



# Identificación automática de cuentas bot en proyectos *Open-Source*

Miguel Ángel Fernández Sánchez

Máster en Data Science

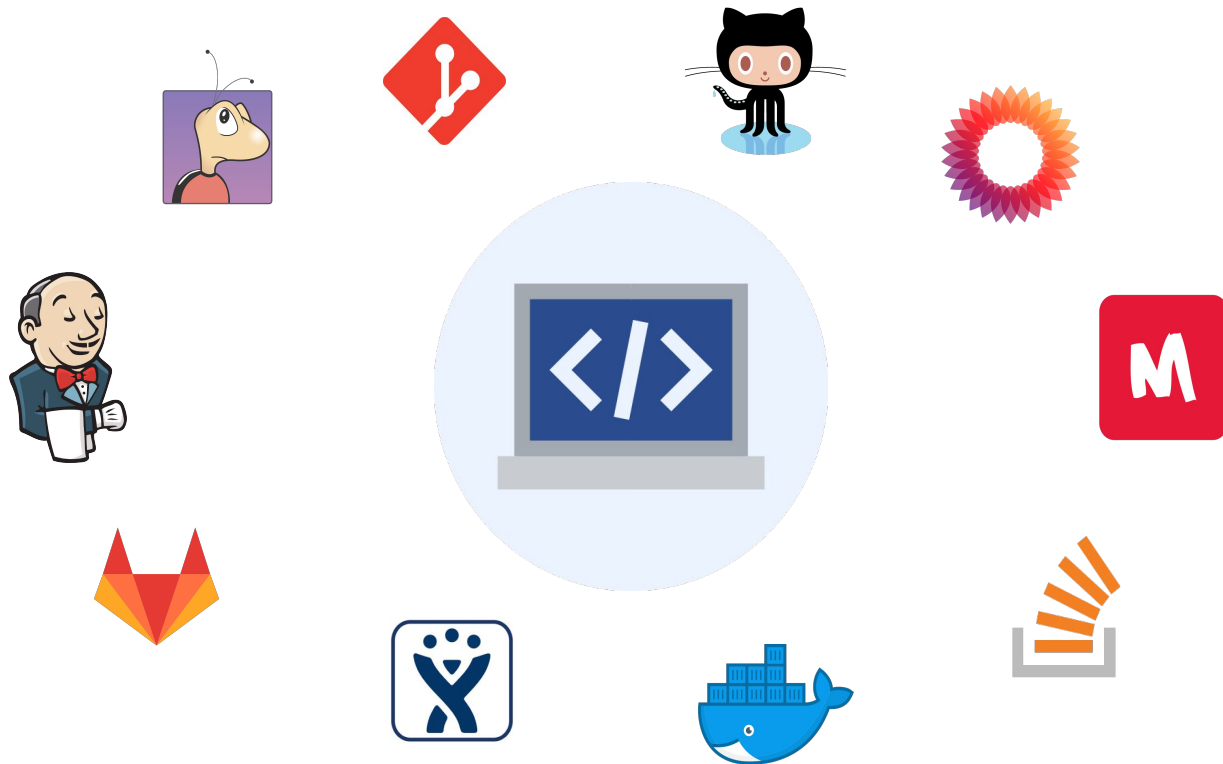
Tutor: Dr. José Felipe Ortega Soto

20 de abril, 2023



Universidad  
Rey Juan Carlos

## /motivación



## /motivación



### Cuentas



jane.doe@urjc.es  
janedoe@libresoft.com

**GitHub**

janedoe



stackoverflow

doe.jane



janedoe

### Organizaciones



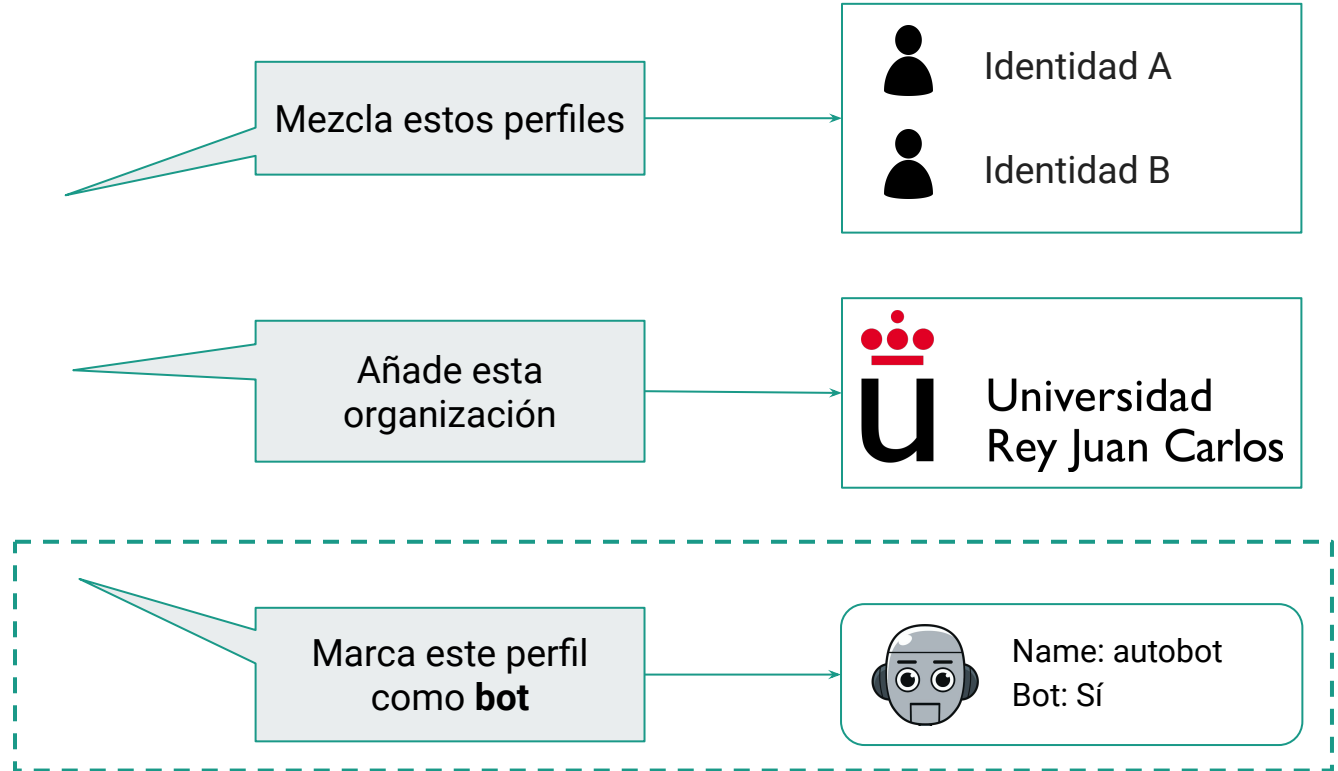
Universidad  
