Estimación de anomalías de temperatura del aire en el departamento de Antioquia

Modelos no supervisados

Valentina Sánchez Castaño

UNIVERSIDAD NACIONAL DE COLOMBIA

Justificación

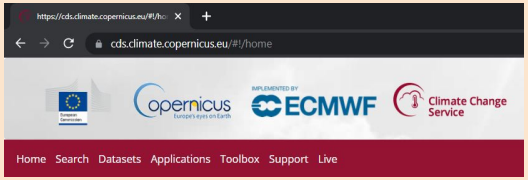
Mapping Monthly Air Temperature in the Tibetan Plateau From MODIS Data Based on Machine Learning Methods

Yongming Xu, Anders Knudby, Yan Shen, and Yonghong Liu

- La temperatura del aire cerca de la superficie es uno de los parámetros meteorológicos más críticos en los estudios ambientales y climáticos.
- Las mediciones puntuales observadas por la estación no son especialmente representativas y pueden introducir sesgos cuando se utilizan para caracterizar la variación de temperatura en grandes áreas, especialmente en regiones subdesarrolladas y montañosas.
- Una buena medición de la temperatura del aire cerca de la superficie es requerida para conocer si una región es sensible al cambio climático.

I. Construcción del DataFrame

Obtención de datos



Evaporación total (m of water equivalent)
Temperatura del suelo (K)
Cobertura de nubes (%)
Velocidad del viento (m/s)
Tipo de cobertura (–)

NASA MODIS
Moderate Resolution Imaging Spectroradiometer

NDVI (–)

UAF
ALASKA SATELLITE FACILITY

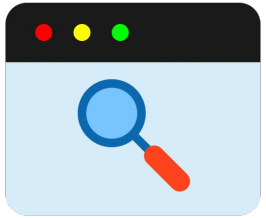
Modelo digital
de elevación

DHIME
DATOS HIDROLÓGICOS Y METEOROLÓGICOS

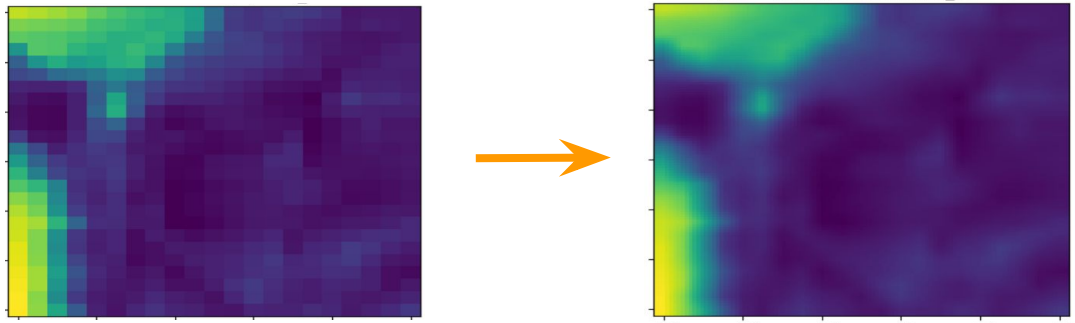
Temperatura
del aire

Obtención de datos

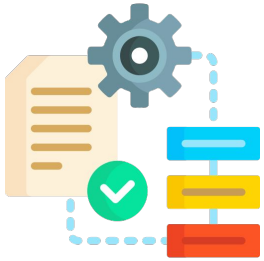
Resolución espacial	Resolución temporal	Cantidad de datos	Variable (Y)
5 km	Mensual	2003-2014	36 estaciones



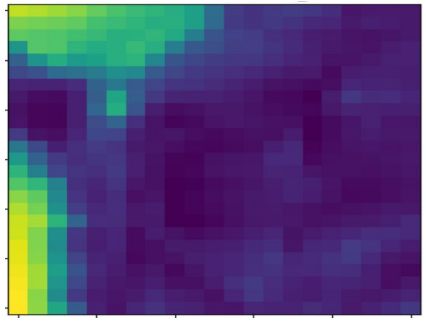
Procesamiento de datos

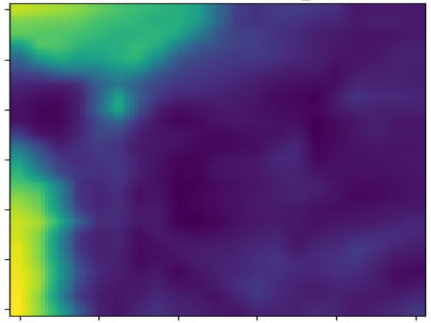


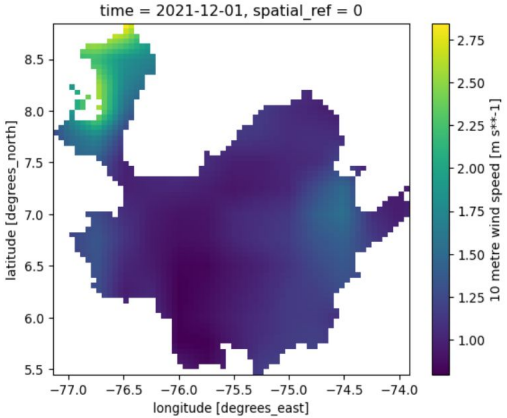
Llevar todas las bases de datos a la misma resolución



