Gender Gap in Spanish WP

Aprendizaje Automático I

Grado en Ciencia e Ingeniería de Datos

Grupo 6: Álvaro Morán Lorente

Sophie Parker

Pablo Pardo Gutiérrez

Universidad
Rey Juan Carlos

Escuela Técnica Superior de
Ingeniería Informática

Índice

1. Introducción
2. Comprensión del negocio/problema y los datos
3. Procesado de datos
4. Modelado
    i. Aprendizaje no supervisado
    ii. Aprendizaje supervisado
5. Evaluación
6. Conclusiones
7. Trabajo a futuro

1. Introducción

- The presentation:
  - Data analysis of the gender of Wikipedia creators
- What we will cover:
  - How we processed the data.
  - What information could be obtained from the data processing
  - How we modeled it.
  - The Machine learning techniques, algorithms, technologies, and plots we used: Naive Bayes, kNN, decision trees, XGBoost, and Random Forest.
  - The conclusions we drew in the end

2. Comprensión del problema y los datos

Table 4. Editors' gender obtained by combining extracted from MediaWiki API and our content coding for the 4,746 coded profiles.

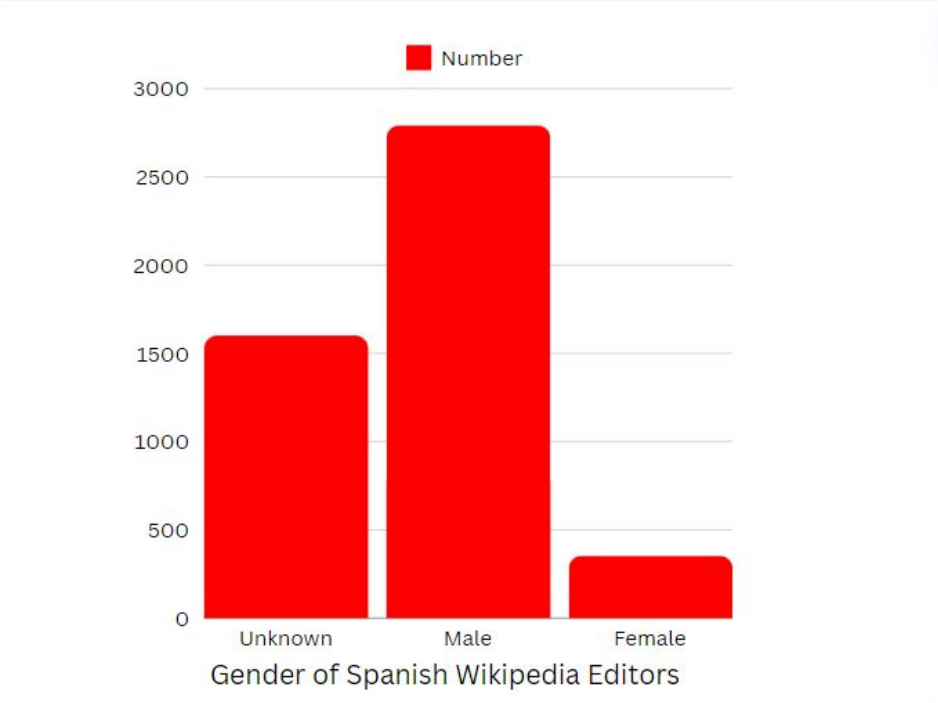
	Gender determined by content coding			
Gender determined by MediaWiki API	**Unknown**	**Men**	**Women**	**Total (API)**
Unspecified	1,601	**1,131**	***172***	2,904
Men	**763**	**898**	0	1,661
Women	***58***	0	***123***	181
Total (manual)	2,422	2,029	295	4,746

https://doi.org/10.1371/journal.pone.0246702.t004

- Wikipedia generates half a million page views daily
- The data we used analyzes the editing practices and gender of Spanish wikipedia editors
- 4746 Wikipedia writers analyzed
- The gender variable was made from two variables
  - C_man: gender extracted from content coding, C_api: gender extracted from WikiMedia API
- NPages: number of pages edited, NCreated: number of pages created

