

Clasificación supervisada de textos de ficción usando bosques aleatorios

Lic. Santiago García Sánchez

Mg. Diego Marfetán Molina

Mg. Marcos Miguel Prunello

Introducción

- **Aprendizaje automático / *machine learning***: Estudio y aplicación de algoritmos que descubren patrones presentes en datos provistos como entrada.
 - **Aprendizaje no supervisado**: Variables explicativas $X_1, X_2 \dots X_n$
 - **Aprendizaje supervisado**: Variables explicativas $X_1, X_2 \dots X_n$ y variable respuesta Y
- **Clasificación de textos**: Proceso por el cual documentos son asignados a categorías.
 - **Documento**: Unidad de datos textuales (libro, noticia, e-mail, etc.)
 - **Corpus**: Conjunto de documentos.

Objetivos

- **Objetivo principal:** Construir un clasificador que permita categorizar obras de ficción en español dentro de géneros literarios predeterminados.
- **Objetivos secundarios:**
 - Plantear distintos escenarios de análisis y evaluar el desempeño del algoritmo en cada uno de ellos.
 - Determinar qué géneros presentaron un mayor % de unidades correctamente clasificadas.
 - Calcular qué variables resultaron más importantes para el proceso de clasificación.

Metodología

- Conjunto de datos:

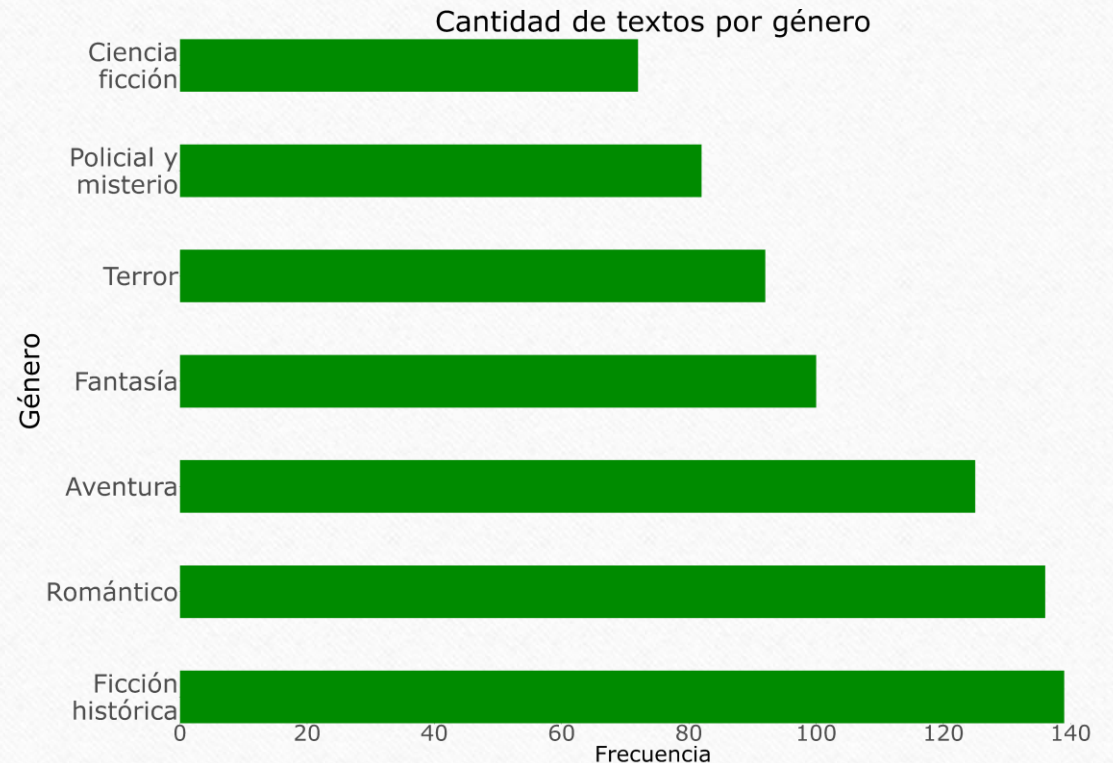
- Obras de ficción en español o traducidas.
- Fuentes: 9 sitios web con libros sin derechos de autor.
- Se recolectan con



- Etiquetado:

- Se toma la etiqueta más frecuente de

goodreads



Análisis de textos

➤ Preprocesamiento:

- Importar los textos al software.
- Limpieza de los documentos.
- Conversión a minúsculas:
 - {A, B,...Z} → {a, b,...z}
- Eliminar palabras no deseadas:
 - Preposiciones, conjunciones, artículos, etc.

