



Kaggle: Predict the pollutant

Rocío Ruiz

Índice

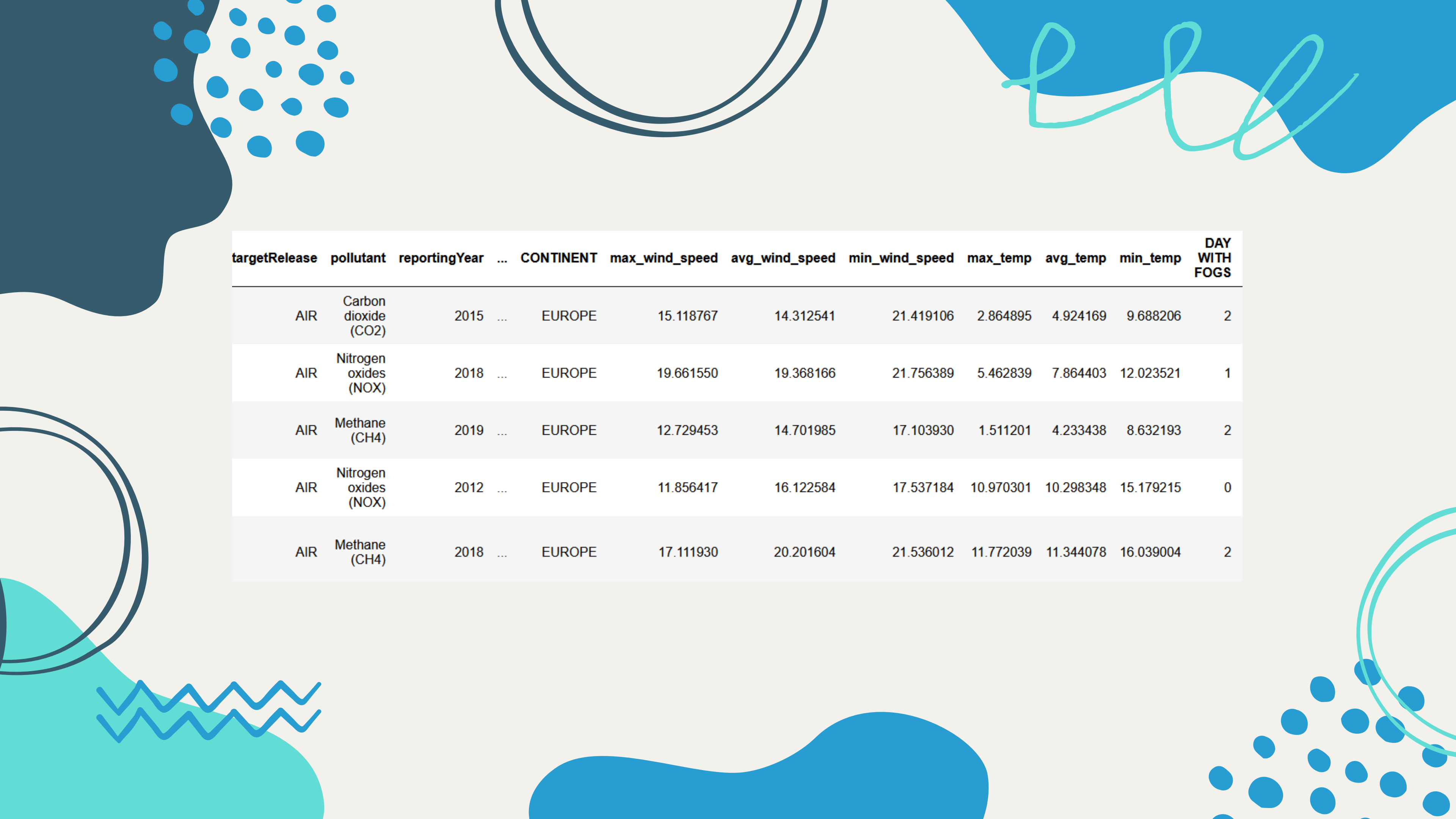
- 1.Dataframe inicial
- 2.Observación relaciones variables
- 3.Arbol de decisión- Feature importances
- 4.SVM lineal
- 5.Pipeline y Grid
- 6.Análisis mejor modelo
- 7.Análisis segundo mejor modelo
- 8.Conclusiones



1. DataFrame original

Unnamed: 0	countryName	eptrSectorName	EPTRAnnexI	MainActivityLabel	FacilityInspireID	facilityName	City
0	0	Germany	Mineral industry	Installations for the production of cement cli...	https://registry.gdi-de.org/id/de.ni.mu/062217...	Holcim (Deutschland) GmbH Werk Höver	Sehnde
1	1	Italy	Mineral industry	Installations for the production of cement cli...	IT.CAED/240602021.FACILITY	Stabilimento di Tavernola Bergamasca	TAVERNOLA BERGAMASCA
2	2	Spain	Waste and wastewater management	Landfills (excluding landfills of inert waste ...	ES.CAED/001966000.FACILITY	COMPLEJO MEDIOAMBIENTAL DE ZURITA	PUERTO DEL ROSARIO
3	3	Czechia	Energy sector	Thermal power stations and other combustion in...	CZ.MZP.U422/CZ34736841.FACILITY	Elektrárny Prunéřov	Kadaň
4	4	Finland	Waste and wastewater management	Urban waste-water treatment plants	http://paikkatiedot.fi/so/1002031/pf/Productio...	TAMPEREEN VESI LIIKELAITOS, VIINIKANLAHDEN JÄT...	Tampere

