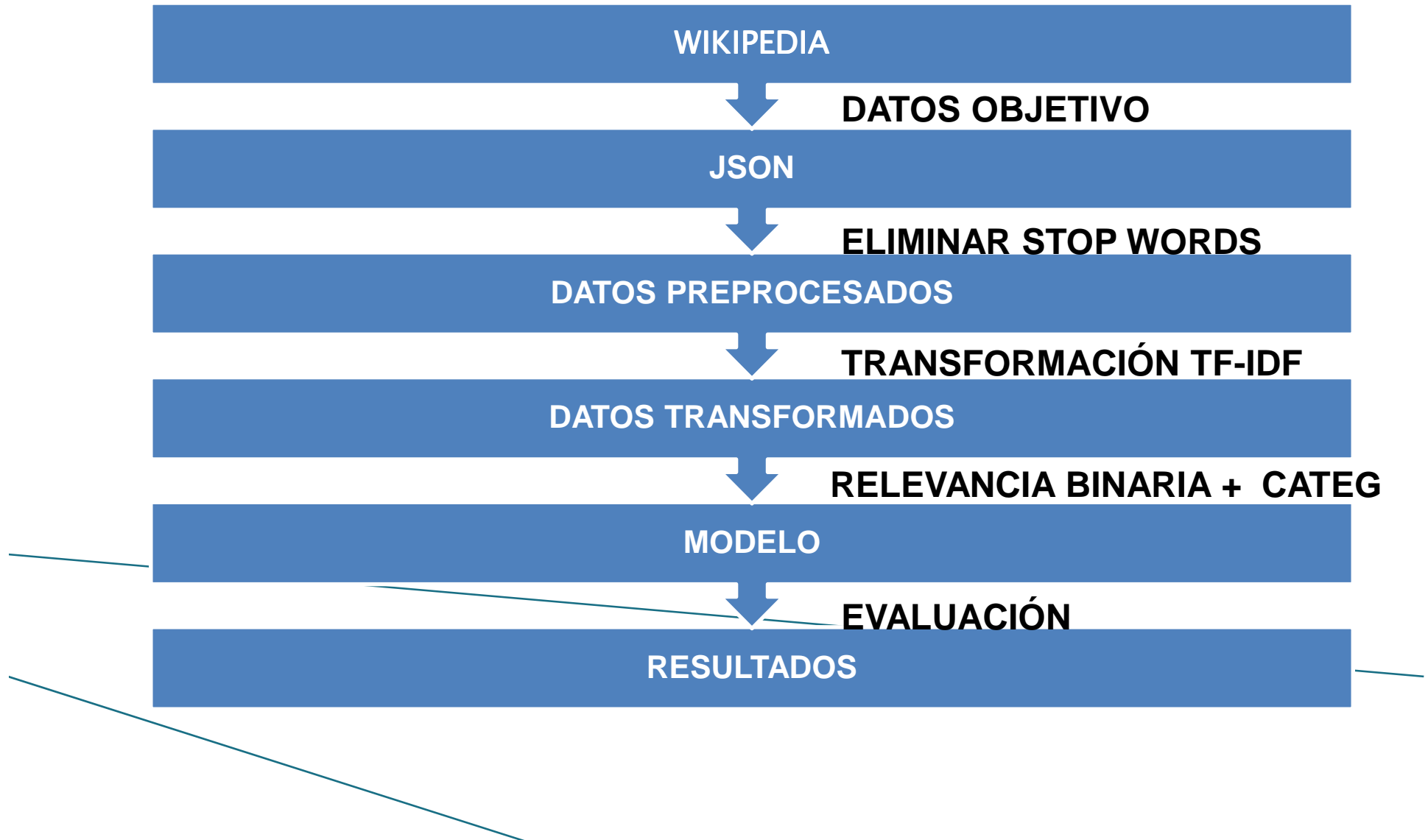
dataset	ejemplos	etiquetas	cardinalidad	densidad	combinaciones
deportes	705	10	1.2	0.12	30
software	714	12	2.21	0.18	33
profesiones	344	11	1.014	0.09	15
bebidas	189	11	2.06	0.19	33
comidas	1892	28	1.06	0.037	85

Tabla 5.1: Métricas descripción de los datasets

- Naives Bayes, CART, Logistic, Knn y LinearSVC
- Hamming Loss, Precision, Recall, Distancia Simétrica, Macro-F1, Micro-F1, Accuracy



# Experimentos



# Experimentos

	Relevancia Binaria				
Métrica	Naives Bayes	Knn	Logistic	SVM	DT
Hamming loss	0.119	0.062	0.092	0.064	0.099
Precision	0.180	0.768	0.702	0.825	0.618
Recall	0.011	0.695	0.286	0.584	0.548
Distancia Simétrica	0.092	0.038	0.072	0.043	0.058
Macro-F1	0.009	0.676	0.234	0.552	0.546
Micro-F1	0.022	0.729	0.423	0.683	0.570
Accuracy	0.009	0.567	0.211	0.468	0.327

Tabla 5.2: Resultados para el dataset deportes

	Relevancia Binaria				
Métrica	Naives Bayes	Knn	Logistic	SVM	DT
Hamming loss	0.083	0.067	0.051	0.048	0.079
Precision	0.749	0.832	0.782	0.847	0.808
Recall	0.682	0.801	0.815	0.837	0.796
Distancia Simétrica	0.012	0.007	0.005	0.004	0.008
Macro-F1	0.270	0.391	0.340	0.388	0.418
Micro-F1	0.767	0.828	0.866	0.875	0.801
Accuracy	0.541	0.539	0.655	0.661	0.501