Rey Juan Carlos

**From:** September 1st, 2015  
**To:** January 31st, 2017



**From:** February 1st, 2017  
**To:** Present day

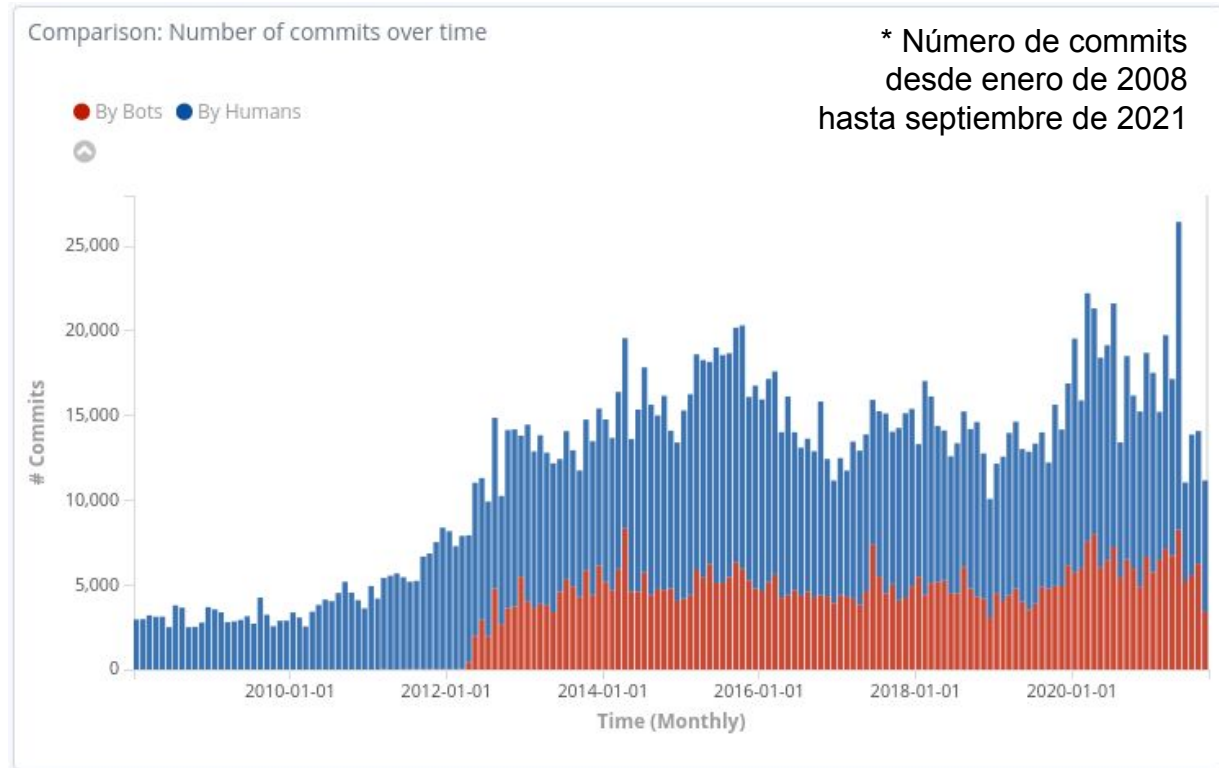
## /motivación



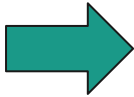
## /problema


● Commits  
de humanos

● Commits  
de bots



/problema



 Hatstall [Profiles](#) [Organizations](#) [About](#)

## Profile info / 221fbe1e6f273891523748ff9261b388edfadc00

Name dependabot[bot]

e-mail None

Bot? ☒

Country None

[Edit](#)

