

Introducción

Se denomina aprendizaje automático a la disciplina que estudia algoritmos que buscan patrones en conjuntos complejos de datos.

Una de las áreas de aplicación es la clasificación de textos, proceso por el cuál una serie de documentos son asignados dentro de determinadas categorías.

El objetivo de este trabajo es investigar el uso de técnicas de aprendizaje supervisado para clasificar automáticamente obras de ficción (como novelas o cuentos) dentro de sus respectivos géneros literarios.

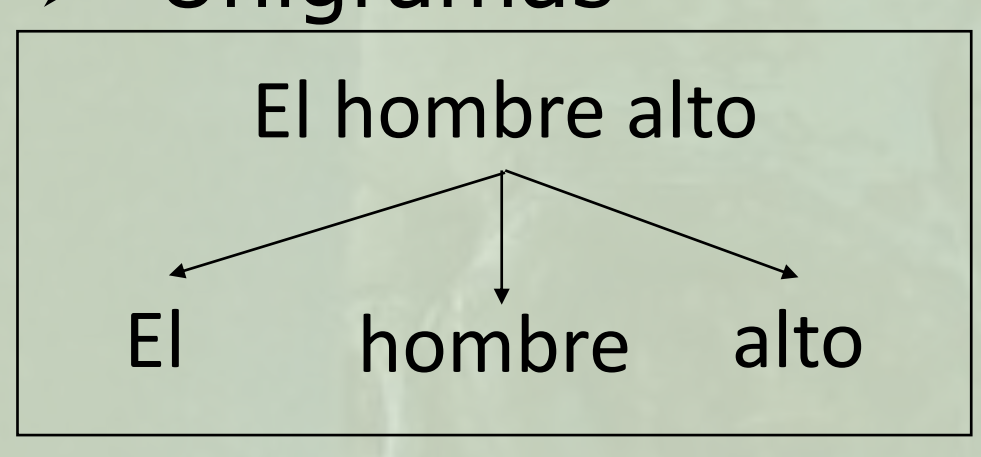


Análisis de textos

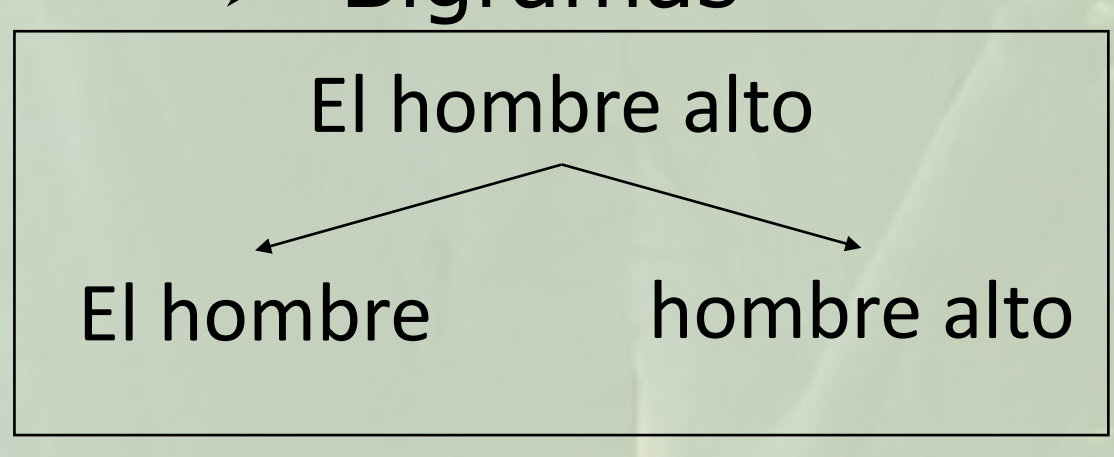
Considerando distintas formas de construir variables predictoras a partir de los textos, se construyen 8 escenarios de análisis.