2. Comprensión del problema y los datos

- Variable objetivo: gender
- Gender values:
  - 0 (unknown), 1 (male), and 2 (female)
    - 1601 unknown
    - 2792 male
    - 353 female
- ~8:1 M-F ratio according to the variable “gender”
- This can cause issues
  - Misrepresentation of women
  - Underrepresentation of women



3. Procesado de datos

gender	Target	Categorical	Gender	0 (unknown), 1 (male), 2 (female)	no
C_api	Target	Categorical	Gender	gender extracted from WikiMedia API, codes as female / male / unknown	no
C_man	Target	Integer	Gender	gender extracted from content coding, coded as 1 (male) / 2 (female) / 3 (unknown)	no
E_NEds	Feature	Integer		I index of stratum IJ (0,1,2,3)	no
E_Bpag	Feature	Integer		J index of stratum IJ (0,1,2,3)	no
firstDay	Feature	Date		first edition in the Spanish Wikipedia (YYYYMMDDHHMMSS)	no
lastDay	Feature	Date		last edition in the Spanish Wikipedia (YYYYMMDDHHMMSS)	no
NEds	Feature	Integer		total number of editions	no
NDays	Feature	Integer		number of days (lastDay-firstDay+1)	no
NActDays	Feature	Integer		number of days with editions	no

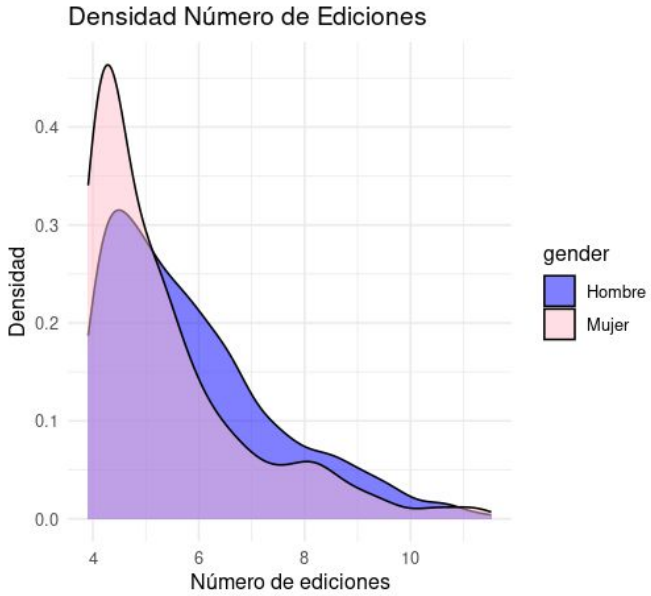
Variable Name	Role	Type	Demographic	Description	Units	Missing Values
NPages	Feature	Integer		number of different pages edited		no
NPcreated	Feature	Integer		number of pages created		no
pagesWomen	Feature	Integer		number of edits in pages related to women		no
wikiprojWomen	Feature	Integer		number of edits in WikiProjects related to women		no
ns_user	Feature	Integer		number of edits in namespace user		no
ns_wikipedia	Feature	Integer		number of edits in namespace wikipedia		no
ns_talk	Feature	Integer		number of edits in namespace talk		no
ns_userTalk	Feature	Integer		number of edits in namespace user talk		no
ns_content	Feature	Integer		number of edits in content pages		no
weightIJ	Feature	Continuous		correcting weight for stratum IJ		no
NIJ	Feature	Integer		number of elements in stratum IJ		no

- Hay variables categóricas redundantes que explican lo mismo que otras continuas. Estas primeras son más difíciles de interpretar.

- Hemos fragmentado las variables que indican fechas en otras variables. También hemos tenido que aplicar logaritmos a una gran cantidad de variables porque sus valores eran en la gran mayoría 0 o eran cercanos.