➤ Lematización:

{amigo, amiga, amigos, amigas} → amigo

{alto, alta, altos, altas} → alto

{hago, haces, hacemos....} → hacer

➤ Segmentación / tokenización:

- Token: Unidad significativa de texto:
 - Caracteres.
 - Palabras.
 - n -gramas (Secuencia de n palabras consecutivas)

- Segmentación en palabras / unigramas:



- Segmentación en bigramas:



Análisis de textos

➤ Construcción de atributos:

❖ **Primer enfoque:** Frecuencia de términos (T.F. o F.T.): Frecuencia absoluta del término t en el documento d :

- $tf(t, d) = fd(t)$
- Enfoque bolsa de palabras (*bag of words*):

*“Hay un perro, un gato y
un pez”*



Palabra	Frecuencia
Un	3
Hay	1
Perro	1
Gato	1
Y	1
Pez	1

Análisis de textos

➤ Construcción de atributos:

❖ Segundo enfoque: Frecuencia de término - frecuencia inversa de documento (TF-IDF):

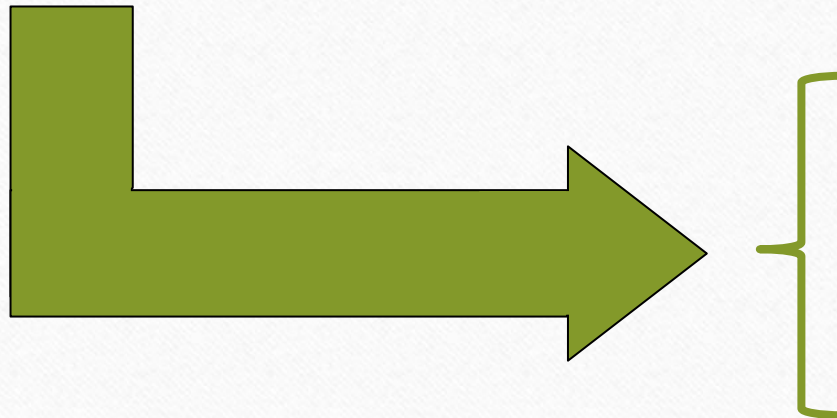
- $tf-idf(t,d) = tf_{td} \times idf_t$
- Donde: $idf_t = \ln\left(\frac{n^\circ \text{ total de documentos en el corpus}}{n^\circ \text{ de documentos que contienen a } t}\right)$
- Mayores valores → Términos muy frecuentes en unos pocos documentos.
- Menores valores → Términos que aparecen en muchos documentos o con baja frecuencia.

➤ Filtro:

- Se piensa que los nombres propios pueden llevar al sobreajuste del modelo.
- Se recurre al diccionario de la RAE eliminando términos que no figuren en él.
- Para cada escenario, se añade una variante en la que se aplicó el filtro.

➤ **Total:** 8 escenarios.