Procesamiento de datos

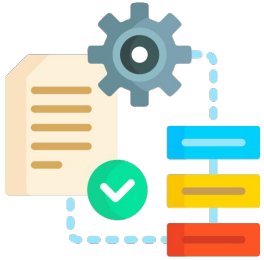




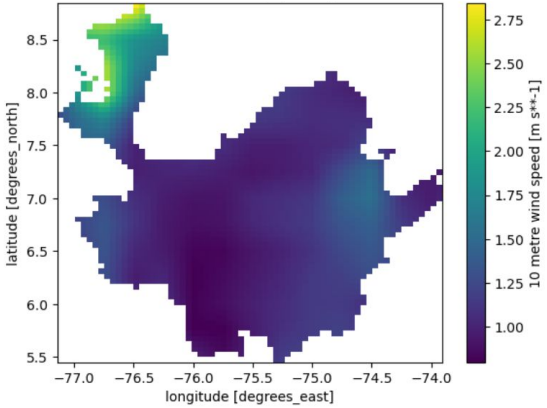


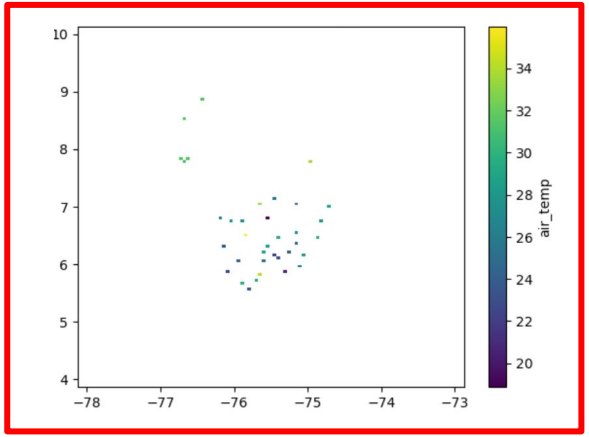
Llevar todas las bases de datos a la misma resolución

Recortar los datos según la zona de interés



Procesamiento de datos





Reducción considerable de los datos por la cantidad de estaciones disponibles

Ocho variables independientes

DataFrame

	LandCover	Wind	DEM	NDVI	Clouds	Temp	Eva	AirTempCategory
405	40.0	4.404931	39.0	0.69	0.473909	27.549194	-0.003469	0
406	40.0	4.108960	49.0	0.77	0.480379	27.544739	-0.003545	0
1328	40.0	3.921639	14.0	0.78	0.538886	27.015350	-0.004009	0
1395	50.0	3.284423	45.0	0.77	0.595392	26.776520	-0.003976	0
1396	11.0	2.966187	79.0	0.78	0.636385	26.358978	-0.004067	0
...	...	...	...	...	...	...	...	...
664657	50.0	1.179073	2368.0	0.81	0.785024	16.397888	-0.003265	0
664716	50.0	0.957614	737.0	0.77	0.786089	20.228333	-0.003138	0
664847	30.0	0.910709	805.0	0.79	0.795860	17.989594	-0.003124	0
664909	50.0	0.887851	1225.0	0.85	0.841868	16.256653	-0.003198	0
664977	40.0	0.908780	2469.0	0.82	0.827497	15.593140	-0.003231	0

5134 rows × 8 columns

Ocho variables independientes

DataFrame

	LandCover	Wind	DEM	NDVI	Clouds	Temp	Eva	AirTempCategory
405	40.0	4.404931	39.0	0.69	0.473909	27.549194	-0.003469	0
406	40.0	4.108960	49.0	0.77	0.480379	27.544739	-0.003545	0
1328	40.0	3.921639	14.0	0.78	0.538886	27.015350	-0.004009	0
1395	50.0	3.284423	45.0	0.77	0.595392	26.776520	-0.003976	0
1396	11.0	2.966187	79.0	0.78	0.636385	26.358978	-0.004067	0
...	...	...	...	...	...	...	...	...
664657	50.0	1.179073	2368.0	0.81	0.785024	16.397888	-0.003265	0
664716	50.0	0.957614	737.0	0.77	0.786089	20.228333	-0.003138	0
664847	30.0	0.910709	805.0	0.79	0.795860	17.989594	-0.003124	0
664909	50.0	0.887851	1225.0	0.85	0.841868	16.256653	-0.003198	0
664977	40.0	0.908780	2469.0	0.82	0.827497	15.593140	-0.003231	0

5134 rows × 8 columns

Total de datos disponibles

Ocho variables independientes

DataFrame

	LandCover	Wind	DEM	NDVI	Clouds	Temp	Eva	AirTempCategory
405	40.0	4.404931	39.0	0.69	0.473909	27.549194	-0.003469	0
406	40.0	4.108960	49.0	0.77	0.480379	27.544739	-0.003545	0
1328	40.0	3.921639	14.0	0.78	0.538886	27.015350	-0.004009	0
1395	50.0	3.284423	45.0	0.77	0.595392	26.776520	-0.003976	0
1396	11.0	2.966187	79.0	0.78	0.636385	26.358978	-0.004067	0
...	...	...	...	...	...	...	...	...
664657	50.0	1.179073	2368.0	0.81	0.785024	16.397888	-0.003265	0
664716	50.0	0.957614	737.0	0.77	0.786089	20.228333	-0.003138	0
664847	30.0	0.910709	805.0	0.79	0.795860	17.989594	-0.003124	0
664909	50.0	0.887851	1225.0	0.85	0.841868	16.256653	-0.003198	0
664977	40.0	0.908780	2469.0	0.82	0.827497	15.593140	-0.003231	0