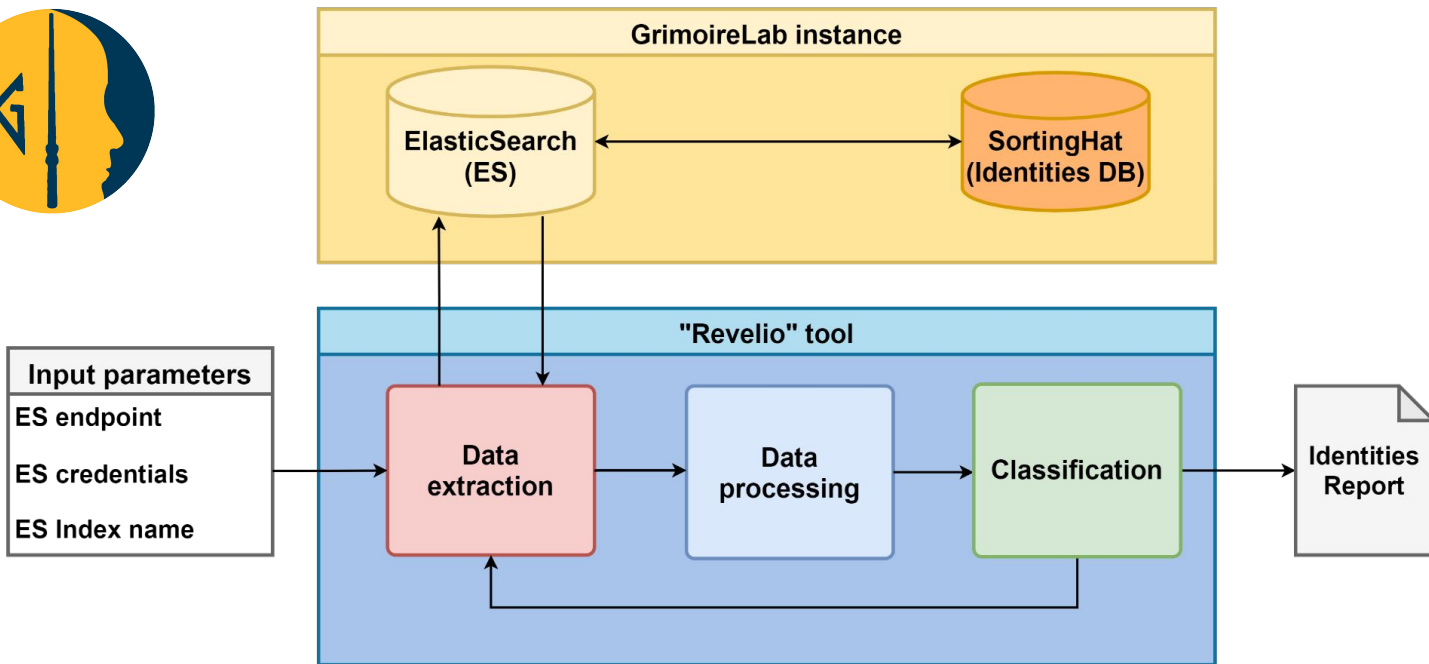
### Enrollments [Add](#)

Organization	Dates	
Github	1900-01-01 to 2100-01-01	<a href="#">Update</a> <a href="#">Un-enroll</a>

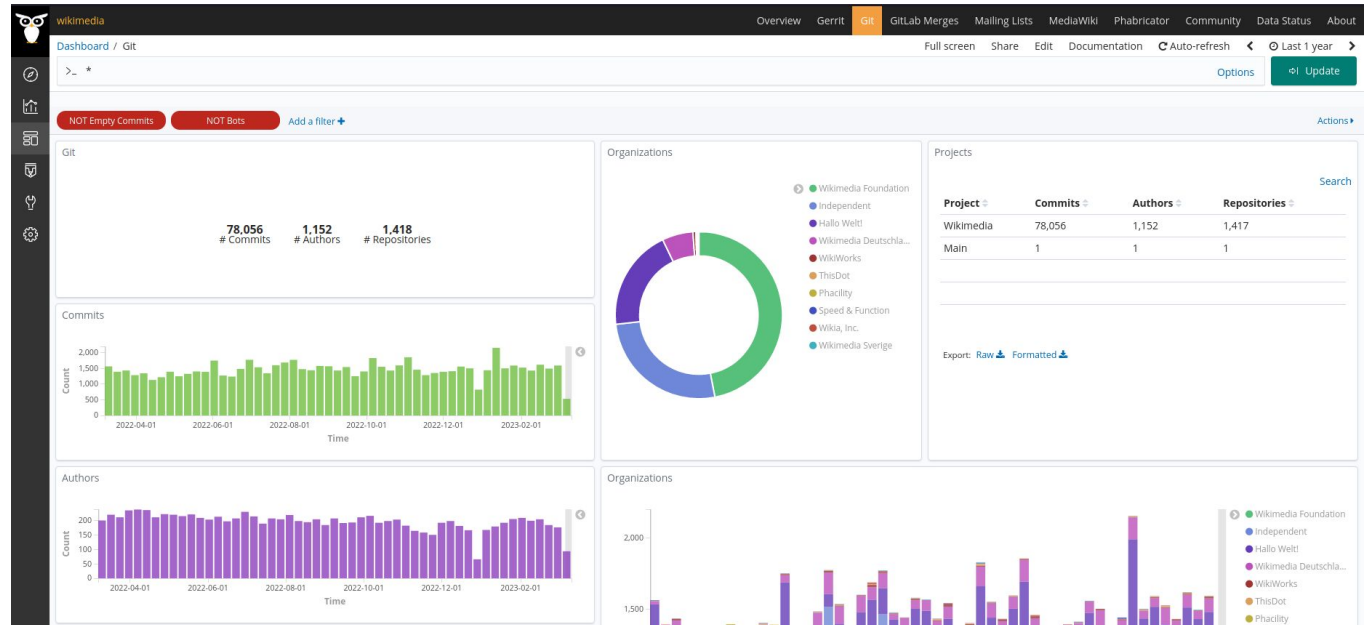
### Profile Identities [Add](#)