Segmentación del texto

- Unigramas



Bigramas



Estadísticas

- Frecuencia de términos (TF)
Nº de veces que aparece un unigrama o bigrama en un texto.

TF-IDF

TF ponderado. Penaliza unigramas o bigramas que aparecen en muchos textos.

Filtro

Se eliminan nombres propios utilizando el diccionario de la RAE, para evitar un posible efecto de sobreajuste.

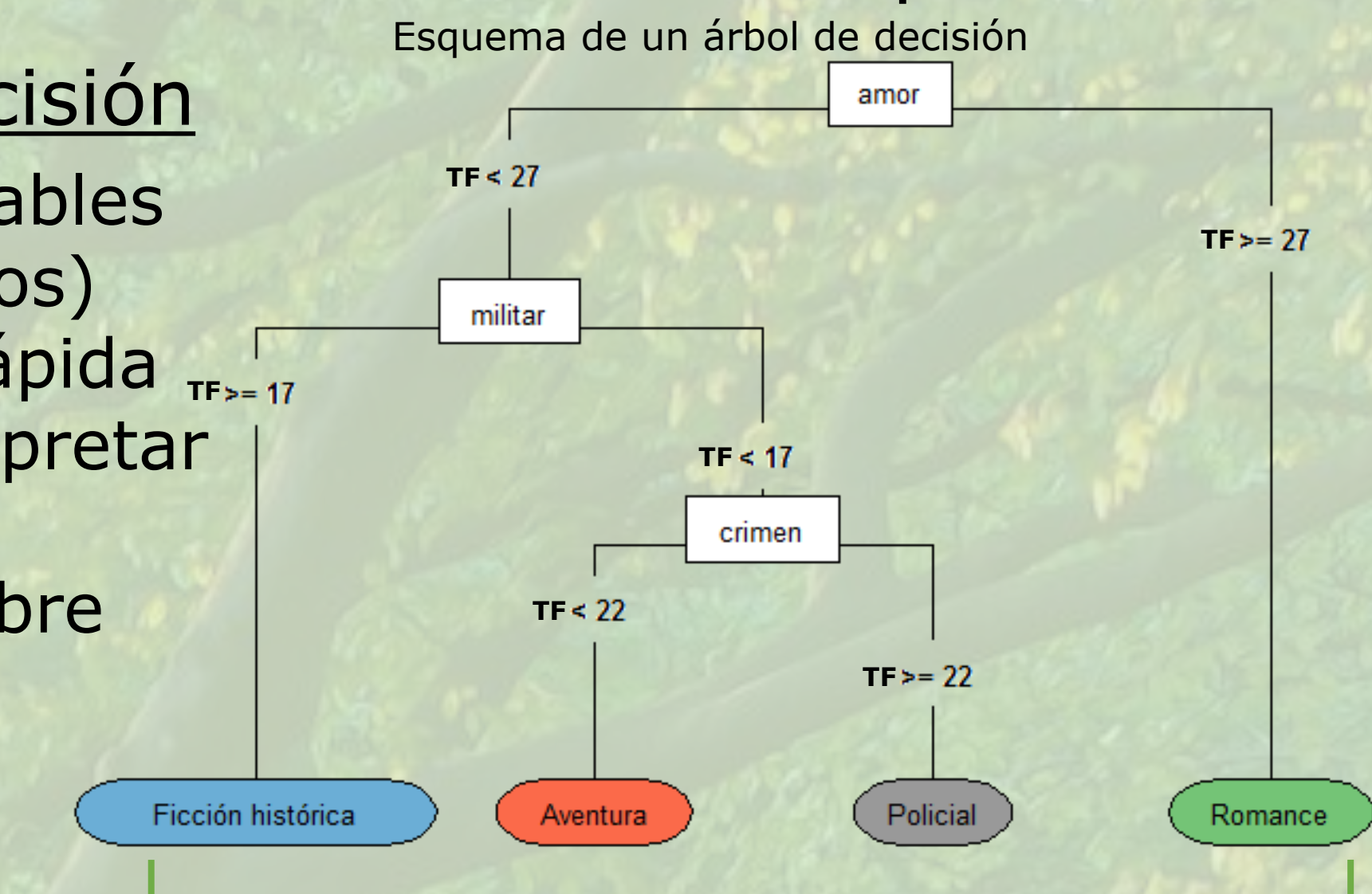
Escenarios de análisis			
Escenario	Segmentación	Estadística	Filtro
A	Unigramas	Frecuencia	No
B	Bigramas	Frecuencia	No
C	Unigramas	TF-IDF	No
D	Bigramas	TF-IDF	No
E	Unigramas	Frecuencia	Si
F	Bigramas	Frecuencia	Si
G	Unigramas	TF-IDF	Si
H	Bigramas	TF-IDF	Si