Tabla 5.3: Resultados para el dataset software

# Experimentos

	Relevancia Binaria				
Métrica	Naives Bayes	Knn	Logistic	SVM	DT
Hamming loss	0.0953	0.0389	0.0886	0.0558	0.0672
Precision	0.2554	0.8186	0.3637	0.7765	0.7079
Recall	0.0599	0.7288	0.1260	0.4785	0.6560
Distancia Simétrica	0.088	0.033	0.079	0.049	0.045
Macro-F1	0.1043	0.6292	0.133	0.3982	0.4960
Micro-F1	0.1119	0.7907	0.2229	0.6339	0.6629
Accuracy	0.0609	0.7241	0.1279	0.4766	0.4909

Tabla 5.4: Resultados para el dataset profesiones

	Relevancia Binaria				
Métrica	Naives Bayes	Knn	Logistic	SVM	DT
Hamming loss	0.0391	0.0243	0.0351	0.0249	0.03558
Precision	0.0556	0.7963	0.4451	0.8339	0.5754
Recall	0.0154	0.5201	0.1072	0.4115	0.5306
Distancia Simétrica	0.0356	0.0212	0.0319	0.0217	0.0249
Macro-F1	0.0171	0.5547	0.1266	0.4573	0.4697
Micro-F1	0.0264	0.6251	0.1920	0.5628	0.5364
Accuracy	0.0609	0.5079	0.1073	0.3969	0.3430

Tabla 5.5: Resultados para el dataset comidas

# Experimentos

	Relevancia Binaria				
Métrica	Naives Bayes	Knn	Logistic	SVM	DT
Hamming loss	0.144	0.118	0.139	0.105	0.124
Precision	0.413	0.663	0.460	0.702	0.674
Recall	0.39	0.613	0.407	0.564	0.652
Distancia Simétrica	0.0428	0.02752	0.04715	0.0279	0.02579
Macro-F1	0.12	0.299	0.133	0.254	0.362
Micro-F1	0.50	0.659	0.523	0.666	0.662
Accuracy	0.06	0.328	0.095	0.290	0.216

Tabla 5.6: Resultados para el dataset bebidas

# Experimentos


	Relevancia Binaria				
Métrica	Naives Bayes	Knn	Logistic	SVM	DT
Hamming loss	0	3	0	2	0
Precision	0	1	0	4	0
Recall	0	2	0	1	2
Distancia Simétrica	0	3	0	1	1
Macro-F1	0	3	0	1	1
Micro-F1	0	3	0	2	0
Accuracy	0	3	0	2	0

Tabla 5.7: Tabla algoritmo ganador métrica

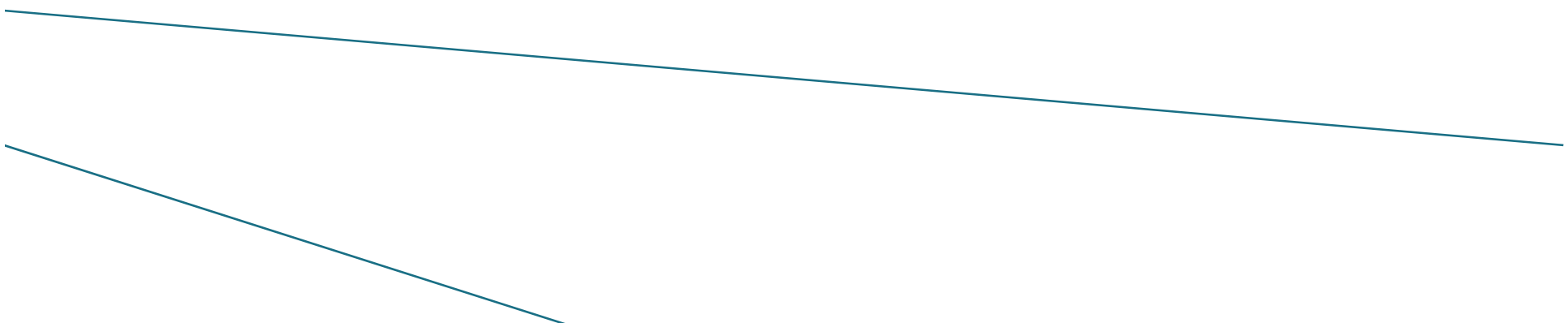
	Relevancia Binaria				
	Naives Bayes	Knn	Logistic	SVM	DT
Victorias	0	18	0	13	4

Tabla 5.8: Tabla victorias totales algoritmo

# Conclusiones

- ▶ Sencillez del uso de la API de la Wikipedia
  - ▶ Dificultad para encontrar conjuntos de datos medianos (a nivel estructural) con cardinalidad de etiqueta mayor que 2
  - ▶ Buen funcionamiento de la métrica definida
  - ▶ Problemas de desbalanceo y ruido
  - ▶ Bases experimentales de la clasificación jerárquica multi-etiqueta establecidas
- 

# Trabajo Futuro

- ▶ Comparación con otros trabajos relacionados a nivel experimental
  - ▶ Definir más métricas basadas en distancias
  - ▶ Definir algoritmos que minimicen el valor de la métrica
  - ▶ Considerar otros datasets
  - ▶ Colofón: Recomendador de etiquetas para mediawiki
- 
- Two decorative teal lines are positioned at the bottom of the slide. The upper line starts on the left and slopes downwards to the right. The lower line also starts on the left and slopes downwards to the right, positioned below the first line.

Muchas gracias por su atención