Name	email	Username	Source	
None	None	dependabot[bot]	github2	
None	None	dependabot[bot]	github	<a href="#">unmerge</a>
dependabot[bot]	support@github.com	None	git	<a href="#">unmerge</a>
dependabot[bot]	None	None	github2	<a href="#">unmerge</a>
dependabot[bot]	49693333+dependabot[bot]@users.noreply.github.com	None	git	<a href="#">unmerge</a>

/propuesta



## /eligiendo la comunidad a analizar





## /detalles de un commit de Git

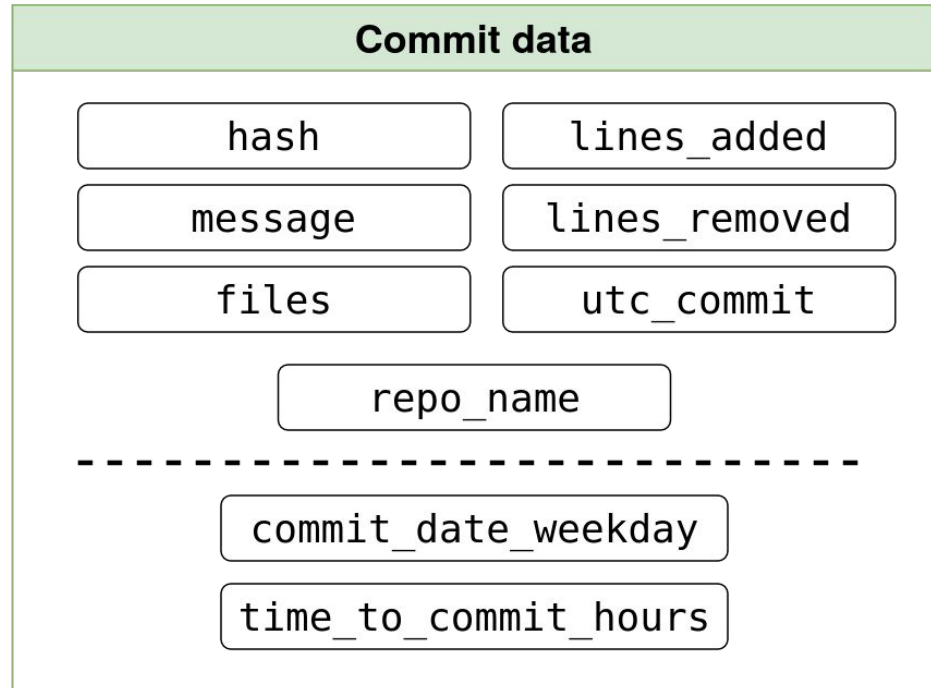
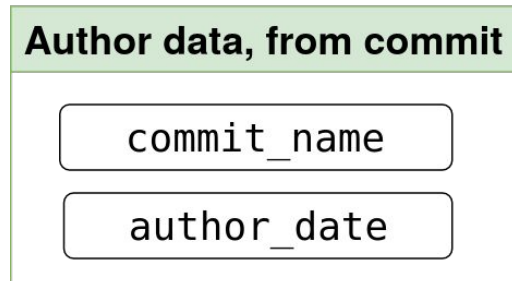
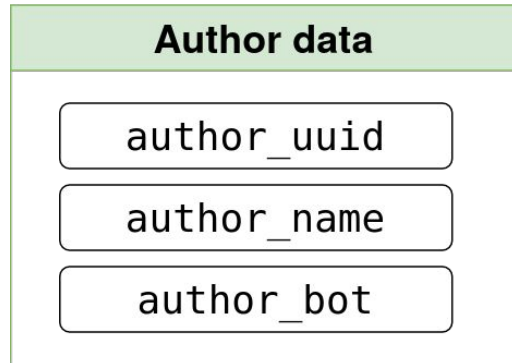
The image shows a screenshot of a Git commit page with several callout boxes explaining different parts of the interface:

- Autor del commit**: Points to the commit author's name, `mafesan`.
- Mensaje del commit**: Points to the commit message, `[schema] Support filtering individuals by last updated date`.
- Identificador del commit**: Points to the commit hash, `66c3b16`.
- Fecha del commit**: Points to the commit date, `Dec 3, 2020`.
- Ficheros modificados**: Points to the list of changed files, `2 changed files with 372 additions and 2 deletions.`
- Líneas añadidas**: Points to the number of lines added, `84`.
- Líneas eliminadas**: Points to the number of lines deleted, `290`.

