Kaggle: Predict the pollutant

Rocío Ruiz

Índice

1. DataFrame original

	Unnamed: 0	countryName	eprtrSectorName	EPRTRAnnexIMainActivityLabel	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)
```

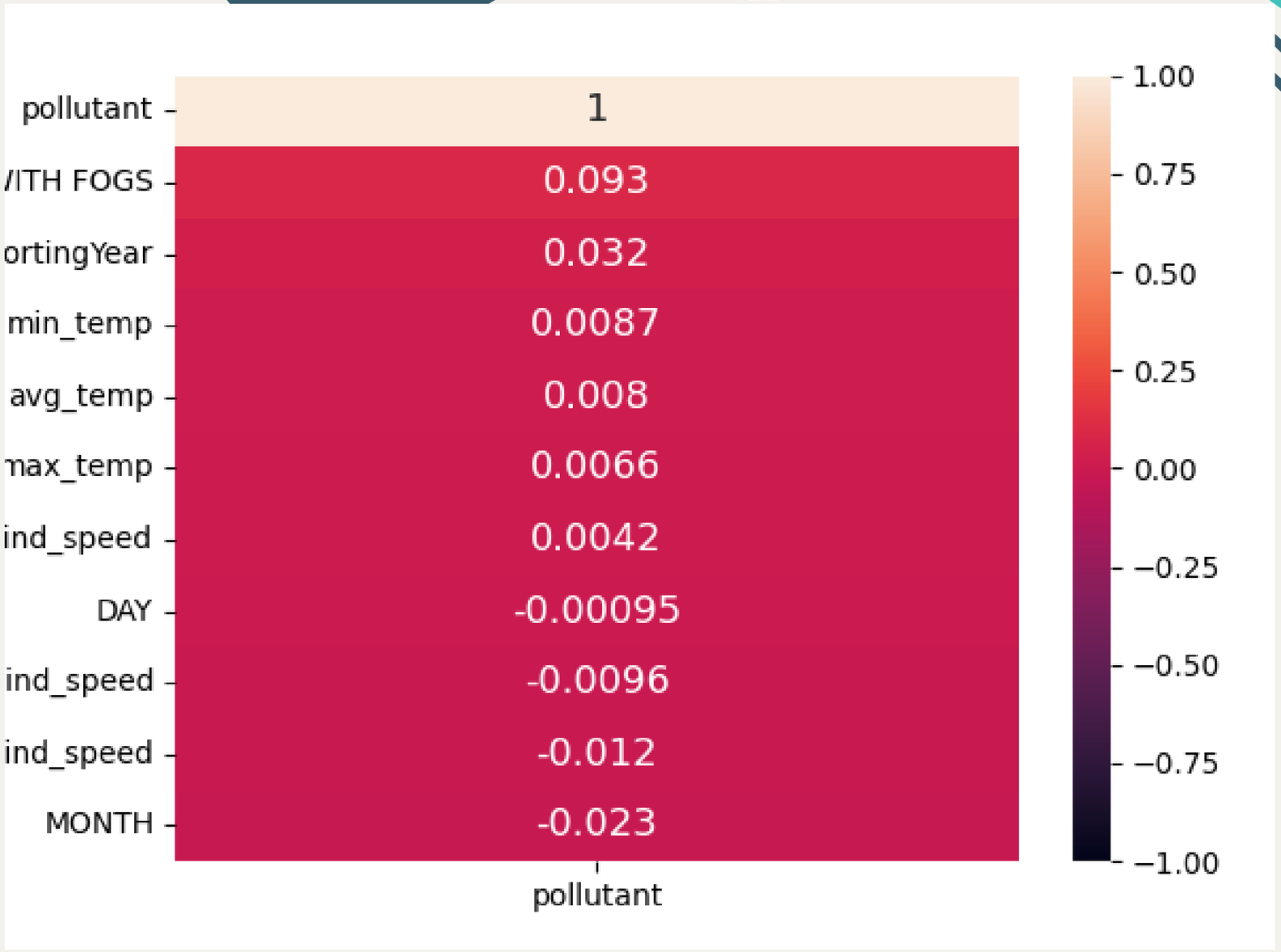
Limpieza de datos:

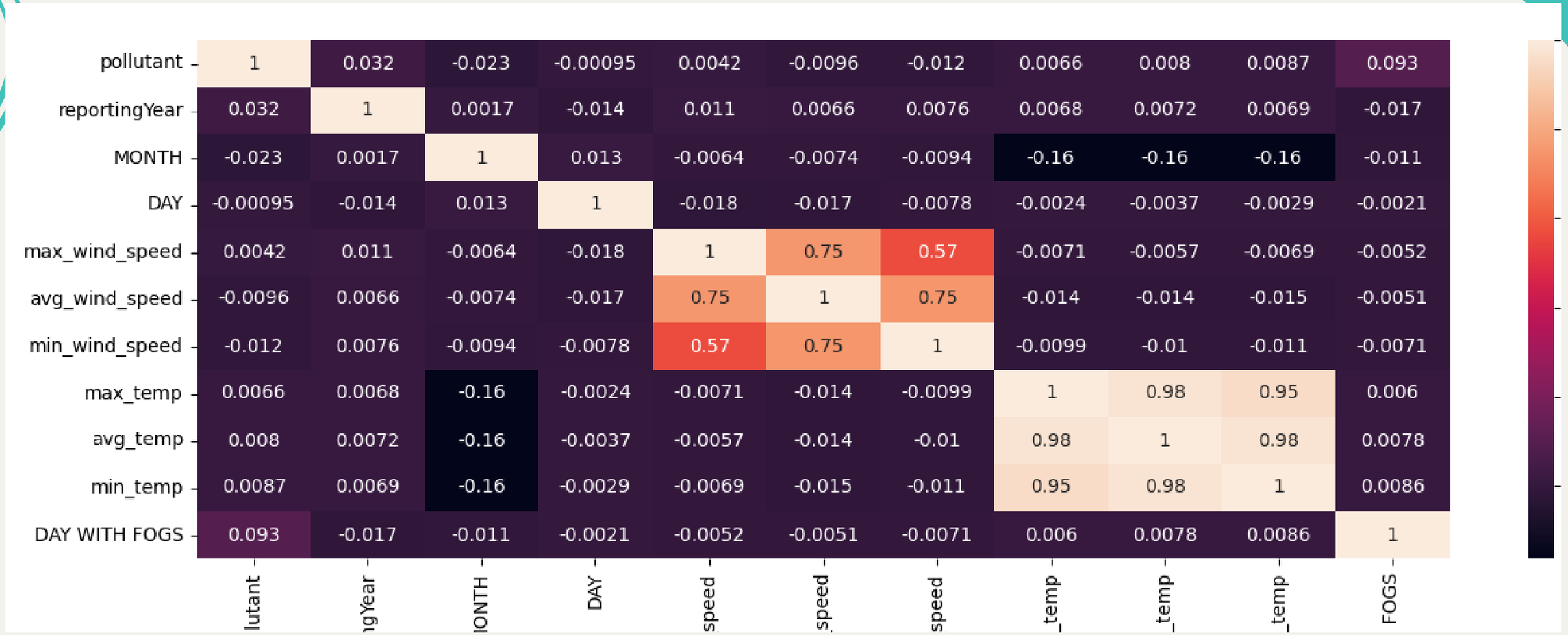
Eliminar columnas:

- Unnamed: 0