5 rows × 22 columns



targetRelease	pollutant	reportingYear	...	CONTINENT	max_wind_speed	avg_wind_speed	min_wind_speed	max_temp	avg_temp	min_temp	DAY WITH FOGS
AIR	Carbon dioxide (CO2)	2015	...	EUROPE	15.118767	14.312541	21.419106	2.864895	4.924169	9.688206	2
AIR	Nitrogen oxides (NOX)	2018	...	EUROPE	19.661550	19.368166	21.756389	5.462839	7.864403	12.023521	1
AIR	Methane (CH4)	2019	...	EUROPE	12.729453	14.701985	17.103930	1.511201	4.233438	8.632193	2
AIR	Nitrogen oxides (NOX)	2012	...	EUROPE	11.856417	16.122584	17.537184	10.970301	10.298348	15.179215	0
AIR	Methane (CH4)	2018	...	EUROPE	17.111930	20.201604	21.536012	11.772039	11.344078	16.039004	2

REPORTER
NAME

CITY ID

Mr. Jacob
Ortega

7cdb5e74adcb2ffaa21c1b61395a984f

Ashlee
Serrano

cd1dbabbdba230b828c657a9b19a8963

Vincent
Kemp

5011e3fa1436d15b34f1287f312fbada

Carol Gray

37a6d7a71c4f7c2469e4f01b70dd90c2

Blake Ford

471fe554e1c62d1b01cc8e4e5076c61a

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20127 entries, 0 to 20126
Data columns (total 22 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Unnamed: 0                               20127 non-null  int64