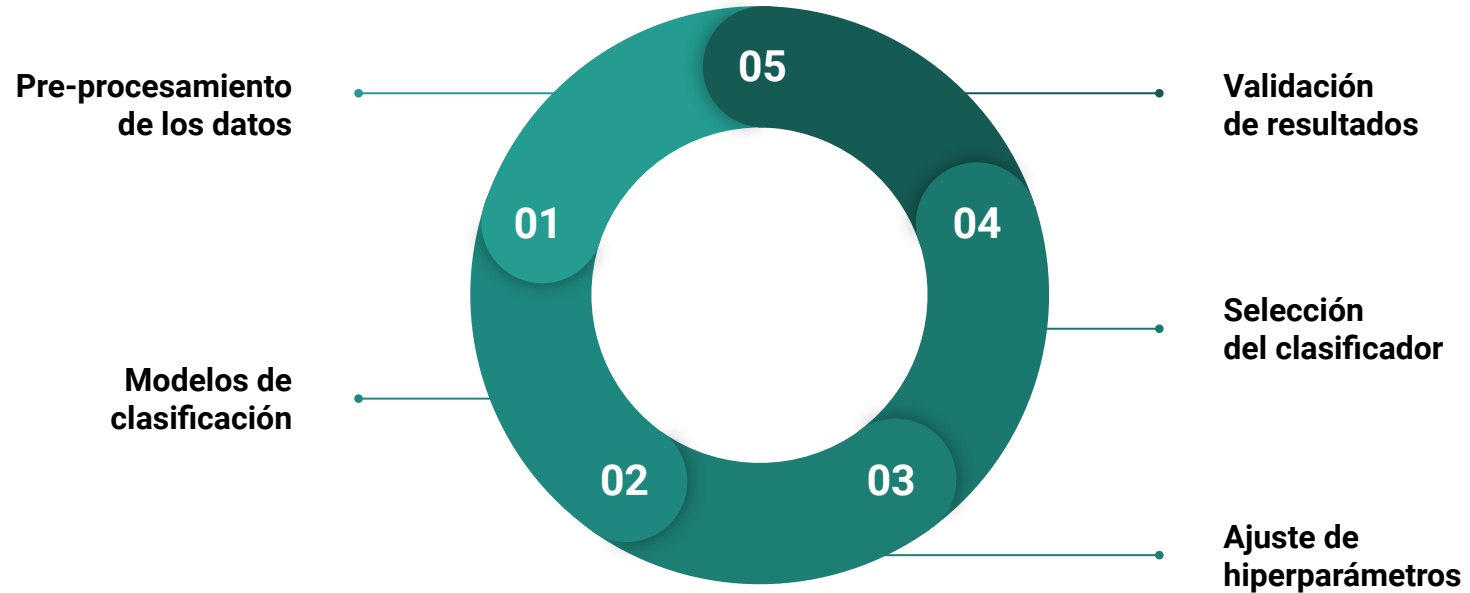
The commit details shown are:

- Commit message: `[schema] Support filtering individuals by last updated date`
- Author: `mafesan`
- Commit hash: `66c3b16`
- Date: `Dec 3, 2020`
- Files changed: `2 changed files with 372 additions and 2 deletions.`
- Files: `sortinghat/core/schema.py` (84 lines added, 290 lines deleted) and `tests/test_schema.py` (290 lines added, 0 lines deleted).