Escenario	Tokenización	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

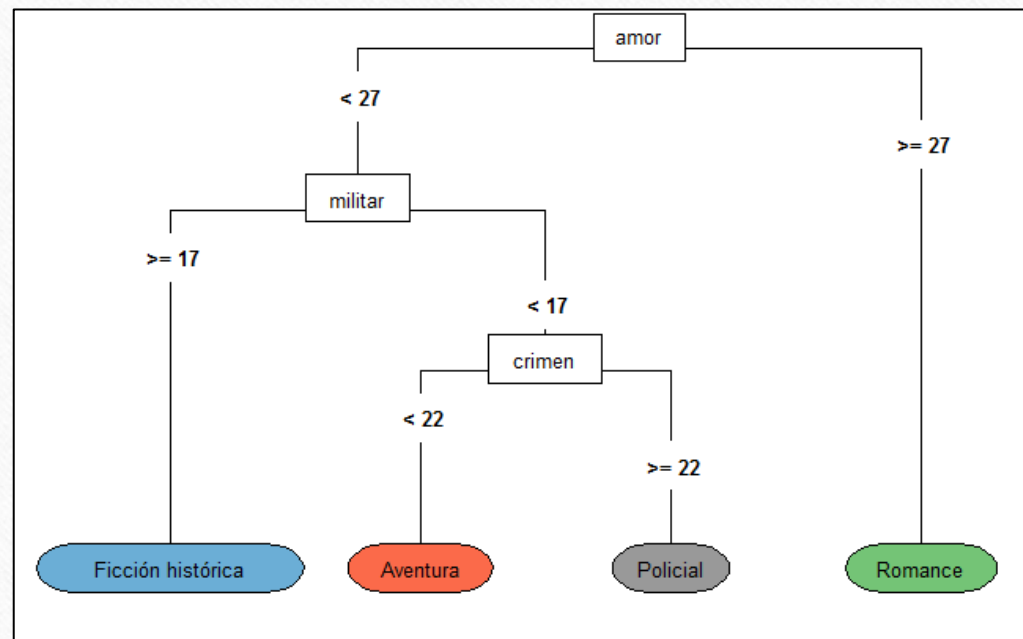


➤ **Matriz de documentos-términos (DTM)**

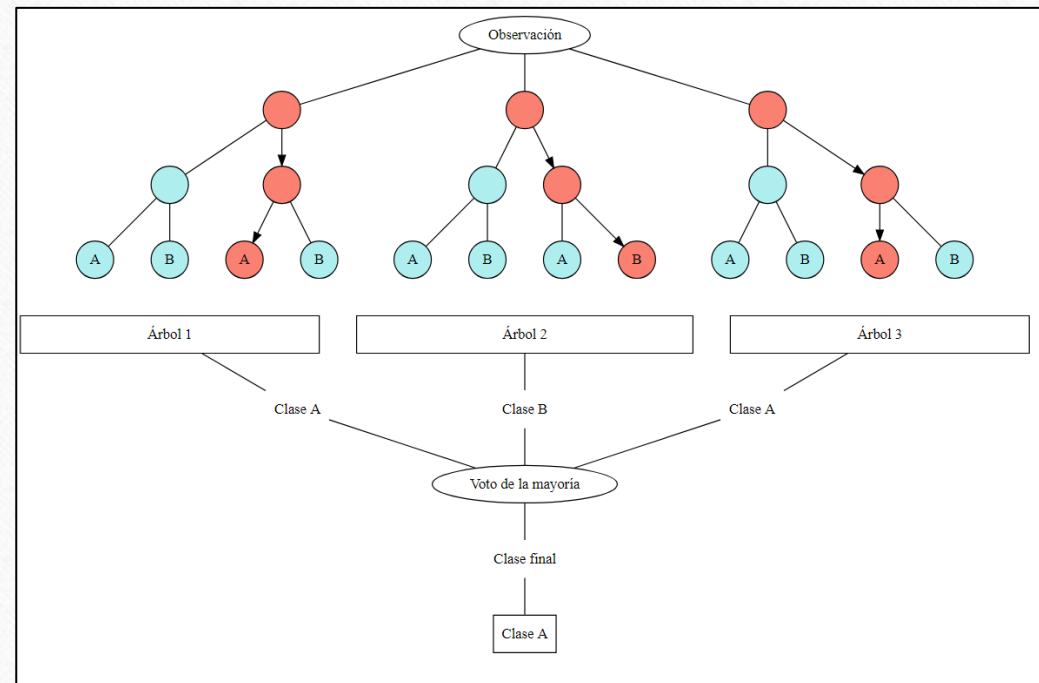
	Token 1	Token 2	...	Token n
Documento 1	Estadística	Estadística	...	Estadística
Documento 2	Estadística	Estadística	...	Estadística
...
Documento m	Estadística	Estadística	...	Estadística