1   countryName                             20127 non-null  object
2   eprtrSectorName                         20127 non-null  object
3   EPRTRAnnexIMainActivityLabel           20127 non-null  object
4   FacilityInspireID                      20127 non-null  object
5   facilityName                           20127 non-null  object
6   City                                    20127 non-null  object
7   targetRelease                          20127 non-null  object
8   pollutant                              20127 non-null  object
9   reportingYear                          20127 non-null  int64
10  MONTH                                  20127 non-null  int64
11  DAY                                   20127 non-null  int64
12  CONTINENT                             20127 non-null  object
13  max_wind_speed                        20127 non-null  float64
14  avg_wind_speed                       20127 non-null  float64
15  min_wind_speed                       20127 non-null  float64
16  max_temp                             20127 non-null  float64
17  avg_temp                             20127 non-null  float64
18  min_temp                             20127 non-null  float64
19  DAY WITH FOGS                         20127 non-null  int64
20  REPORTER NAME                        20127 non-null  object
21  CITY ID                              20127 non-null  object
dtypes: float64(6), int64(5), object(11)
```

Limpieza de datos:

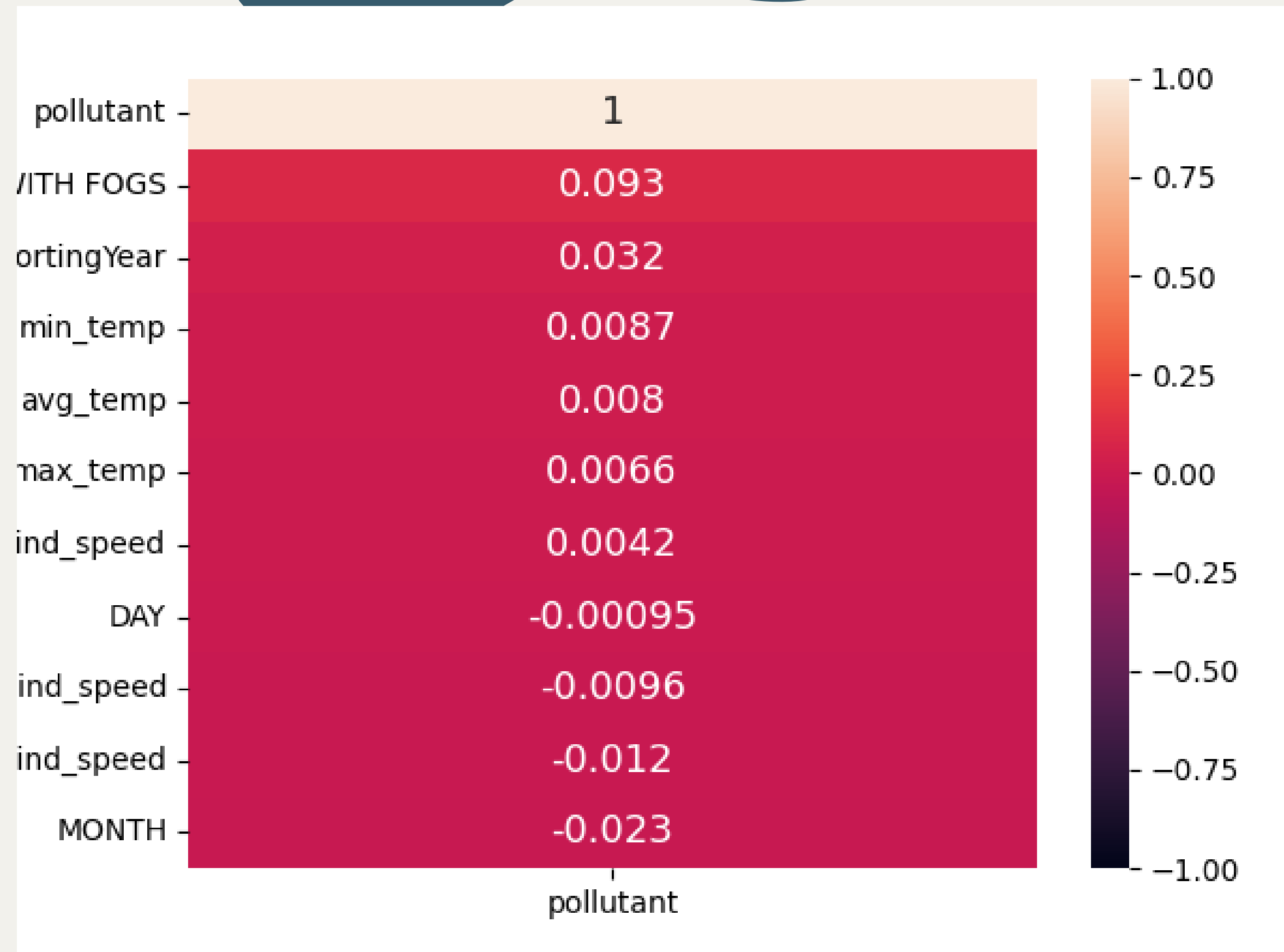
Eliminar columnas:

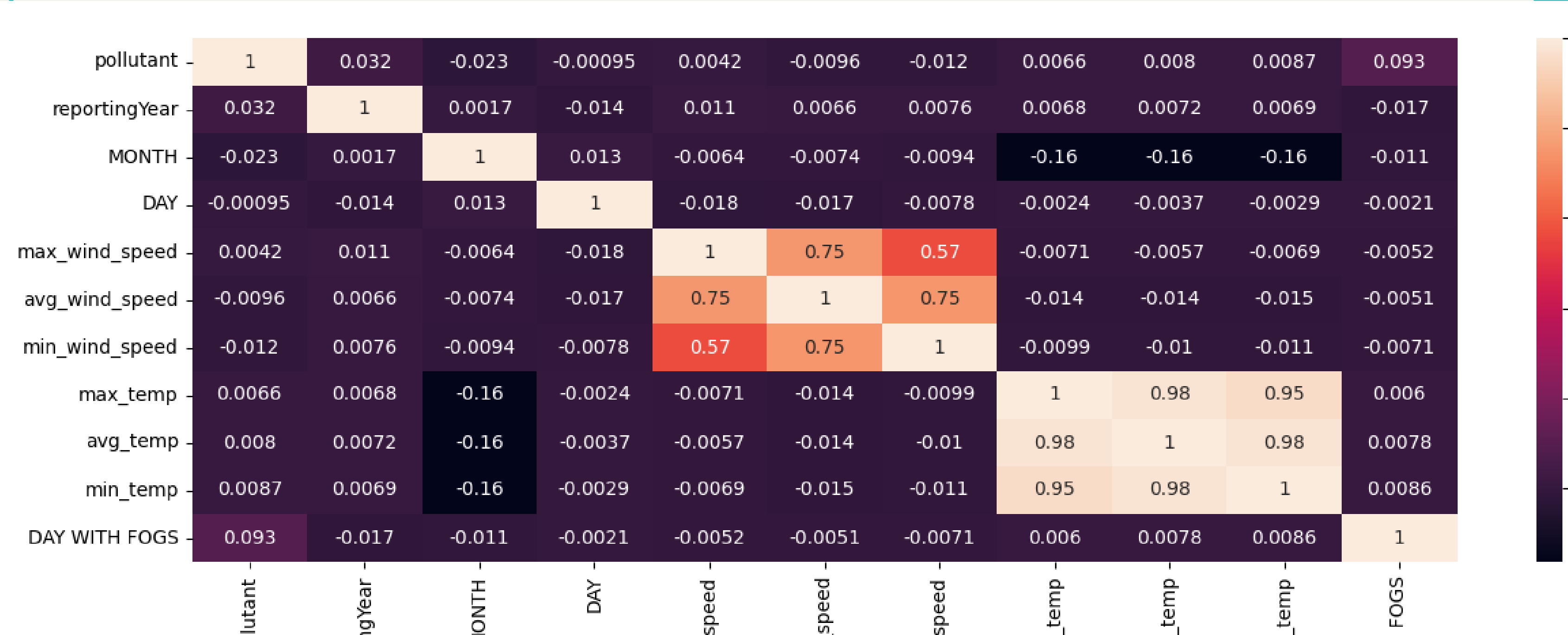
- Unnamed: 0
- TargetRelease
- Continent

Codificar target: 0,1,2



2. Observación relaciones variables