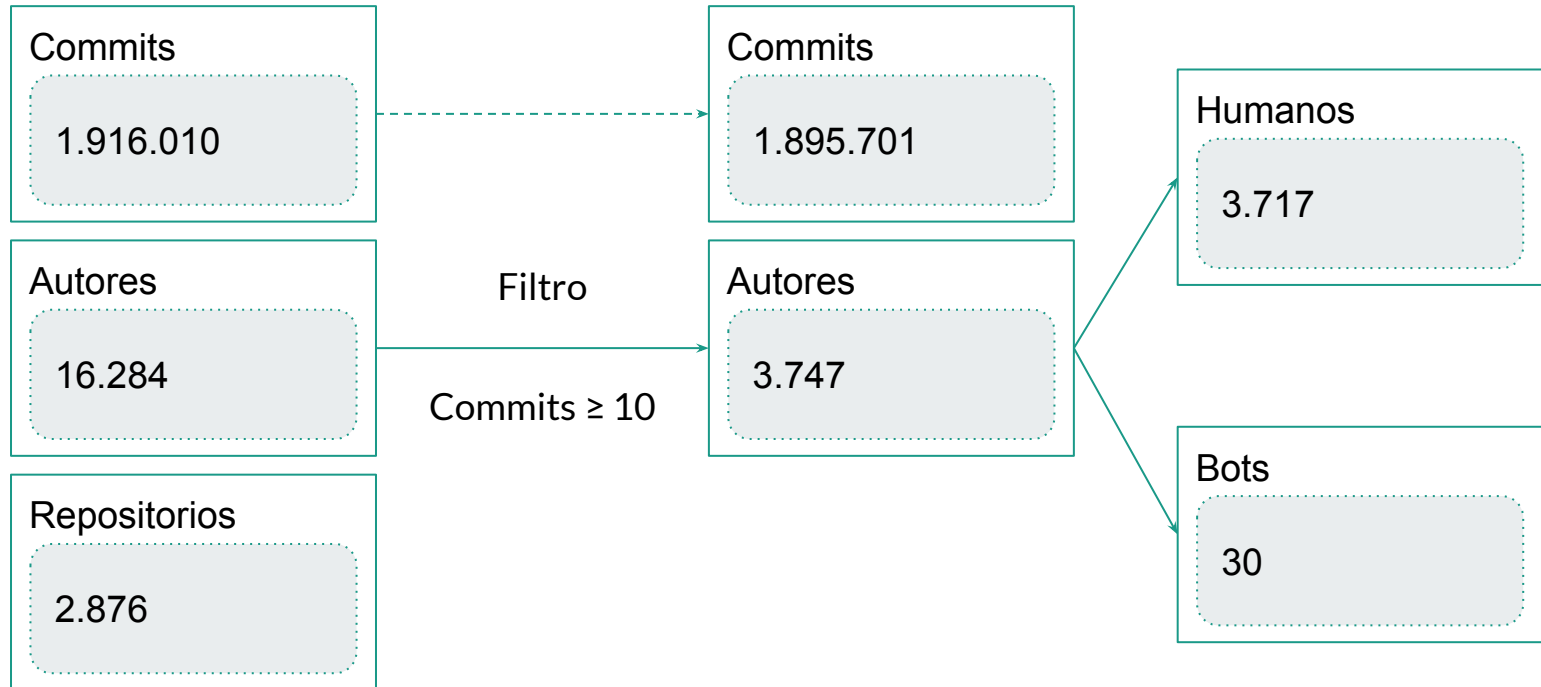
## /selección de campos mediante GQM



## /fase de experimentos



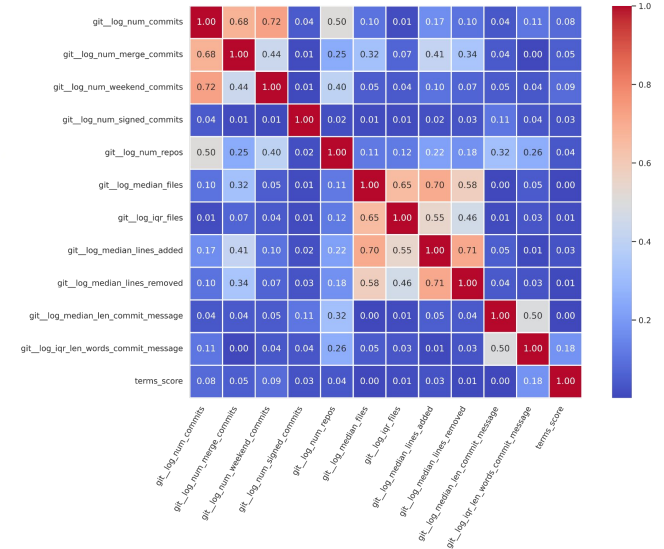
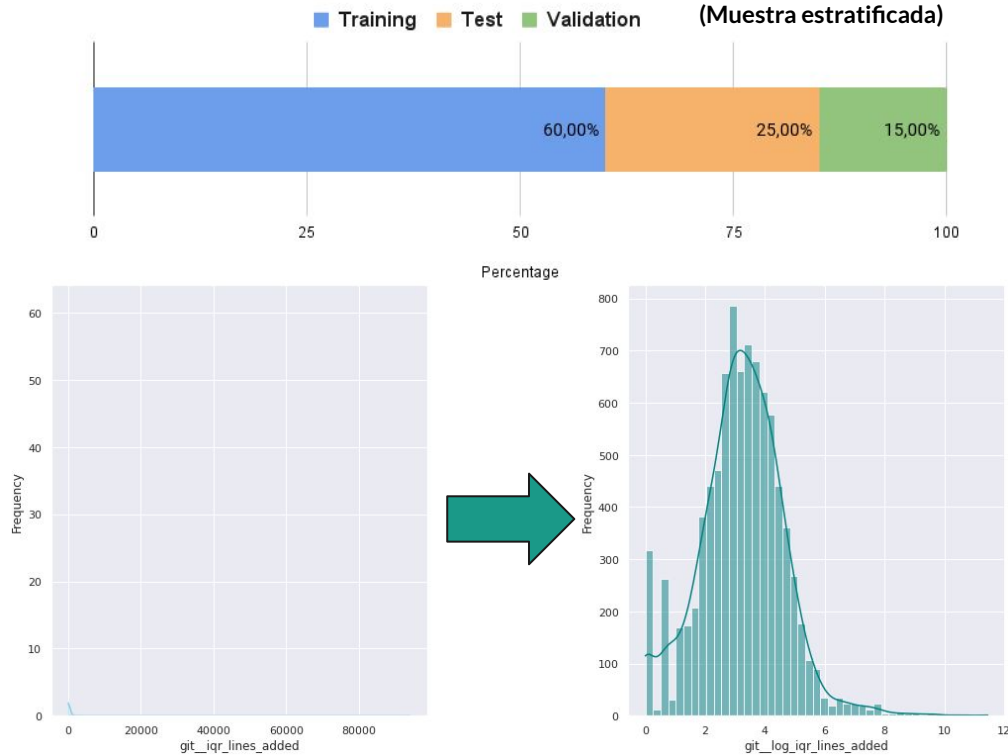
## /análisis exploratorio de datos