Aprendizaje automático

Árboles de decisión

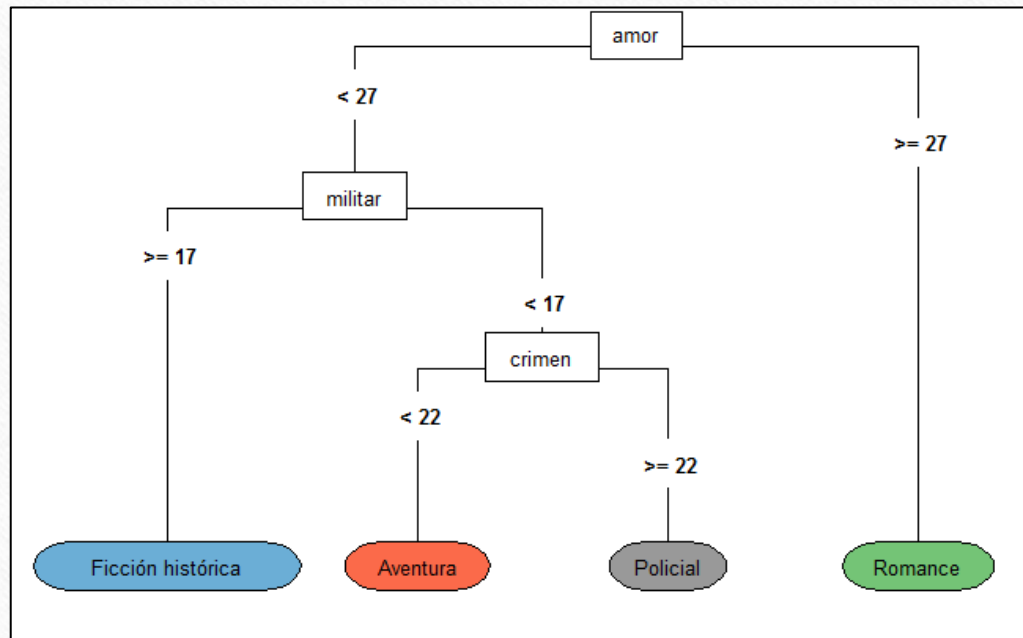


Bosques aleatorios



Aprendizaje automático

Árboles de decisión



- Algoritmo utilizado: **CART**
 - Intenta minimizar el índice de Gini:

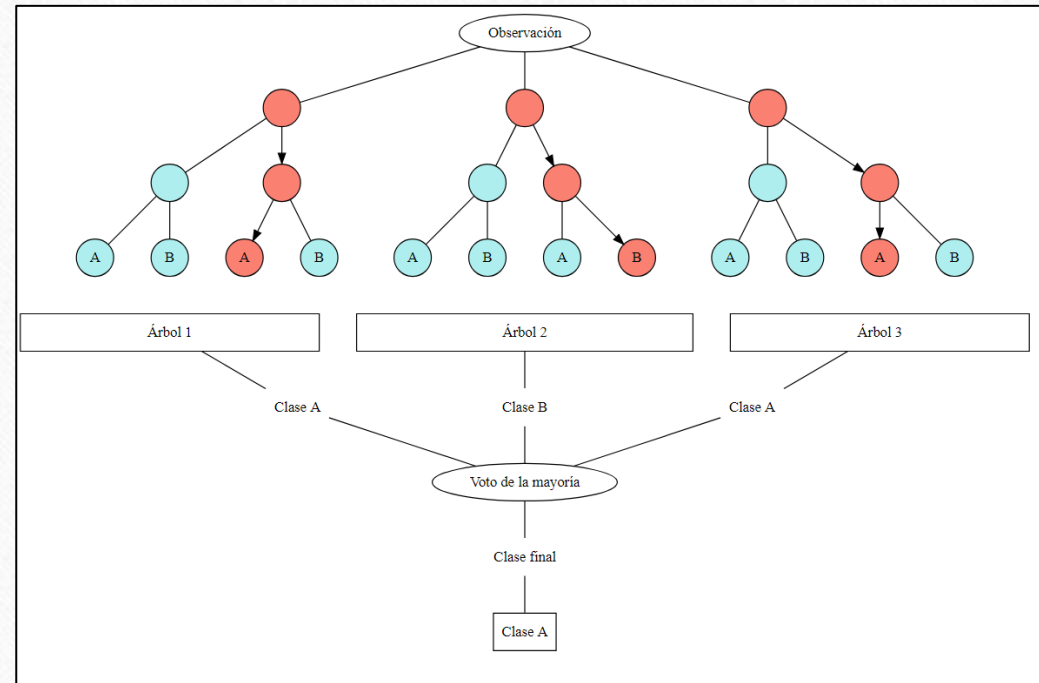
$$G = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk})$$

- ✓ Rápido, flexible, robusto y fácil de interpretar.
- ✗ Peor desempeño comparado a otras técnicas.