Clasificación supervisada

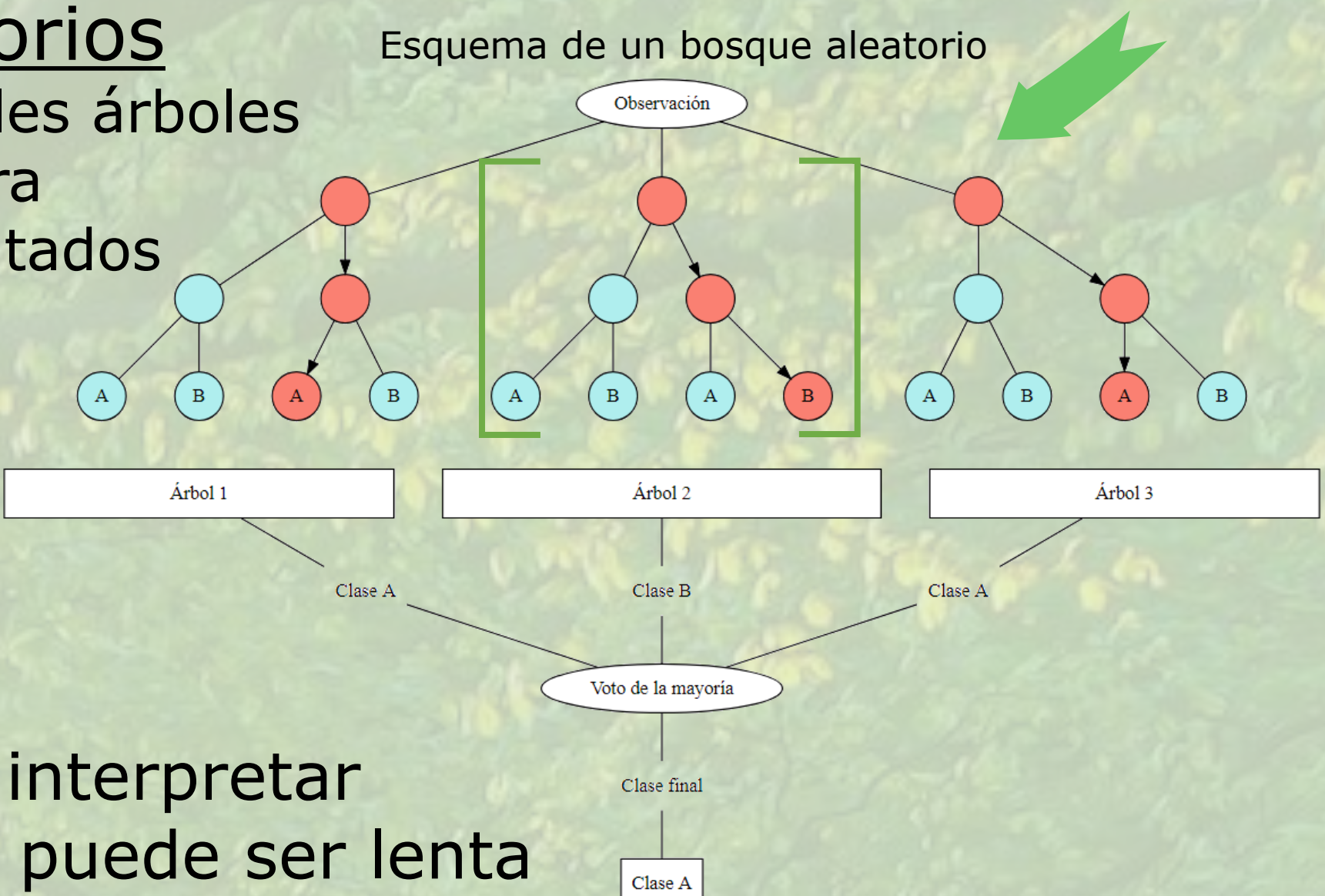
Algoritmos que se aplican cuando cada observación cuenta con una etiqueta previamente especificada. Se busca predecir futuros valores de la variable respuesta en función de una serie de explicativas.