## /dataset de entrada



# /pre-procesamiento de los datos y transformación de variables



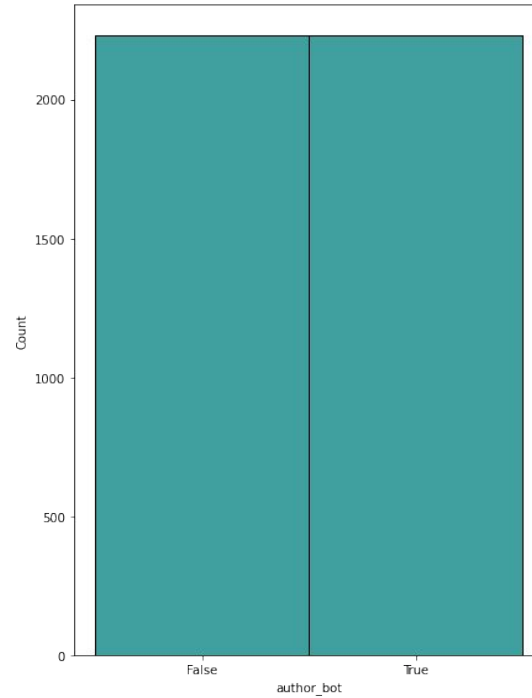
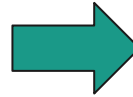
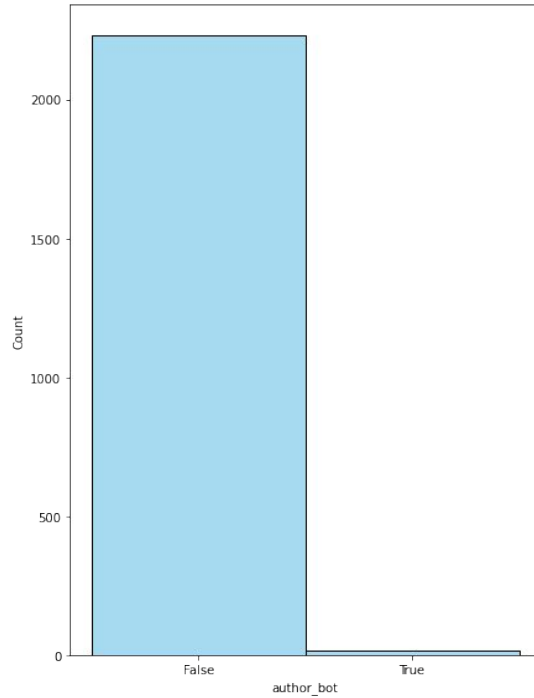
## /generación de variables

Level	Weight	Heuristic terms
1	60	bot, dependency, fix, integration, merge
2	30	auto, build, commit, copy, issue, release, request, review, sync, template, tool, travis
3	10	cd, ci, code, patrol, pr, pull



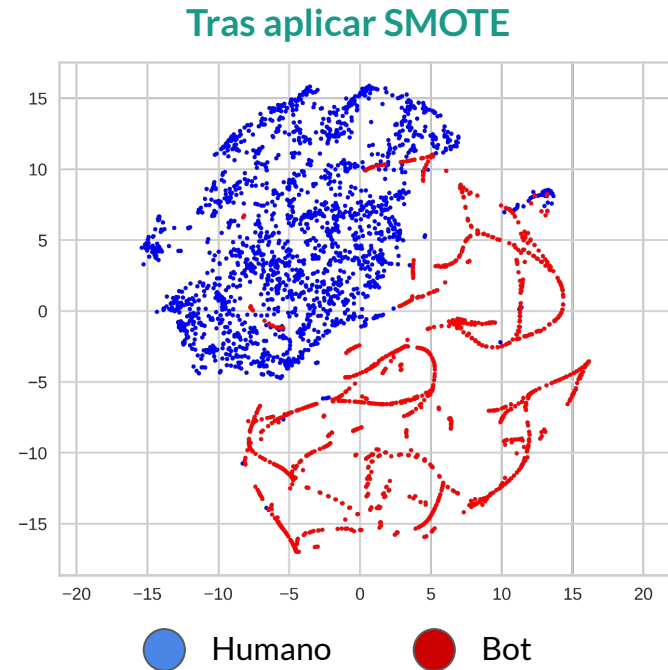
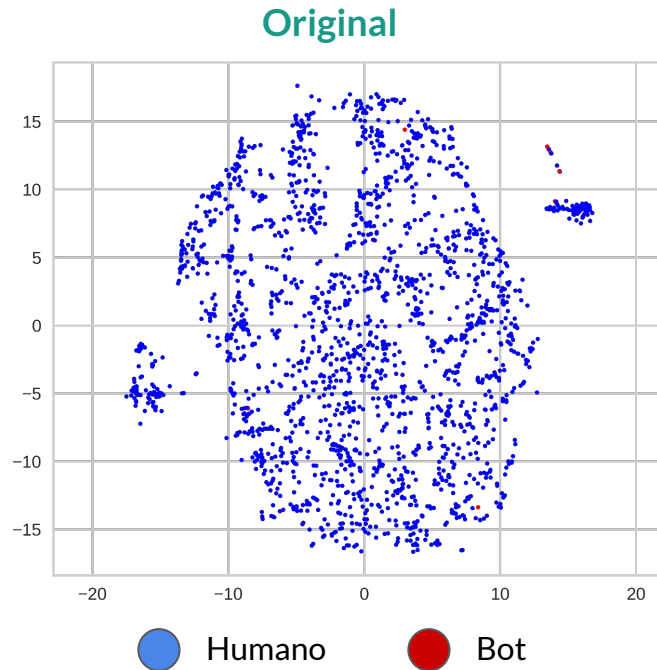
$$Ts(\text{term}) = 60Nl_1 + 30Nl_2 + 10Nl_3$$