- TargetRelease
- Continent

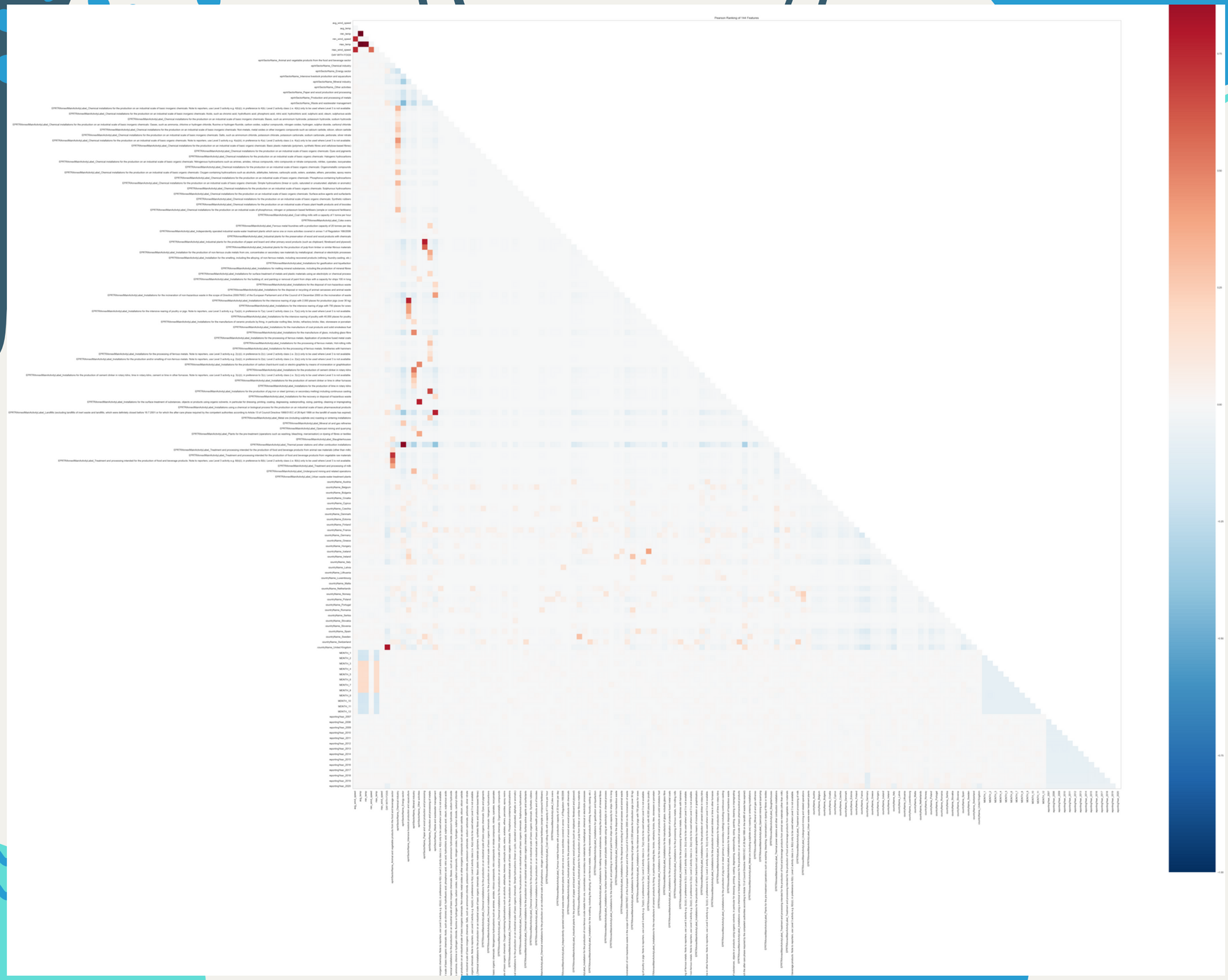
Codificar target: 0,1,2

2. EDA


pollutant
/ITH FOGS
ortingYear
min_temp
avg_temp
max_temp
ind_speed
DAY
ind_speed
ind_speed
MONTH
1
0.093
0.032
0.0087
0.008
0.0066
0.0042
-0.00095
-0.0096
-0.012
-0.023
pollutant
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