5134 rows × 8 columns

Total de datos disponibles

Variable dependiente categórica según la anomalía

2. Análisis exploratorio de los datos

Variables dependiente categórica

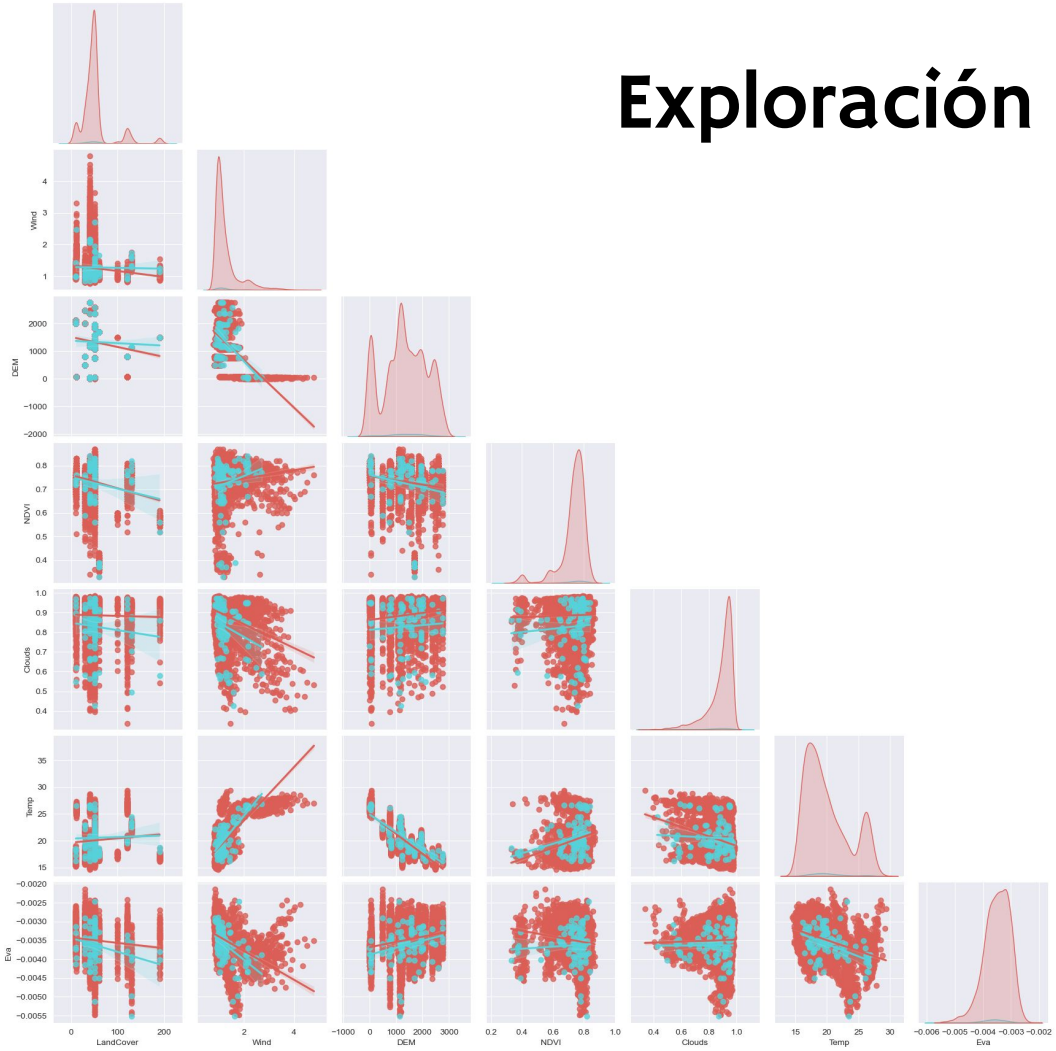
0	1
5028	106

Desbalance

2.11% de los datos

AirTempCategory	LandCover	Wind	DEM	NDVI	Clouds	Temp	Eva
0	51.141806	1.253162	1329.855410	0.731943	0.886798	20.052518	-0.003499
1	53.028302	1.274939	1334.122642	0.725472	0.828695	20.539396	-0.003635

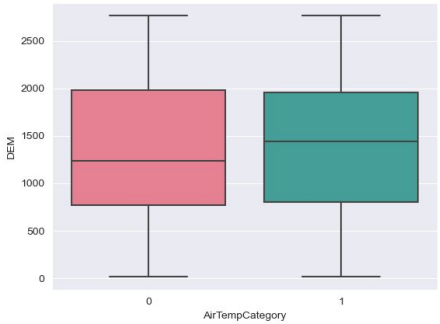
Exploración inicial

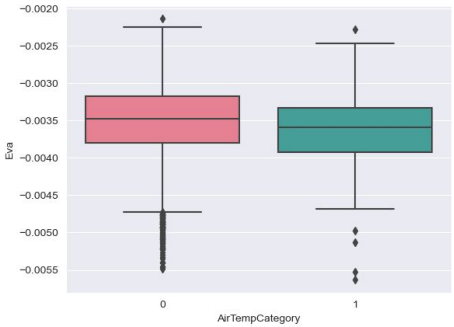


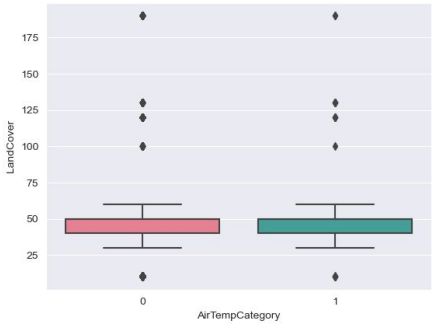
Se evidencia nuevamente el desbalance de los datos

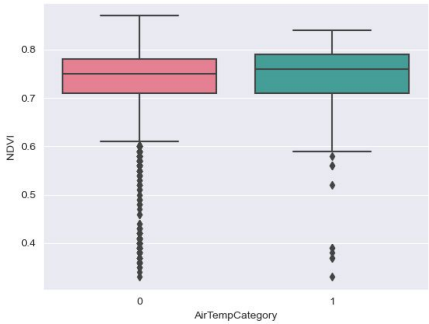
Se logra observar cierta relación entre la variables categoría y variables independientes