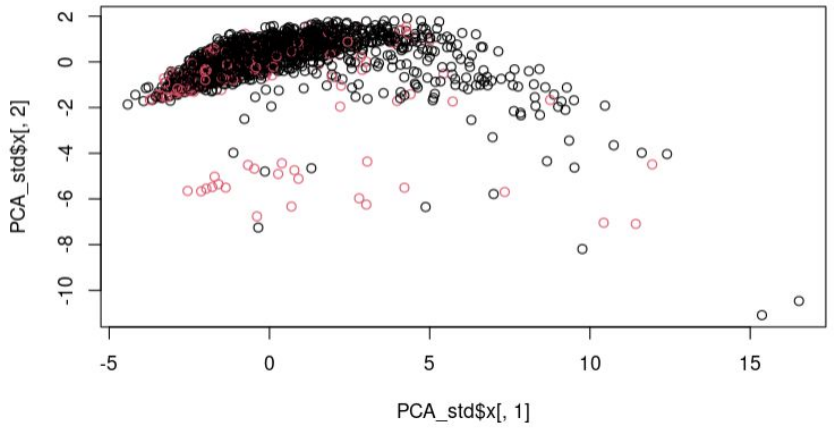
3. Procesado de datos

- Al realizar pruebas de T- student para las variables continuas y chisq para las categóricas obtenemos p-valores bajos en las variables: NDays, wikiprojWomen_bin, logNActdays, logNPages, año_inicial, E_Neds, NIJ, weightIJ, logns_content y logns_talk.
- Variables fundamentales en un futuro: Ndays y wikiprojWomen_bin



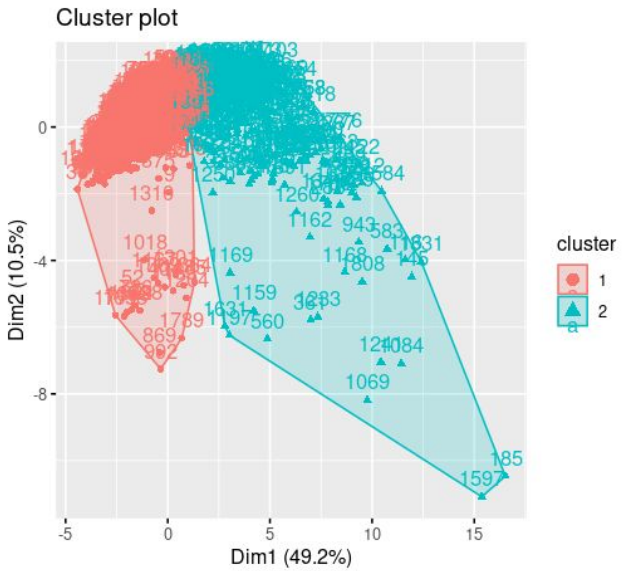
3. Procesado de datos

- En el PCA solo utilizamos variables continuas y transformadas (total de 12 variables)
- Estos gráficos no nos sirven para diferenciar entre grupos pero nos dan pistas sobre qué variables serán relevantes a la hora de hacer los modelos. Nos quedamos con las primeras dos componentes y recogemos un 60% de variabilidad, y en la segunda componente las variables más influyentes son NDays y wikiprojWomen_bin.