pollutant 1 0.032 -0.023 -0.00095 0.0042 -0.0096 -0.012 0.0066 0.008 0.0087 0.093
reportingYear 0.032 1 0.0017 -0.014 0.011 0.0066 0.0076 0.0068 0.0072 0.0069 -0.017
MONTH -0.023 0.0017 1 0.013 -0.0064 -0.0074 -0.0094 -0.16 -0.16 -0.16 -0.011
DAY -0.00095 -0.014 0.013 1 -0.018 -0.017 -0.0078 -0.0024 -0.0037 -0.0029 -0.0021
max_wind_speed 0.0042 0.011 -0.0064 -0.018 1 0.75 0.57 -0.0071 -0.0057 -0.0069 -0.0052
avg_wind_speed -0.0096 0.0066 -0.0074 -0.017 0.75 1 0.75 -0.014 -0.014 -0.015 -0.0051
min_wind_speed -0.012 0.0076 -0.0094 -0.0078 0.57 0.75 1 -0.0099 -0.01 -0.011 -0.0071
max_temp 0.0066 0.0068 -0.16 -0.0024 -0.0071 -0.014 -0.0099 1 0.98 0.95 0.006
avg_temp 0.008 0.0072 -0.16 -0.0037 -0.0057 -0.014 -0.01 0.98 1 0.98 0.0078
min_temp 0.0087 0.0069 -0.16 -0.0029 -0.0069 -0.015 -0.011 0.95 0.98 1 0.0086
DAY WITH FOGS 0.093 -0.017 -0.011 -0.0021 -0.0052 -0.0051 -0.0071 0.006 0.0078 0.0086 1
lutant
gYear
IONTH
DAY
speed
speed
speed
_temp
_temp
_temp
FOGS

3. Arbol de decisión- Feature importances


Pearson Ranking of 144 Features