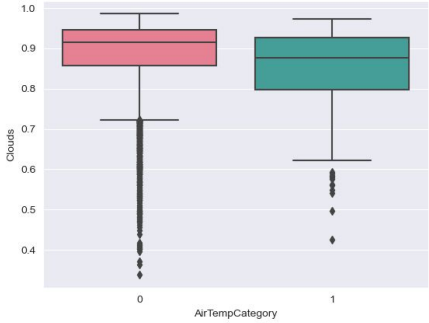
Exploración inicial

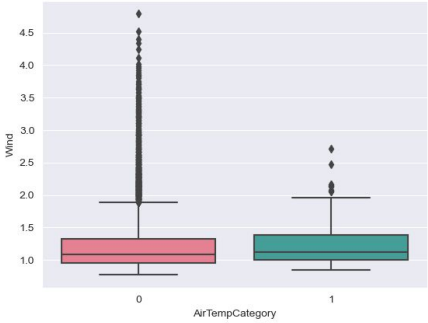


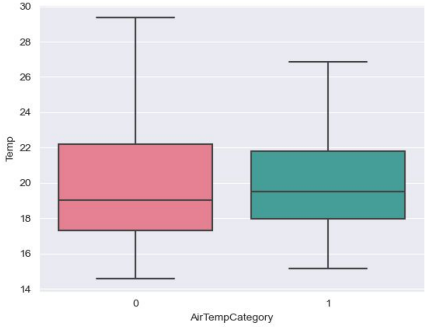




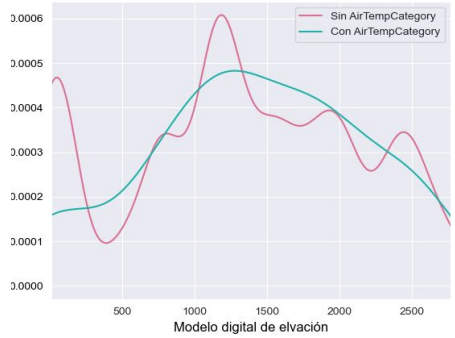


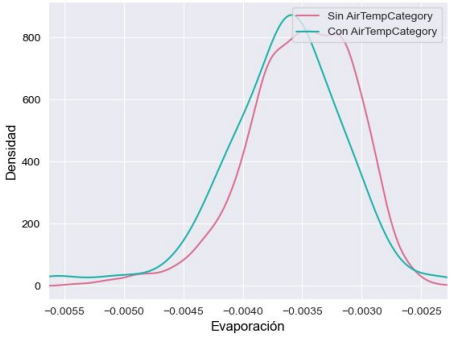


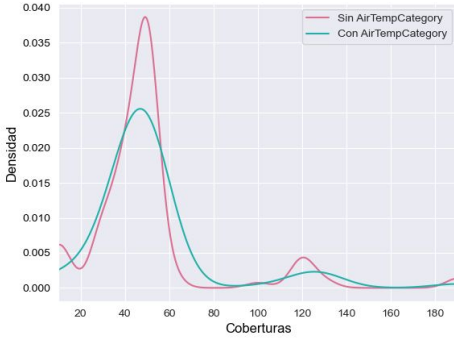


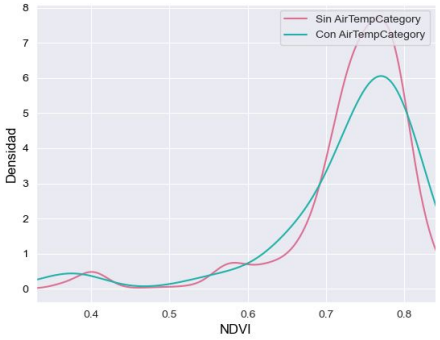


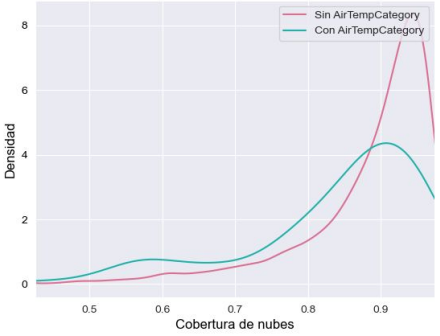
Exploración inicial

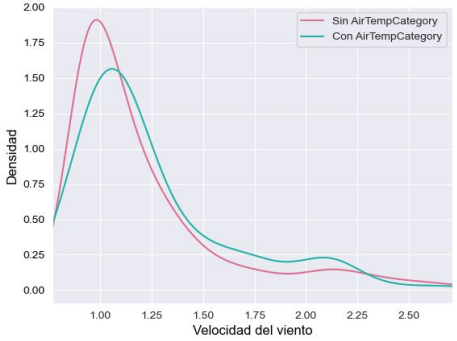


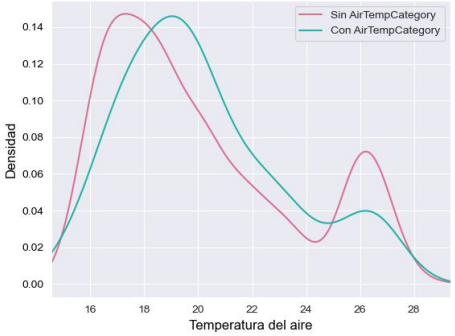




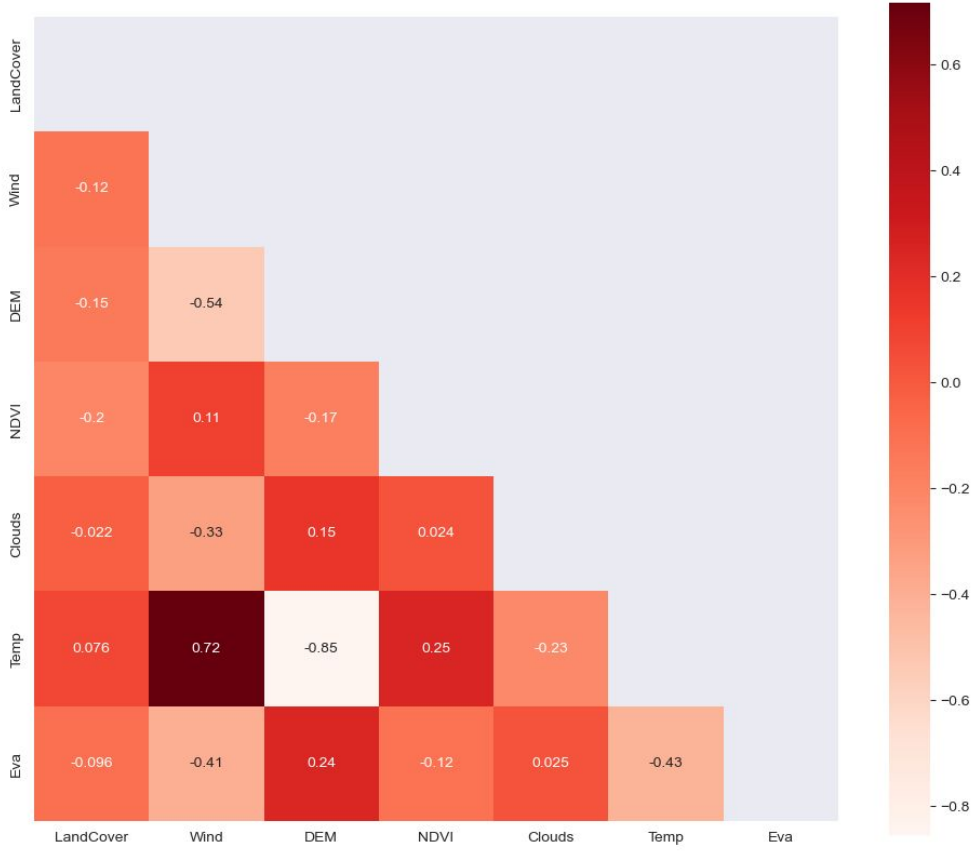








Exploración inicial



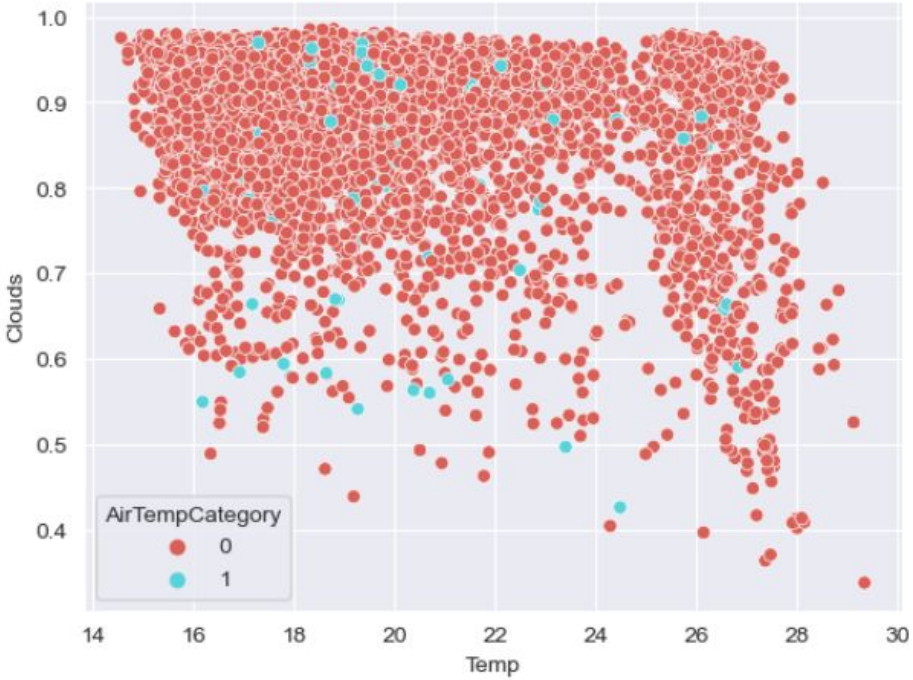
La mayor correlación se presenta entre la temperatura del suelo y el aire.

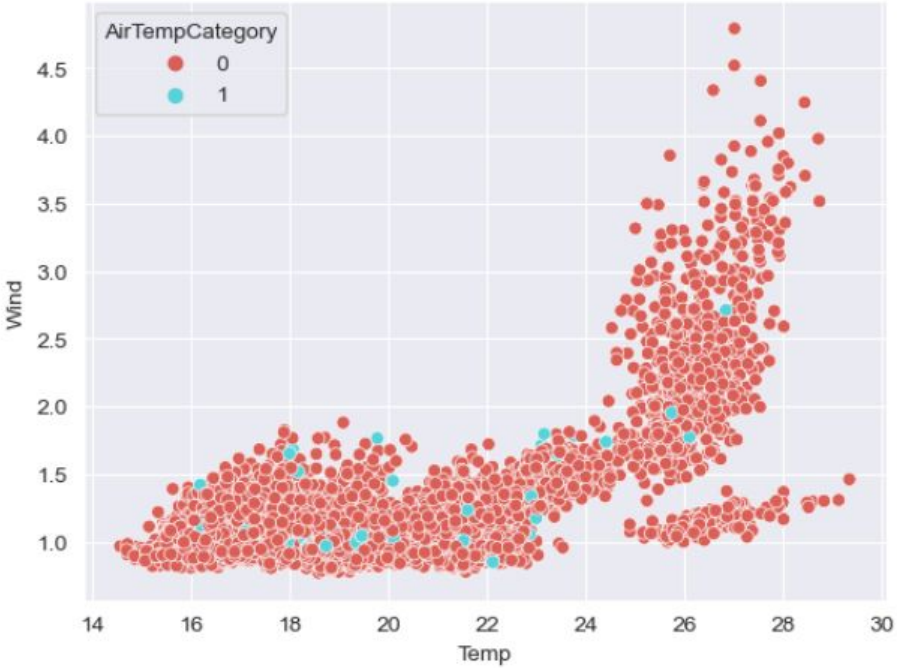
3. Selección de variables

	LandCover	Wind	DEM	NDVI	Clouds	Temp	Eva
Variance Threshold Feature Selection	X	X	X			X	
Variables Univariado	X	X	X			X	
Recursive Feature Elimination		X		X	X	X	
Select From Model				X			
Sequential Feature Selection (forward)	X	X	X	X			
Sequential Feature Selection (backward)				X	X	X	X
ELI5						X	
Modelo Ensamblado		X			X	X	X
Valor P					X		X

4. Desbalance

Exploración inicial





Validación

Regresión Logística

kfold = KFold(n_splits=5)
metric='manhattan'
Scoring ='roc_auc'

894	622
12	13

```
[0.33046415 0.49469697 0.40883191 0.48121469 0.39856603]
La precisión del modelo es: 42.28 %
Counter({0: 906, 1: 635})
```

	precision	recall	f1-score	support
0	0.99	0.59	0.74	1516
1	0.02	0.52	0.04	25
accuracy			0.59	1541
macro avg	0.50	0.55	0.39	1541
weighted avg	0.97	0.59	0.73	1541

Validación

K Neighbors Classifier

kfold = KFold(n_splits=5)
Class_weigh t = 'balanced'
Scoring ='roc_auc'