## /datos desequilibrados: SMOTE





## /distribución de las clases mediante t-SNE



## /clasificadores: evaluación

Predicted	
Real	TN
	FP
Real	FN
	TP

***Precision***

***Recall***

***$F\beta$ -score***

### Modelo de clasificación

Gaussian Naive-Bayes

Complement Naive-Bayes

LinearSVC

KNN

Decision Tree

Random Forest

XGBoost

# /clasificadores: Random Forest

Número de  
estimadores

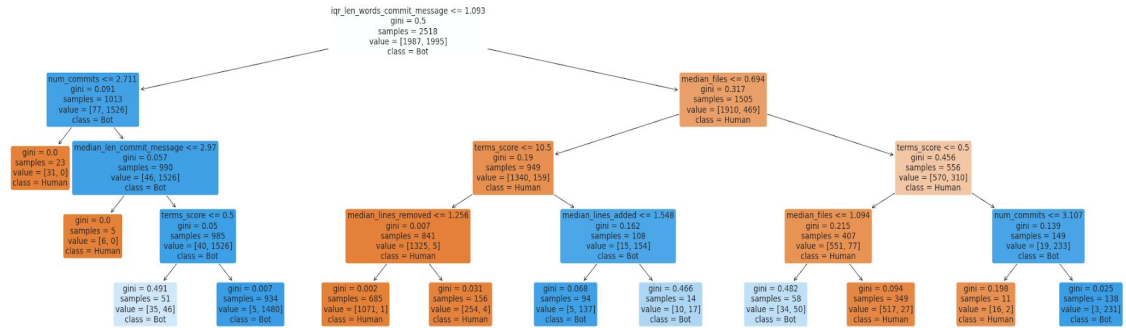
300

Criterio de  
división

Impureza *Gini*

Profundidad  
máxima

4 niveles



Bots



Humanos

## /clasificadores: evaluación

		Predicted	
		Human	Bot
Real	Human	826	3
	Bot	1	6

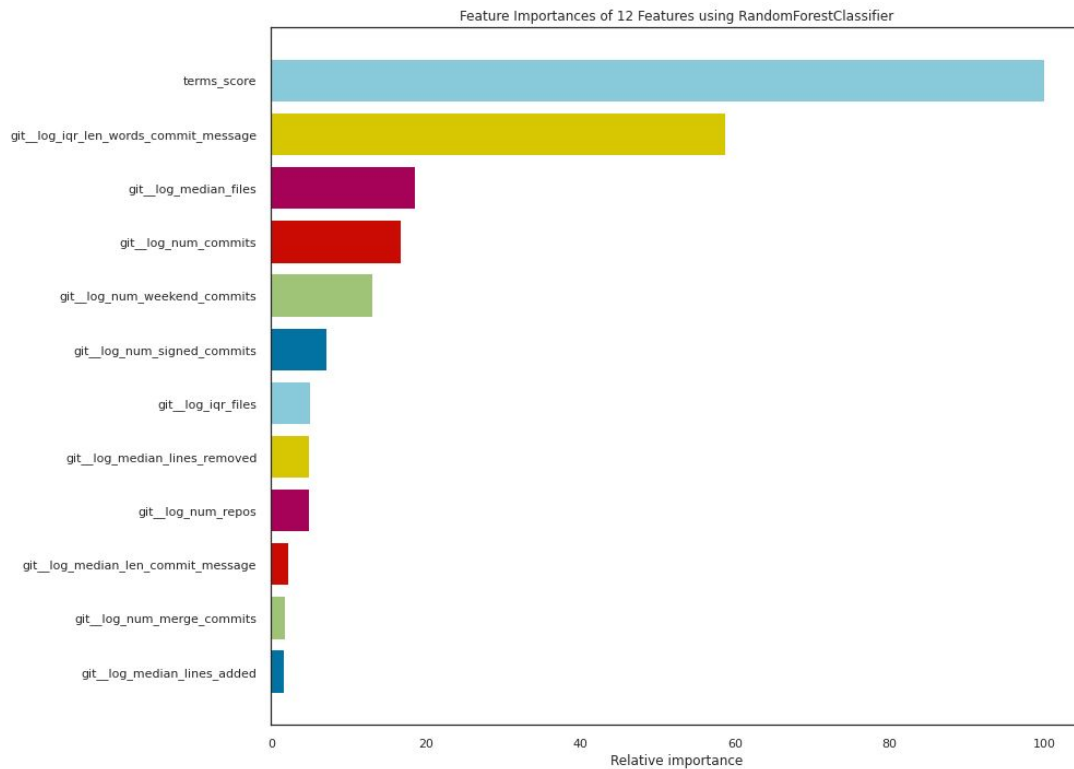
Model name	<i>Precision</i>	<i>Recall</i>	<i>F<math>\beta</math>-score</i>
Random Forest	0.667	0.857	0.811

## /clasificadores: validación

		Predicted	
		Human	Bot
Real	Human	493	6
	Bot	1	3

Model name	<i>Precision</i>	<i>Recall</i>	<i>F<math>\beta</math>-score</i>
Random Forest	0.333	0.75	0.6

## /clasificadores: características más relevantes



## /trabajo futuro

### Mejora y extensión del modelo de clasificación

Otros proyectos / Otras comunidades

Medidas de texto / Huellas digitales

Resumen del usuario / *Concept Learning*

Cuentas mixtas / Clasificación *multiclase*

### Integración con SortingHat/GrimoireLab

Informe de clasificación

Sistema de recomendación

Retroalimentación con el modelo  
tras la respuesta del usuario

/fin



**Turno de preguntas**

**[mafes.github.io/Memoria-TFM](https://mafes.github.io/Memoria-TFM)**