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00

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

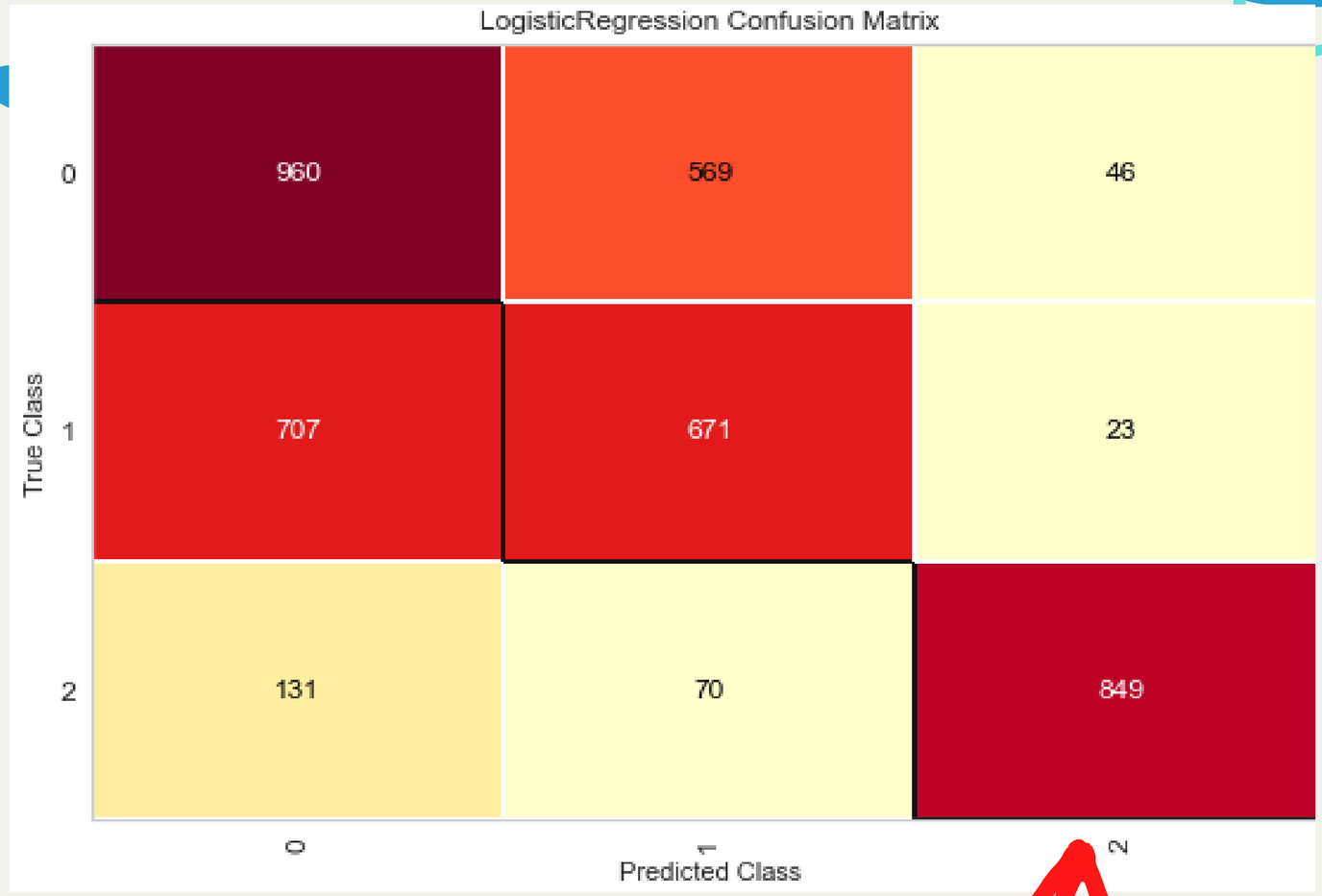
Añadimos:

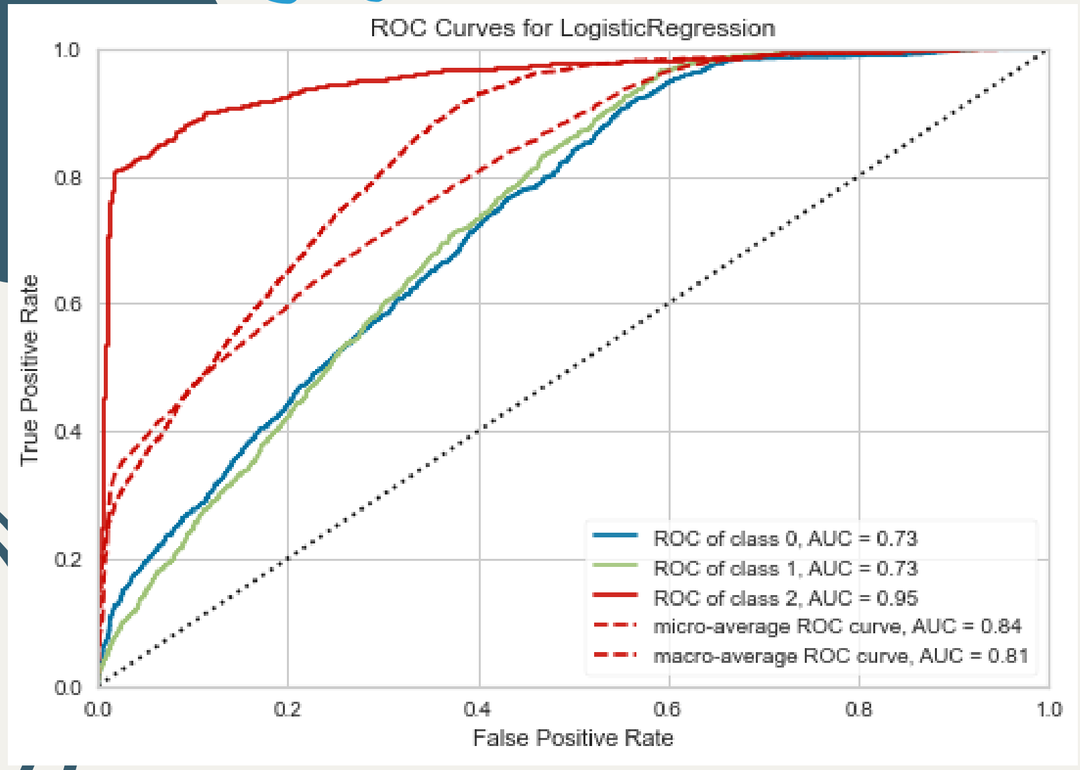
- Regresión logística
- Random Forest
- Support Vector Machine (SVM)

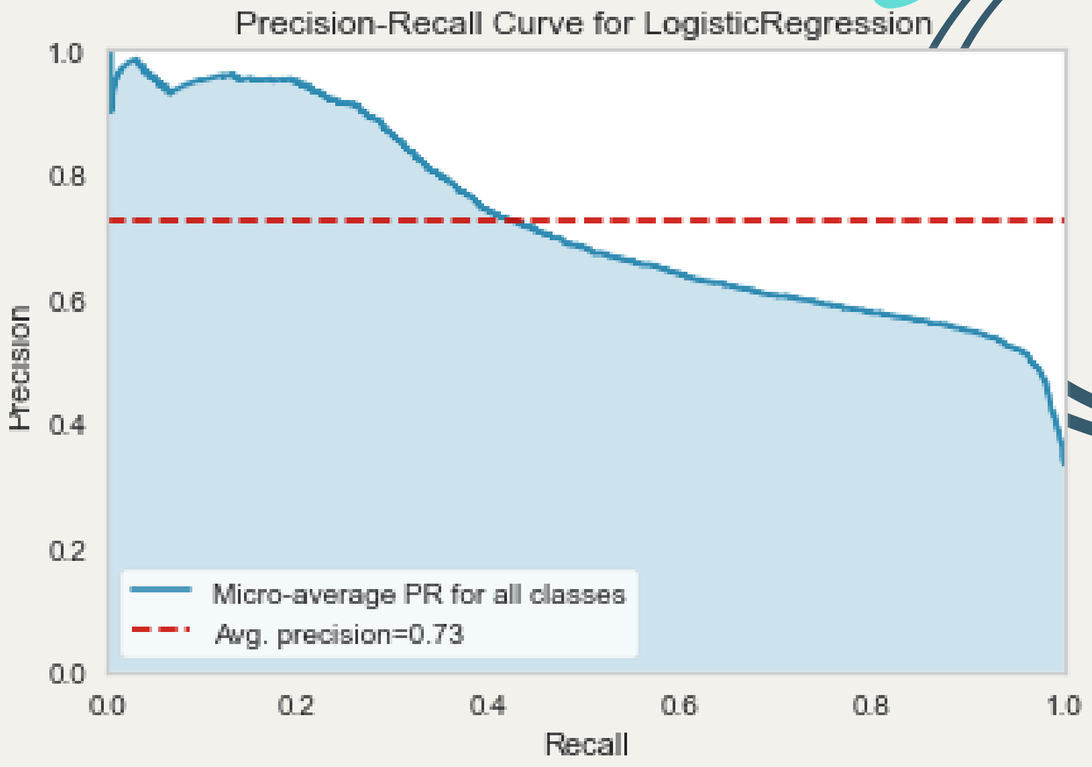
Y el mejor modelo es...