Aprendizaje automático

- Mejora la técnica anterior al combinar los resultados de múltiples árboles.
- ✓ Muy buen desempeño en clasificación.
- ✗ Más complejo y lento de entrenar.
- ✗ Más difícil de interpretar.
 - Se recurre al índice de Gini.

Bosques aleatorios



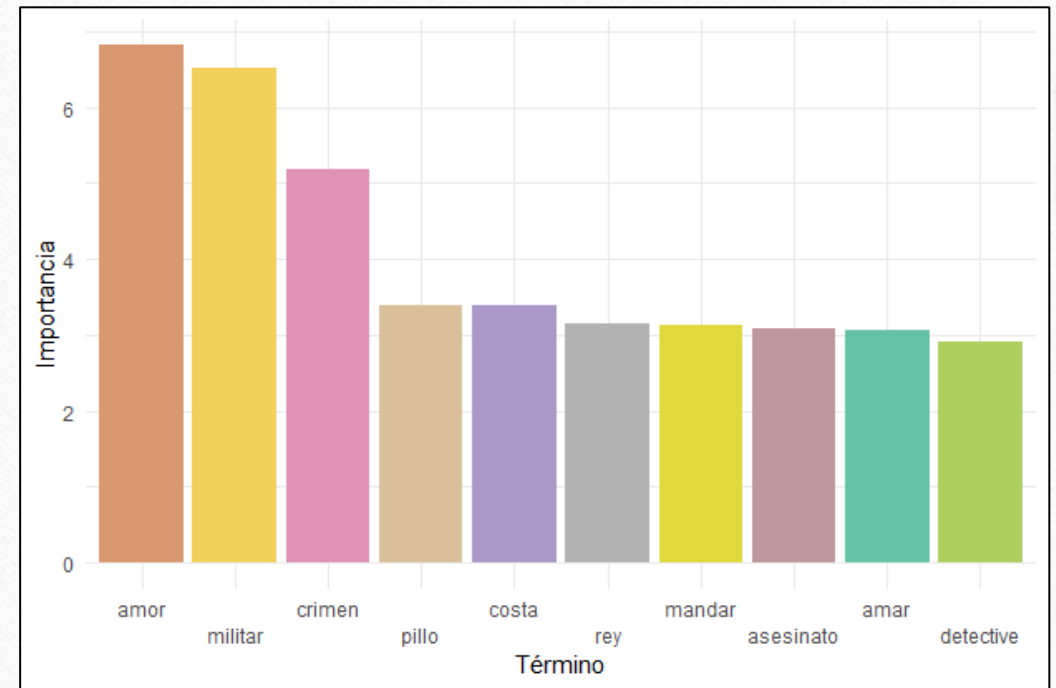
Resultados

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%

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

Resultados

Clase observada	Clase predicha						
	aventura	ciencia ficción	fantástico	ficción histórica	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%)
ficción histórica	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%)	1 (3%)	3 (10%)	5 (17%)	15 (50%)



Conclusiones

- Se logró el objetivo de plantear un modelo para clasificar automáticamente textos de ficción.
 - La mayor precisión alcanzada fue 64%.
- Corpus conformado por 746 obras de ficción pertenecientes a 7 géneros.
- Ocho escenarios de análisis, mejores resultados en general con unigramas, frecuencia de términos y sin filtrar nombres propios.
 - Los bosques aleatorios presentaron resultados muy superiores a los árboles de decisión.
- Precisión muy distinta según clase, variando entre 26% y 82%.

Conclusiones

- En futuras investigaciones, se podrían aplicar nuevos métodos al mismo conjunto de datos y contrastar los resultados.
 - Ejemplos:
 - Nuevos enfoques: Análisis de sentimientos.
 - Nuevas técnicas: Boosting.

Bibliografía y código



<https://github.com/SGS2000/tesina-bosques-aleatorios>