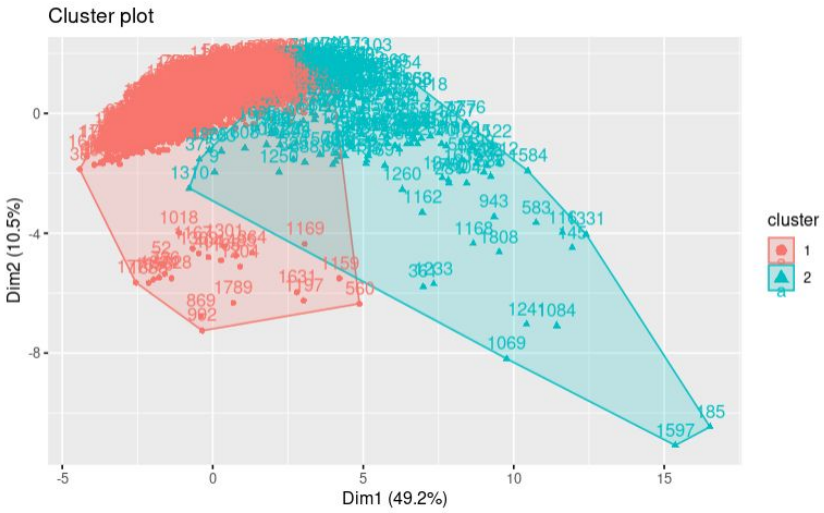


4. Modelado

- Para el método del k-means utilizamos las mismas variables que para el PCA y nos acabamos quedando con 2 clusters. Pero al ver las divisiones con dendrogramas no encontramos explicabilidad porque presenta mucho solapamiento.



K-Means



Cluster Jerárquico

4. Modelado

KNN:

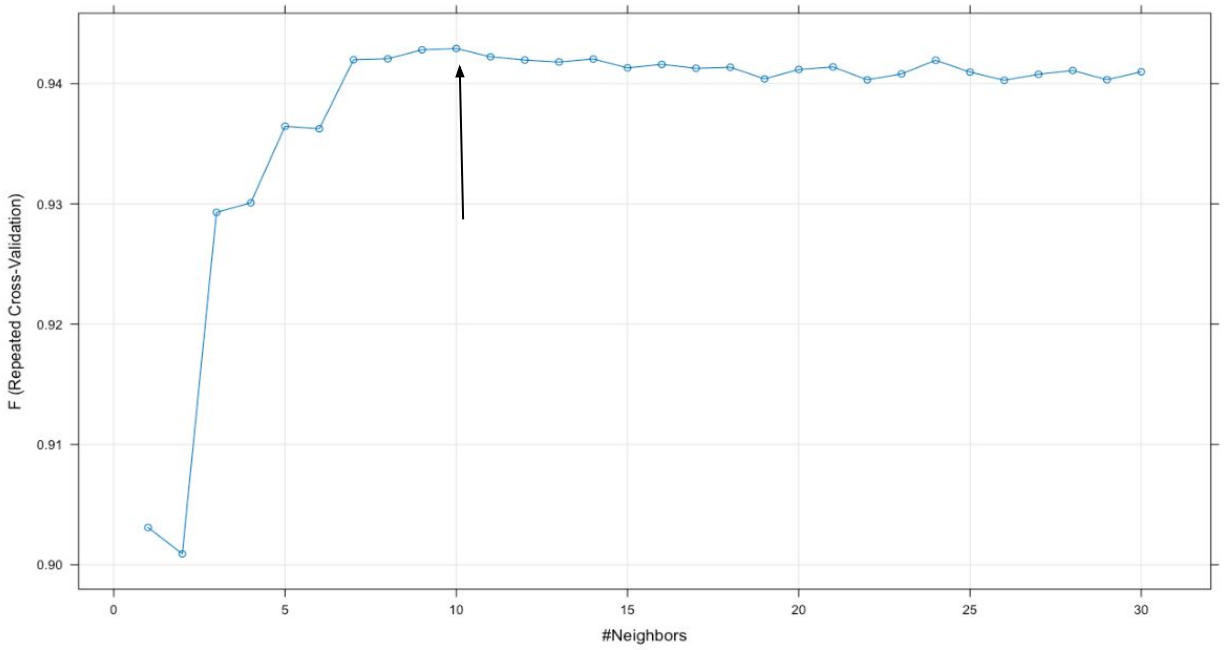
- Mejores k para maximizar F1-Score: k=9 (CV), k=10 (GS)
  - ***F1 : 0.31655 con k=10***

DT:

  - ***F1 : 0.32168***

NAIVE:

  - ***F1 : 0.27848***



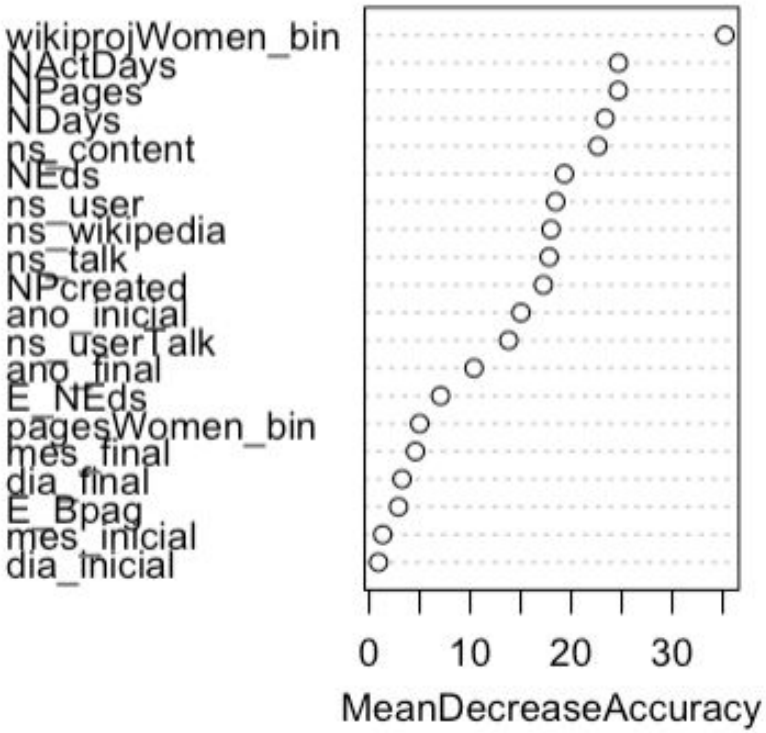
4. Modelado

RF :

500 árboles empleados.