- Árboles de decisión
 - Flexibles (aplicables en muchos casos)
 - Construcción rápida
 - Fáciles de interpretar
 - Desempeño pobre
 - Inestables (tienden al sobreajuste)
- Bosques aleatorios
 - Se construyen múltiples árboles y la predicción se logra combinando sus resultados
 - Muy buen desempeño
 - Conclusiones más robustas
 - Más difíciles de interpretar
 - La construcción puede ser lenta

Esquema de un árbol de decisión



Esquema de un bosque aleatorio



Resultados

Coefficiente Kappa y precisión por escenario

Escenario	Árboles de decisión		Bosques aleatorios	
	Kappa	Precisión	Kappa	Precisión
A	0,33	43%	0,57	64%
B	0,30	41%	0,41	51%
C	0,16	30%	0,51	59%
D	0,24	37%	0,39	50%
E	0,23	36%	0,58	64%
F	0,21	35%	0,38	48%
G	0,33	43%	0,42	52%
H	0,24	37%	0,36	47%

➤ **Precisión:** Porcentaje de unidades bien clasificadas.

➤ **Kappa:** Acuerdo entre el modelo y los datos, teniendo en cuenta el azar ($0 \leq \kappa \leq 1$)

Matriz de confusión del escenario E

	aventura	ciencia ficción	fantástico	historia	policial	romance	terror
aventura	23 (68%)	0 (0%)	2 (6%)	5 (15%)	2 (6%)	1 (3%)	1 (3%)
ciencia ficción	9 (39%)	6 (26%)	1 (4%)	3 (13%)	0 (0%)	2 (9%)	2 (9%)
fantástico	1 (4%)	0 (0%)	11 (42%)	1 (4%)	3 (12%)	9 (35%)	1 (4%)
historia	2 (4%)	0 (0%)	4 (8%)	39 (81%)	1 (2%)	1 (2%)	1 (2%)
policial	0 (0%)	0 (0%)	1 (4%)	1 (4%)	19 (79%)	1 (4%)	2 (8%)
romance	0 (0%)	0 (0%)	2 (5%)	5 (13%)	0 (0%)	32 (82%)	0 (0%)
terror	0 (0%)	1 (3%)	5 (17%)	3 (3%)	3 (10%)	5 (17%)	15 (50%)


Clase observada vs Clase predicha

Precisión por escenario

	Sin filtro		Con filtro	
	TF	TF-IDF	TF	TF-IDF
Unigramas	64%	59%	64%	52%
Bigramas	51%	50%	48%	47%

Conclusiones

- Se logró plantear un modelo para clasificar automáticamente textos de ficción, alcanzando una precisión del 64% con bosques aleatorios.
- Para mejorar el desempeño, se podría aplicar un enfoque multi-etiqueta o utilizar técnicas de mejora (por ejemplo, *boosting*)
- En futuras investigaciones se podría evaluar el uso de nuevas estadísticas o metodologías (por ejemplo, análisis de sentimientos)



Referencias y código