996	15
14	2

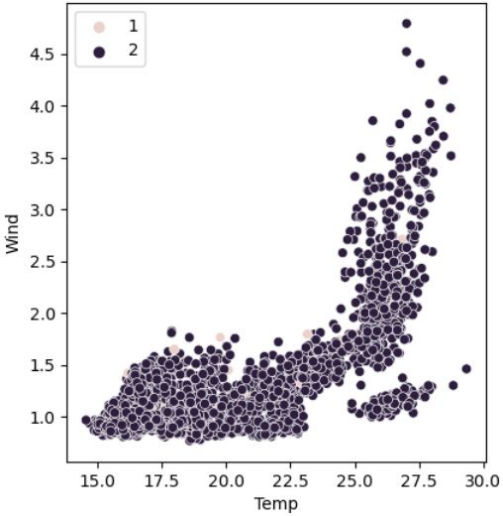
```
[0.48258706 0.47885572 0.53758983 0.59480376 0.53510227 0.47139303
 0.58195135 0.48254364 0.48004988 0.59365475]
La precisión del modelo es: 52.39 %
Counter({0: 1010, 1: 17})
```

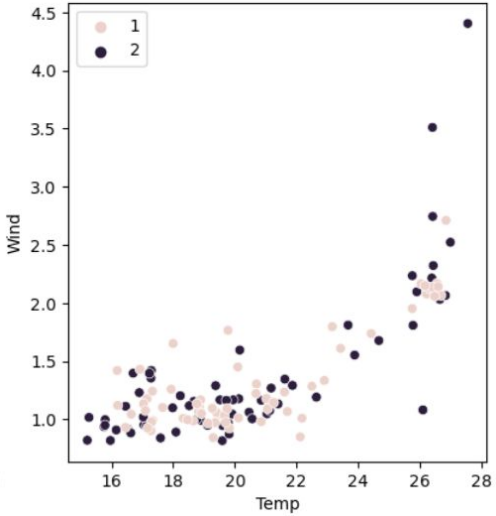
	precision	recall	f1-score	support
0	0.98	0.57	0.72	1499
1	0.03	0.55	0.06	42
accuracy			0.56	1541
macro avg	0.51	0.56	0.39	1541
weighted avg	0.95	0.56	0.70	1541

Técnicas de preprocesamiento

Random Under Sampler

sampling_strategy='majority'



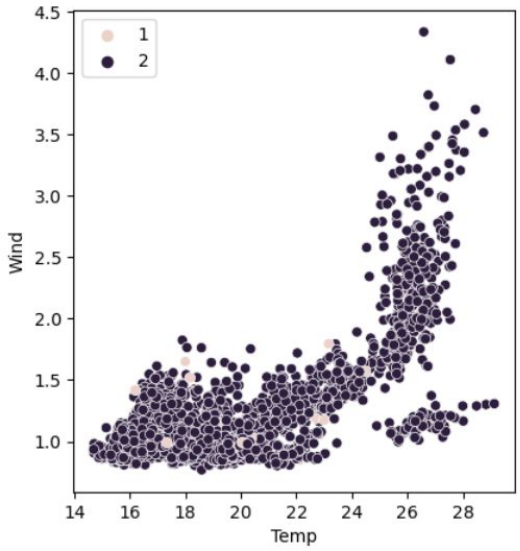


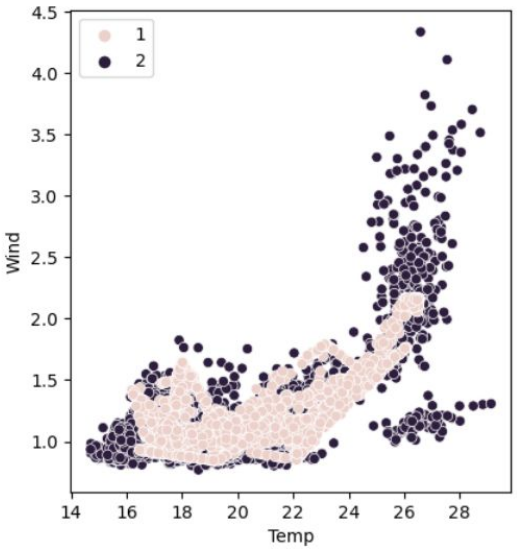
641	858
17	25

	precision	recall	f1-score	support
0	0.97	0.43	0.59	1499
1	0.03	0.60	0.05	42
accuracy			0.43	1541
macro avg	0.50	0.51	0.32	1541
weighted avg	0.95	0.43	0.58	1541

Técnicas de preprocesamiento

SMOTE



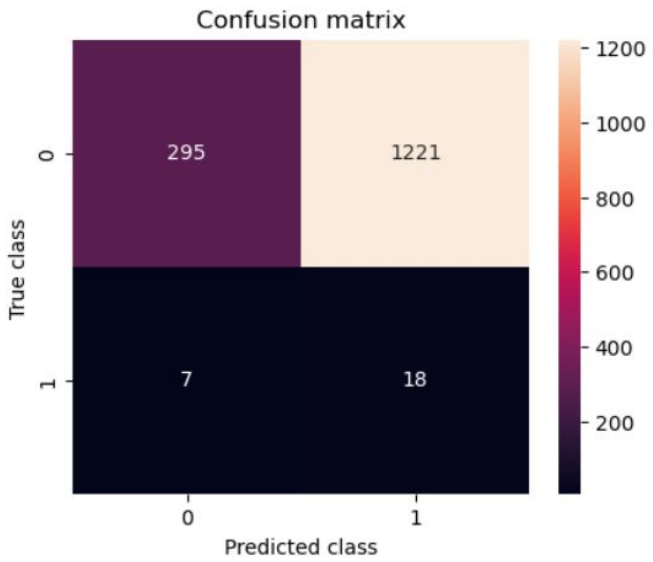


1614	893
35	25

	precision	recall	f1-score	support
0	0.98	0.64	0.78	2507
1	0.03	0.42	0.05	60
accuracy			0.64	2567
macro avg	0.50	0.53	0.41	2567
weighted avg	0.96	0.64	0.76	2567