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

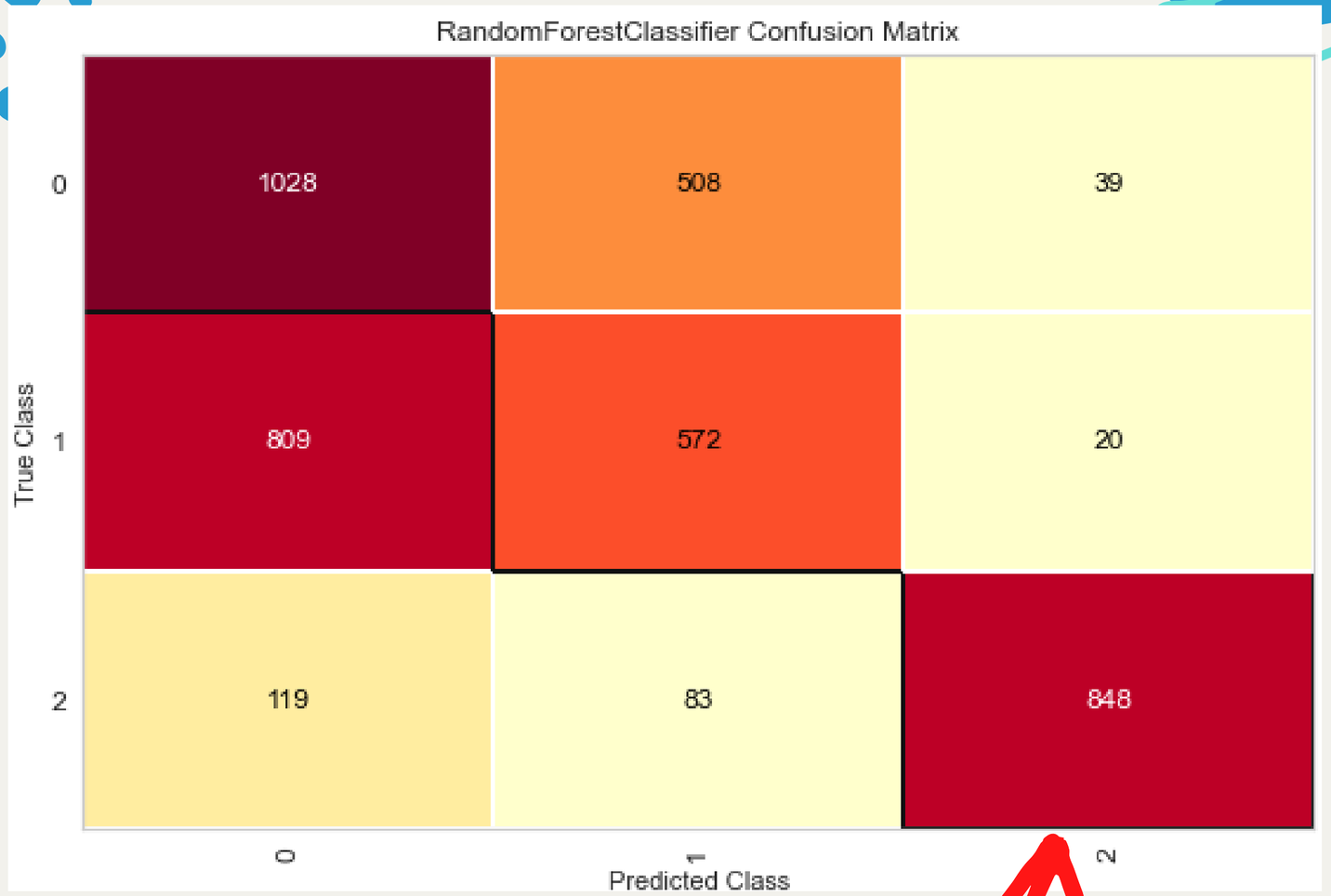
6. Análisis mejor modelo: regresión logística

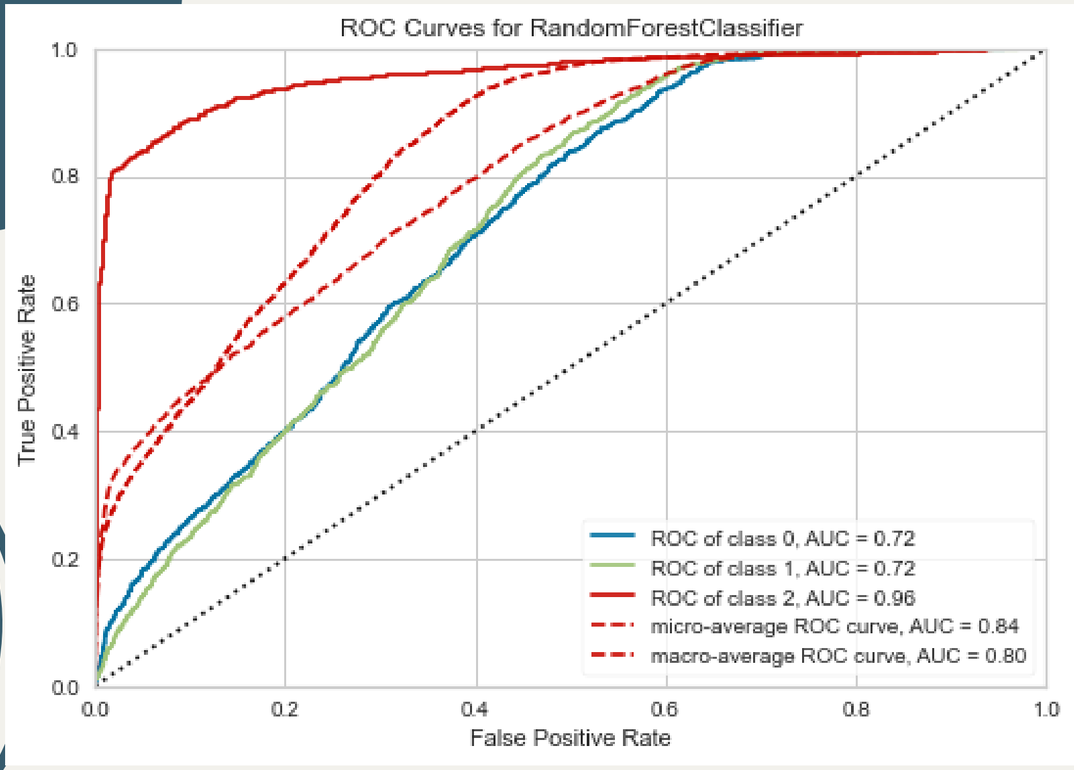

LogisticRegression Confusion Matrix
True Class
0
1
2
960
569
46
707
671
23
131
70
849
0
1
2
Predicted Class


ROC Curves for LogisticRegression
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.73
ROC of class 1, AUC = 0.73
ROC of class 2, AUC = 0.95
micro-average ROC curve, AUC = 0.84
macro-average ROC curve, AUC = 0.81


Precision-Recall Curve for LogisticRegression
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Micro-average PR for all classes
Avg. precision=0.73

7. Análisis segundo mejor modelo: random forest


RandomForestClassifier Confusion Matrix
True Class
Predicted Class
0
1
2
1028
508
39
809
572
20
119
83
848


ROC Curves for RandomForestClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.72
ROC of class 1, AUC = 0.72
ROC of class 2, AUC = 0.96
micro-average ROC curve, AUC = 0.84
macro-average ROC curve, AUC = 0.80


Precision-Recall Curve for RandomForestClassifier
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Micro-average PR for all classes
Avg. precision=0.72

8. Conclusiones

- **Modelo random forest predice mejor que regresión logística los contaminates de tipo 2 (CH4)**
- **Modelo random forest obtiene mejor score en Kaggle que regresión logística**

¡Muchas gracias!