- ***F1 : 0.9755 (train)***
- ***F1 : 0.36257 (test)***

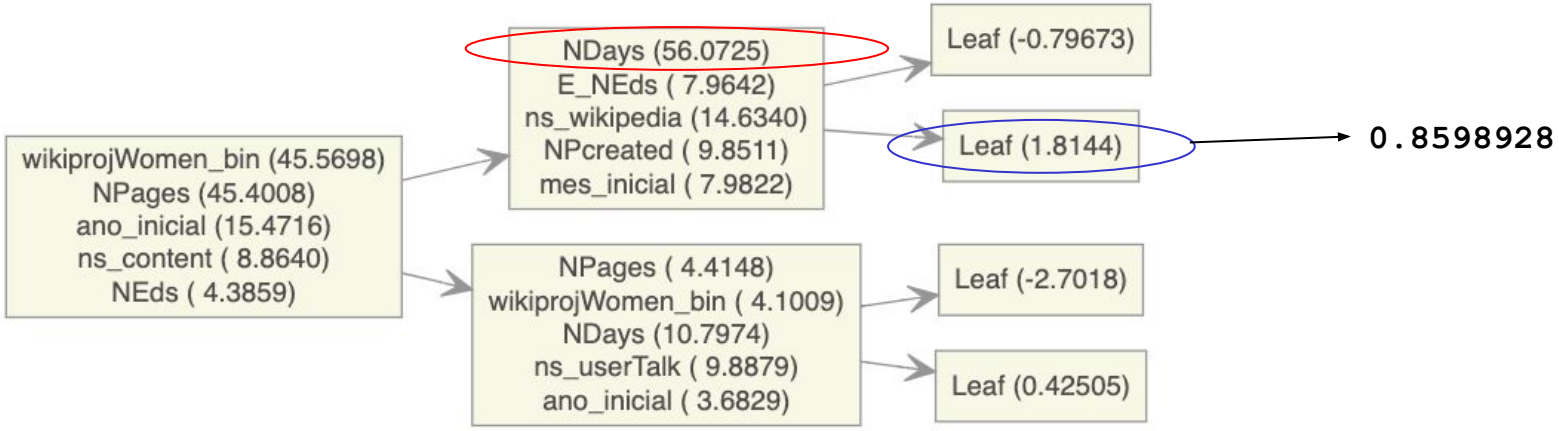
SOBREAJUSTE



4. Modelado

XGBoost

- ***F1 : 0.37037***



5. Evaluación

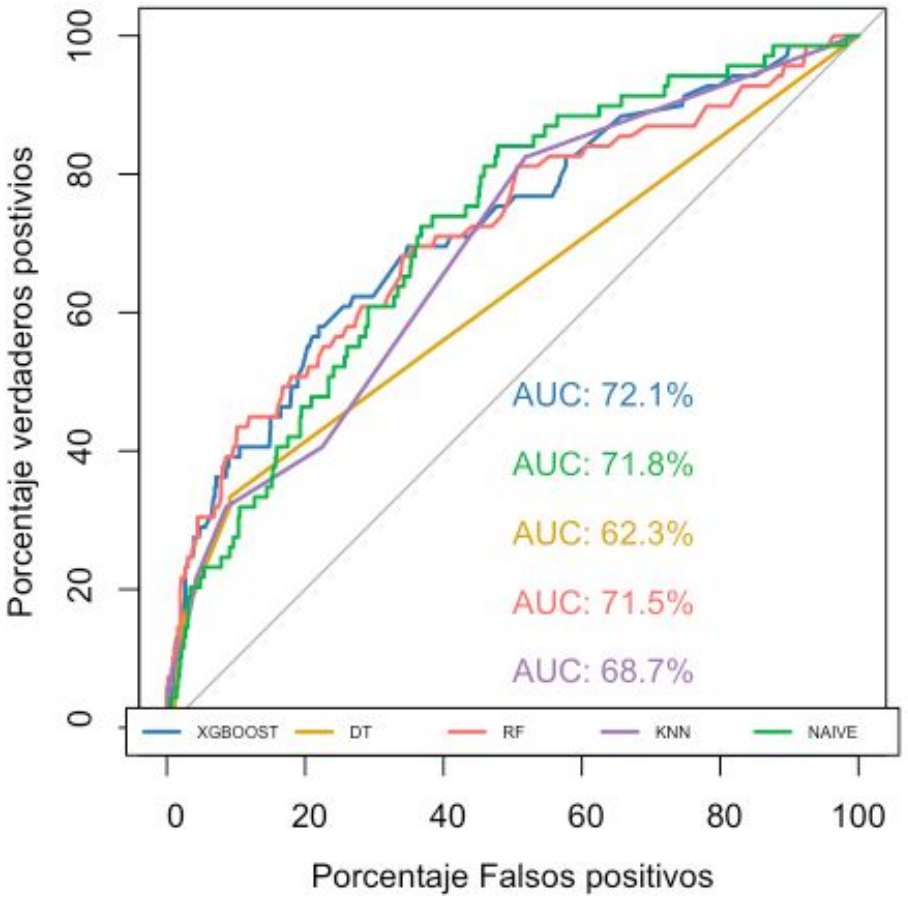
Medida de rendimiento empleada:
F1-Score.

- Modelo con mayor AUC: **XGBoost**.
- Modelo escogido: **Random Forest.**

Naive Bayes: 71,8%

- **Specificity : 0.88129**
- ***F1 : 0.27848***

<u>VALIDACIÓN (RF)</u>
F1 : 0.4085



6. Conclusiones

- Distribuciones no ideales.
- Crear variables a partir de otras puede resultar útil.
- Alta agrupación de datos ➔ Difícil diferenciar grupos.
- Un buen ajuste de hiperparámetros puede ser diferenciador.
- Es tan importante la interpretabilidad de un modelo como su calidad de predicción.

7. Trabajo a futuro

- Potentially compare and contrast with another country's Wikipedia
- Experiment with other partitions
  - 40% 40% 20%
- Try other decision trees with other hyperparameter adjustments
- Run the models with other data agrupations
  - Use categorical variables instead of continuous
  - Number of variables would decrease
  - Potential performance increase of the models
  - The models could be hard to understand or explain
  - Potential performance increase of the models