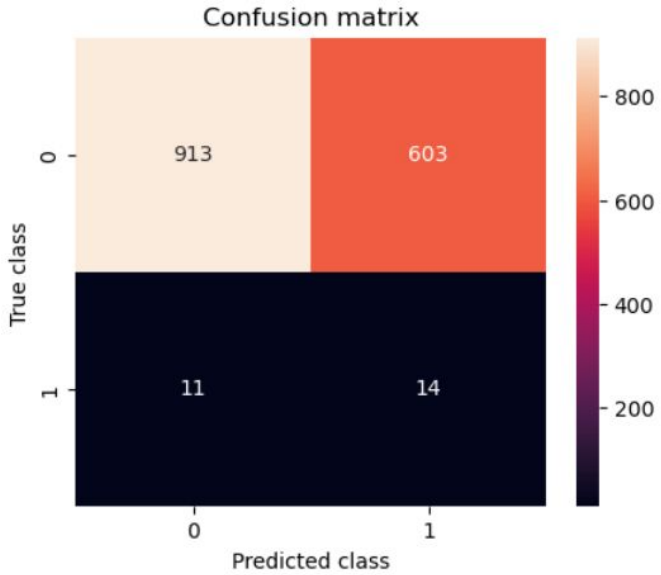
Técnicas de preprocesamiento

NearMiss



	precision	recall	f1-score	support
0	0.98	0.19	0.32	1516
1	0.01	0.72	0.03	25
accuracy			0.20	1541
macro avg	0.50	0.46	0.18	1541
weighted avg	0.96	0.20	0.32	1541

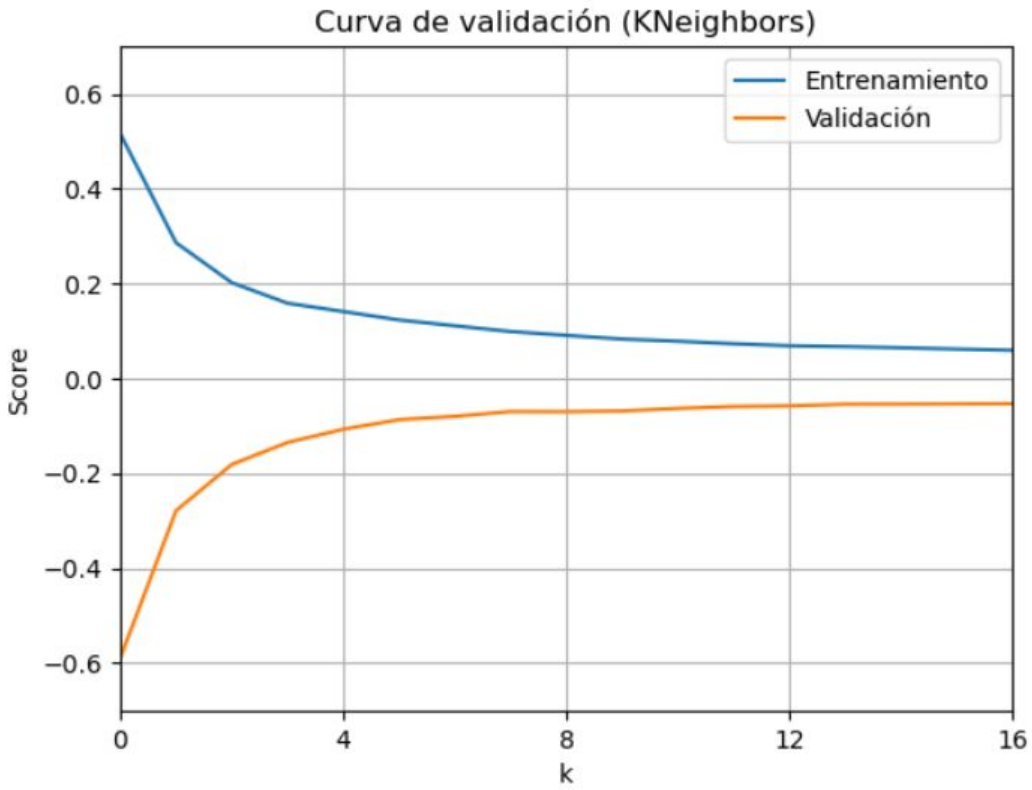
SMOTETomek



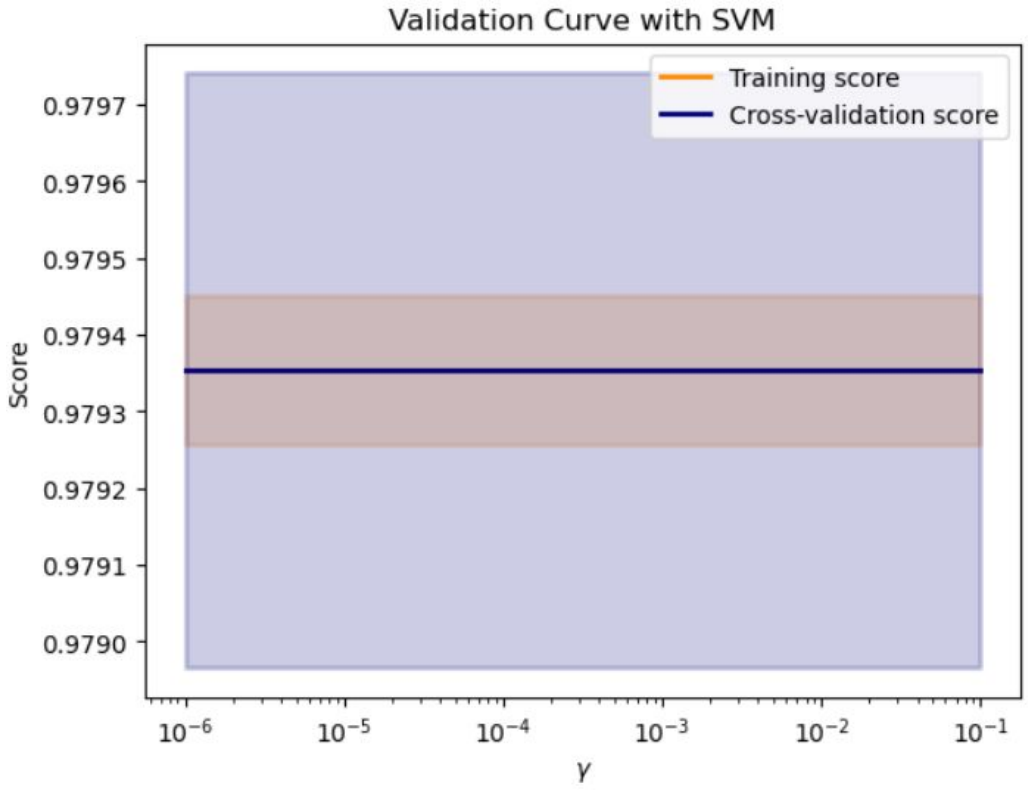
	precision	recall	f1-score	support
0	0.99	0.60	0.75	1516
1	0.02	0.56	0.04	25
accuracy			0.60	1541
macro avg	0.51	0.58	0.40	1541
weighted avg	0.97	0.60	0.74	1541