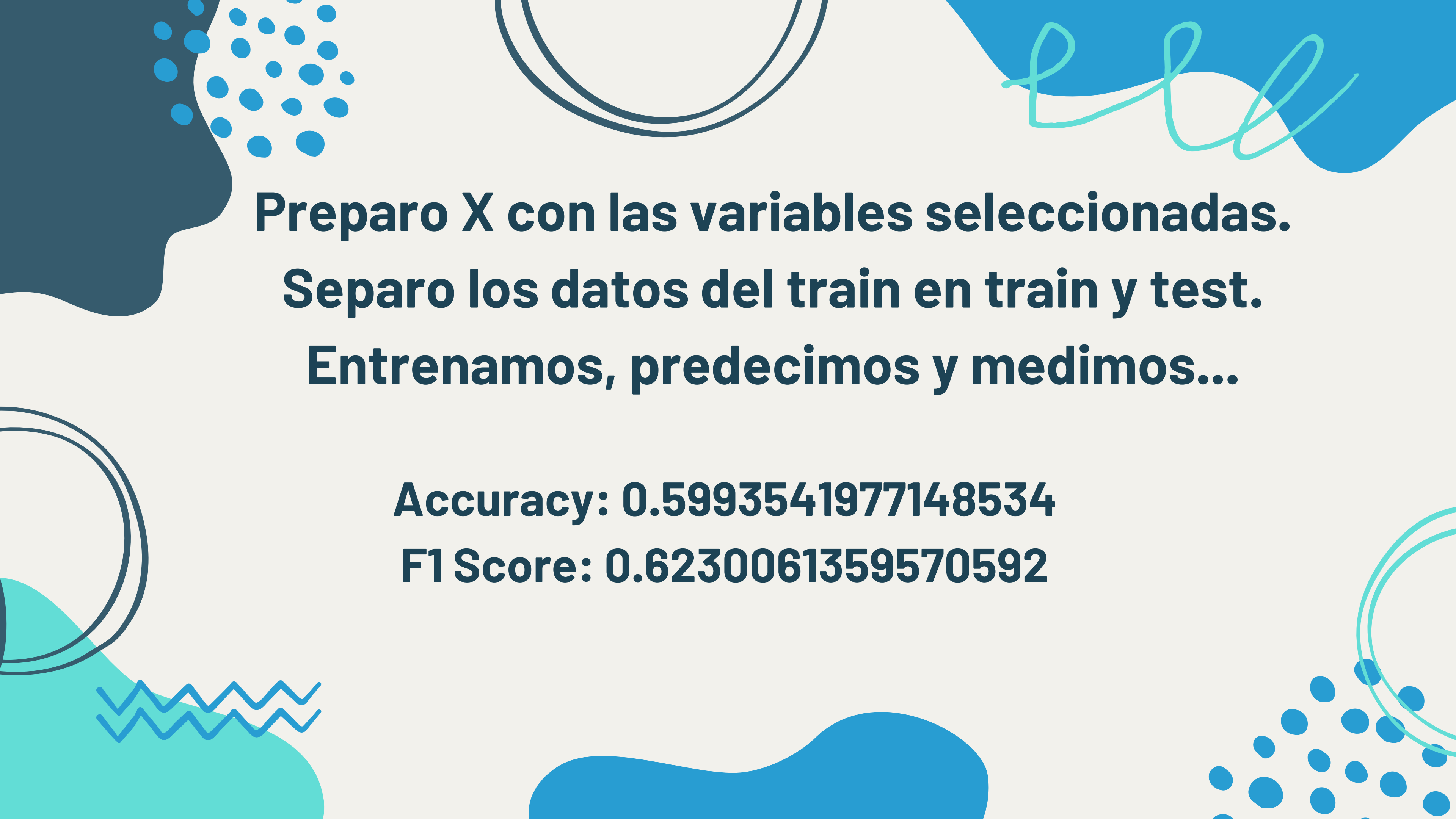




3. Arbol de decisión- Feature importances



4. SVM Lineal



**Preparo X con las variables seleccionadas.
Separo los datos del train en train y test.
Entrenamos, predecimos y medimos...**

**Accuracy: 0.5993541977148534
F1 Score: 0.6230061359570592**



5. Pipeline y Grid



Añadimos:

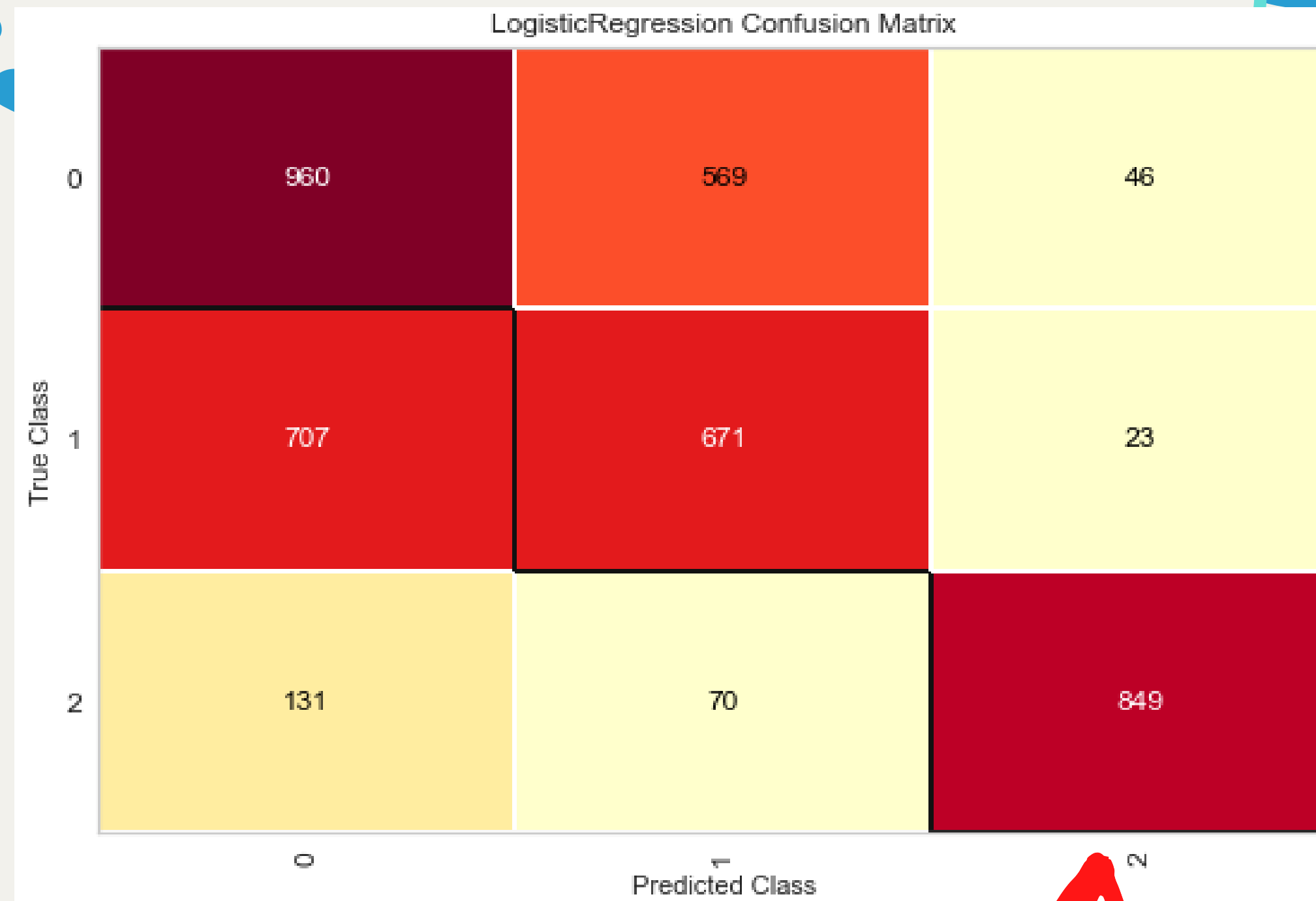
- **Regresión logística**
- **Random Forest**
- **Support Vector Machine (SVM)**

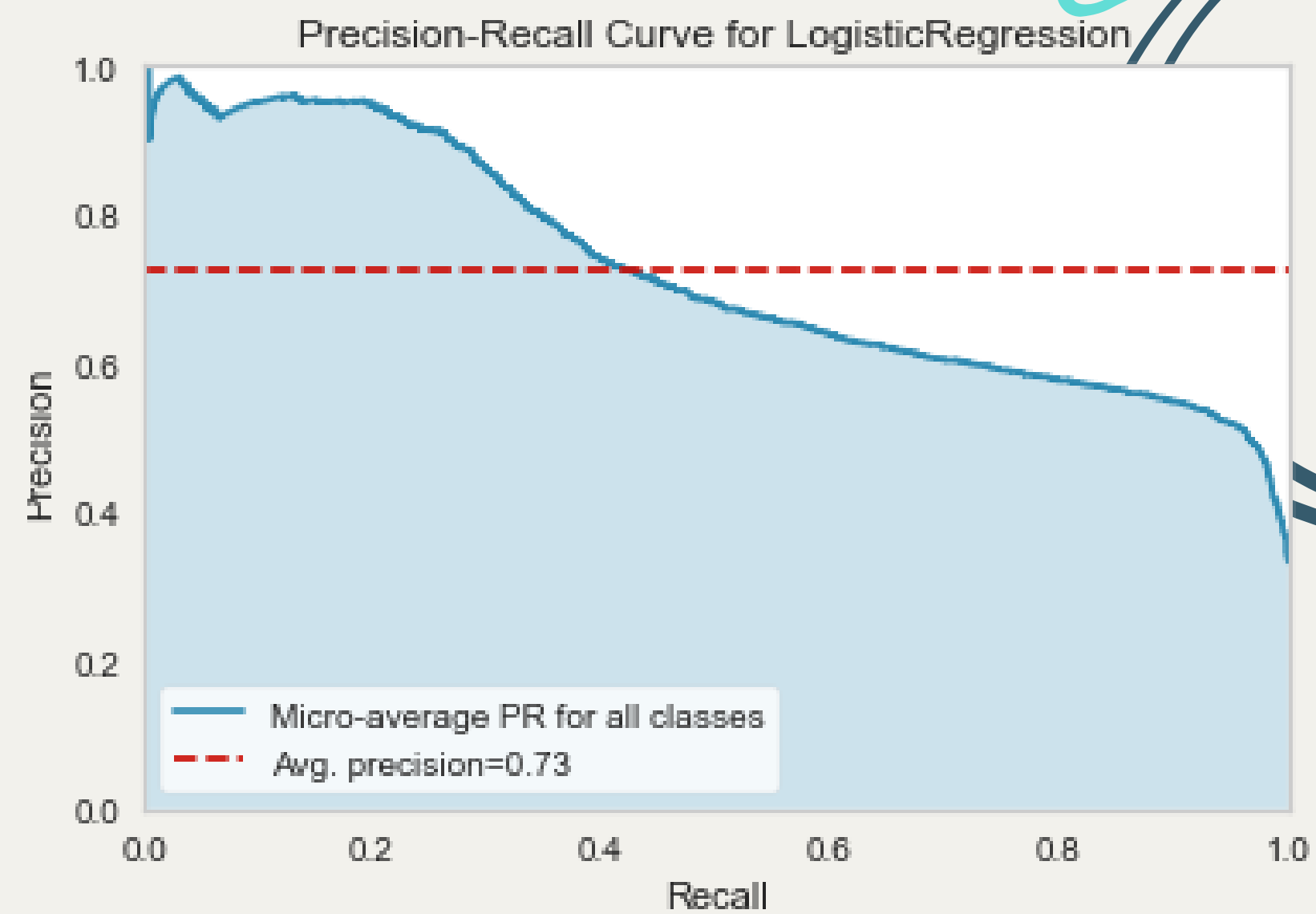
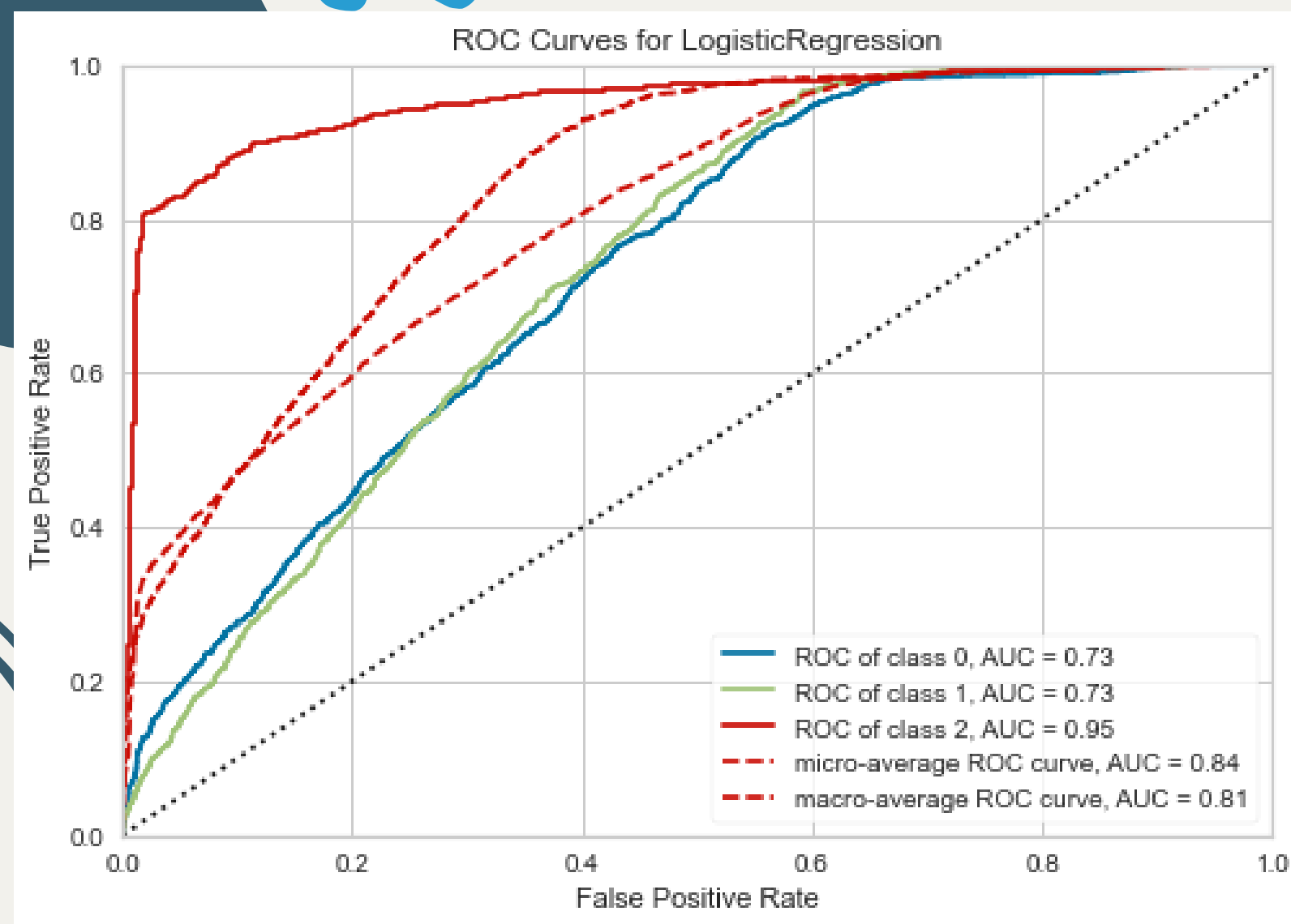
Y el mejor modelo es...

	Grid	Best score
0	gs_reg_log	0.647084
1	gs_rand_forest	0.643148
2	gs_svm	0.547912



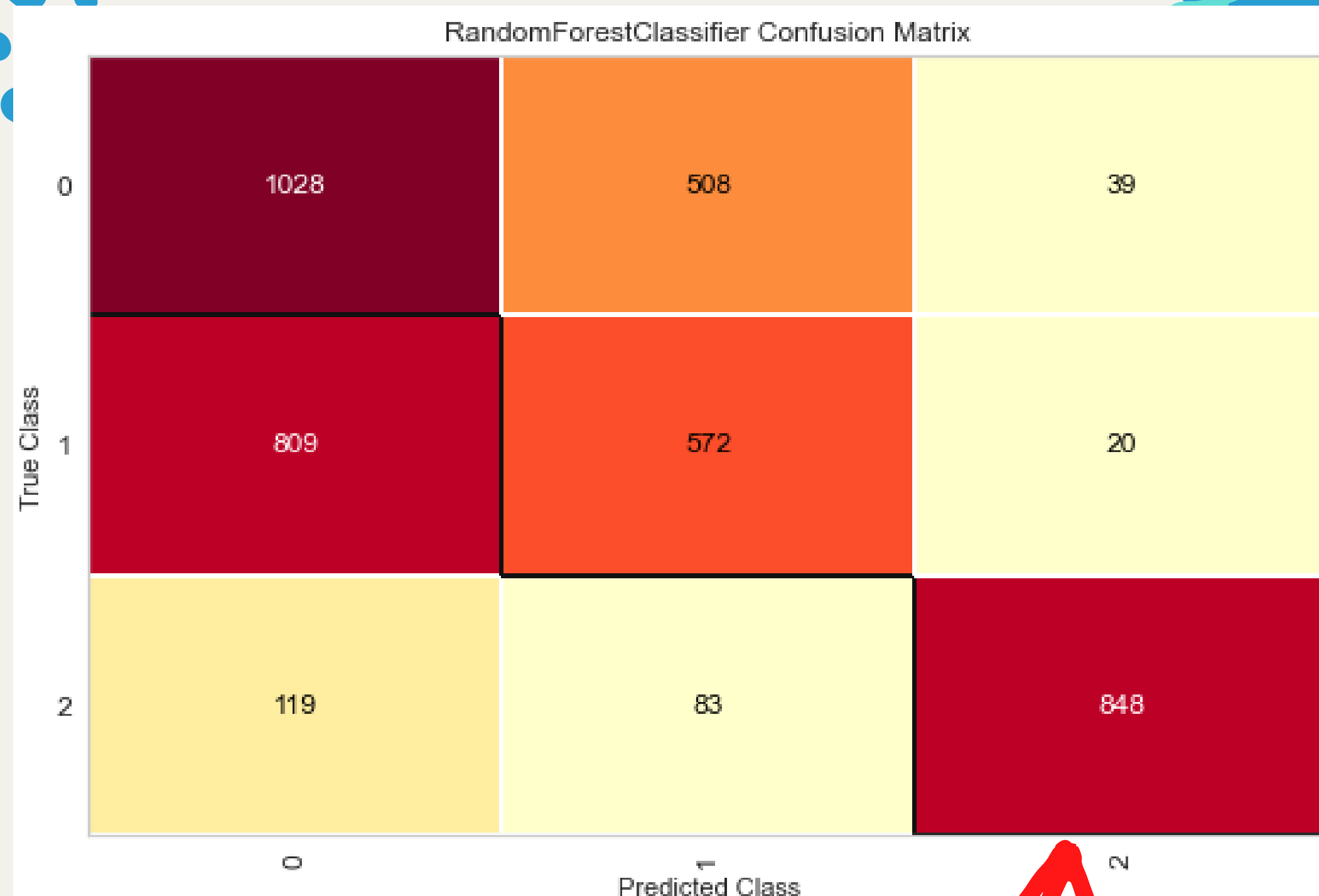
6. Análisis mejor modelo: regresión logística

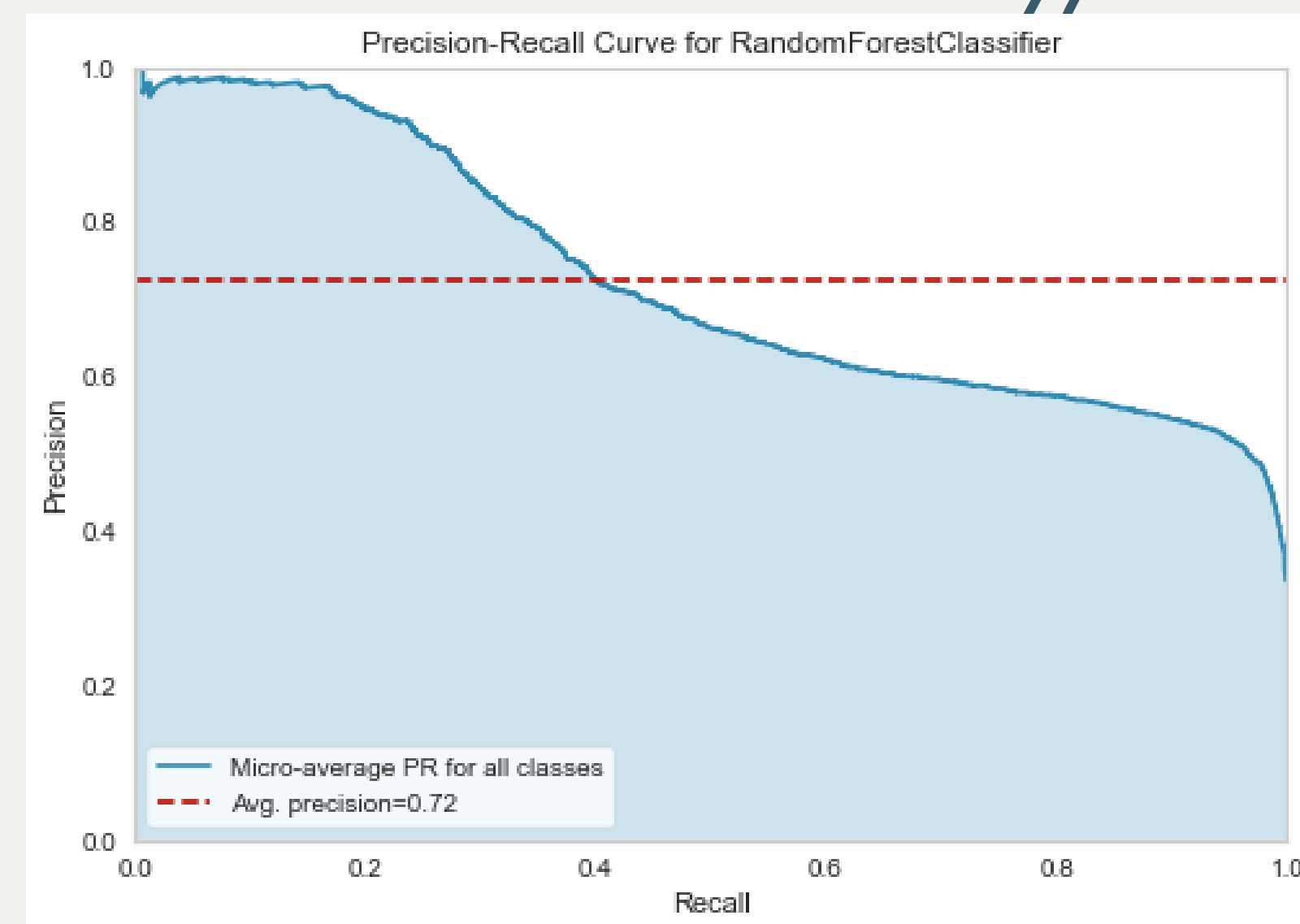
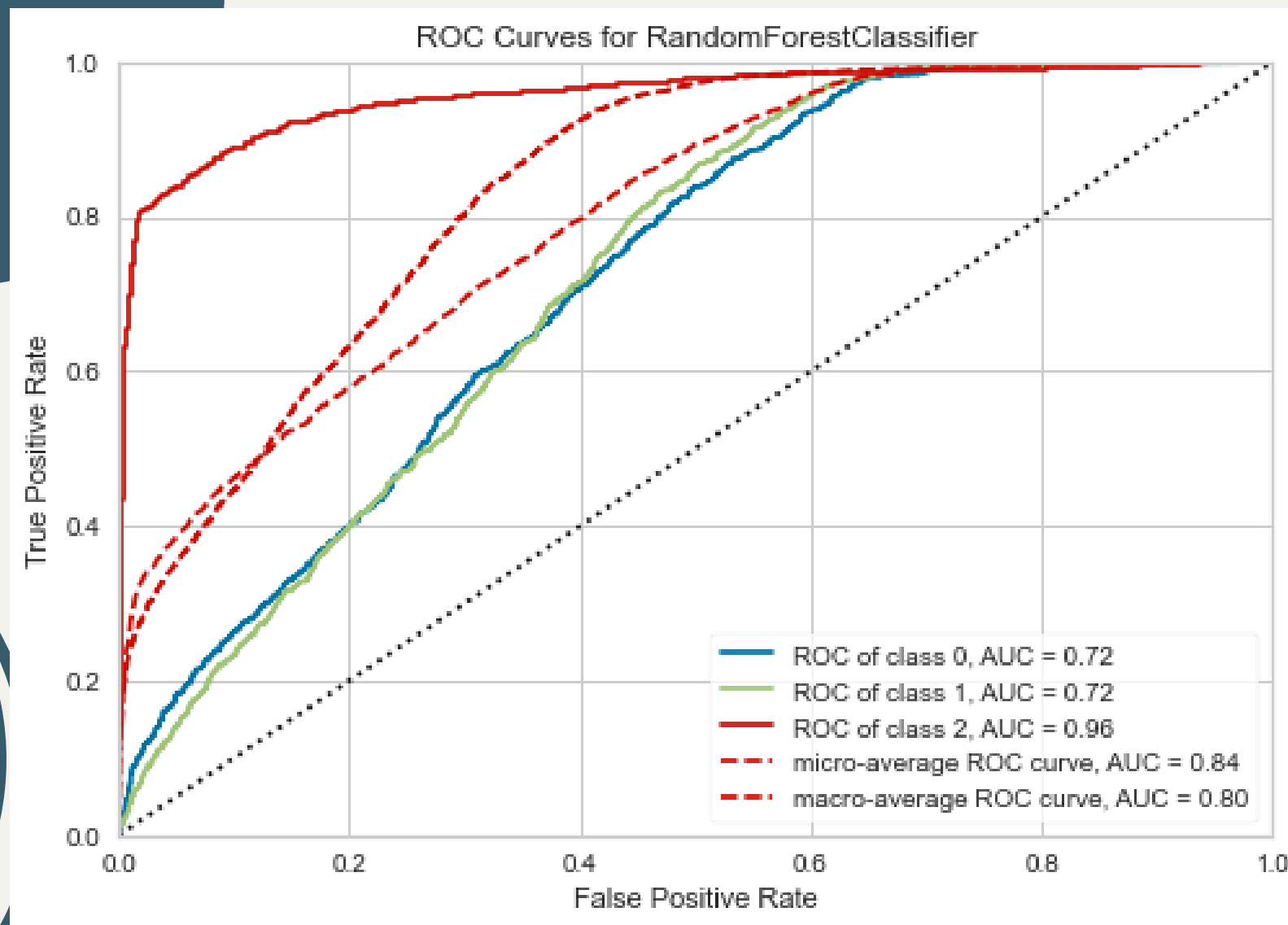






7. Análisis segundo mejor modelo: random forest







8. Conclusiones

- **Modelo random forest predice mejor que regresión logística los contaminates de tipo 2 (NH2)**
- **Modelo random forest obtiene mejor score en Kaggle que regresión logística**



**iMuchas
gracias!**