5. Selección de Hiperparámetros

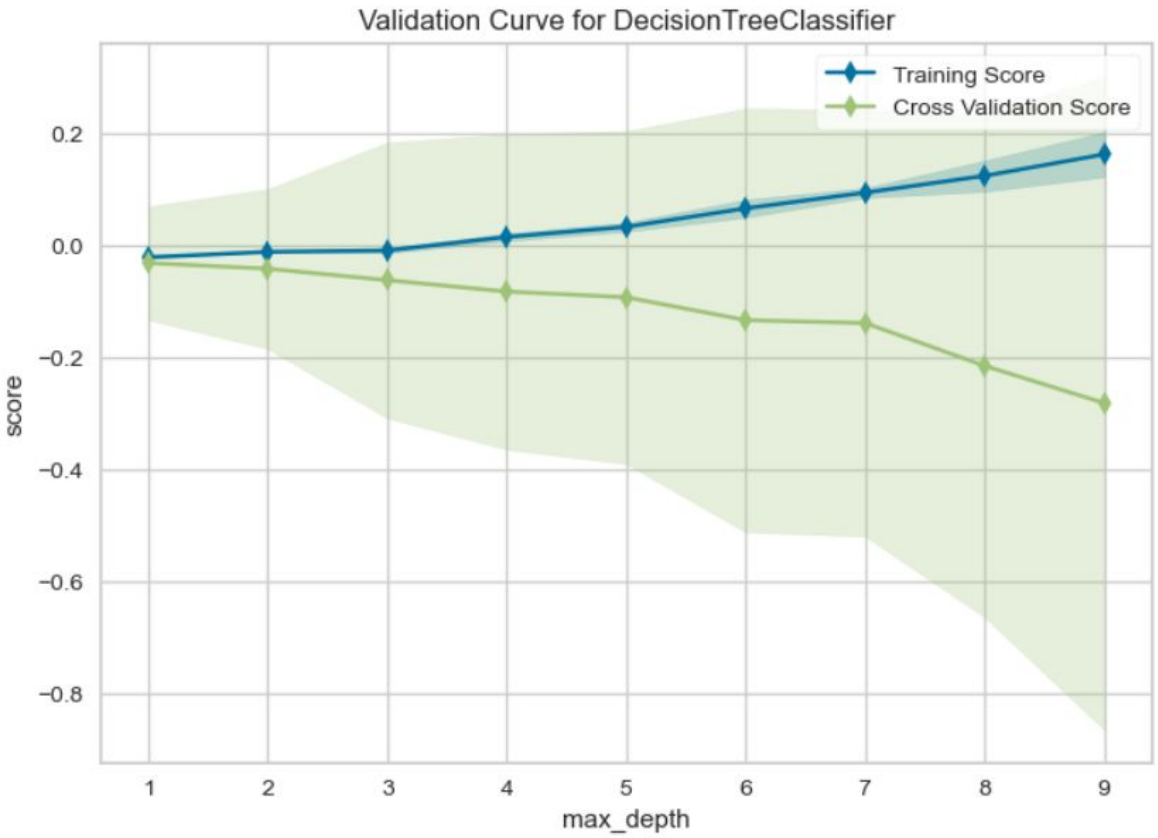
Parámetro K


Curva de validación (KNeighbors)
Entrenamiento
Validación
Score
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
4
8
12
16
k

Parámetro G

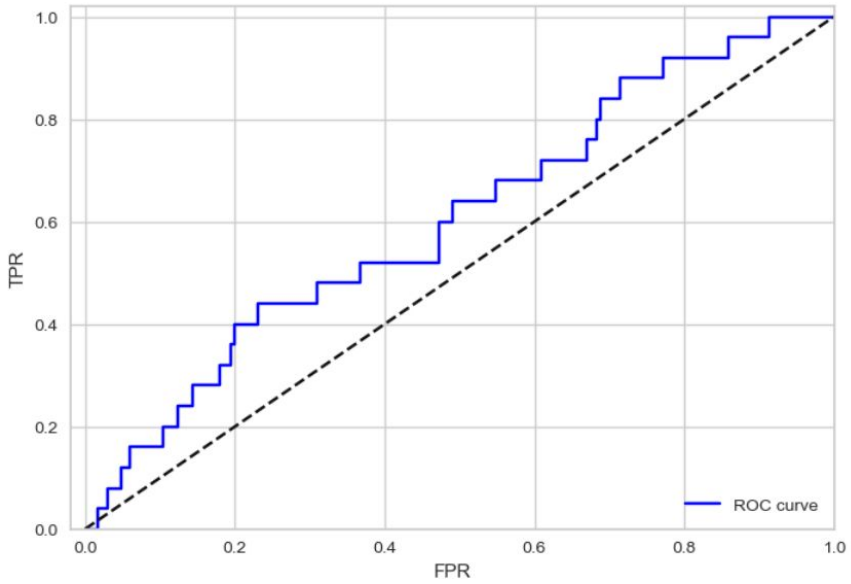

Validation Curve with SVM
Training score
Cross-validation score
Score
0.9797
0.9796
0.9795
0.9794
0.9793
0.9792
0.9791
0.9790
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
γ

Párametro max depth



6. Métricas de Evaluación

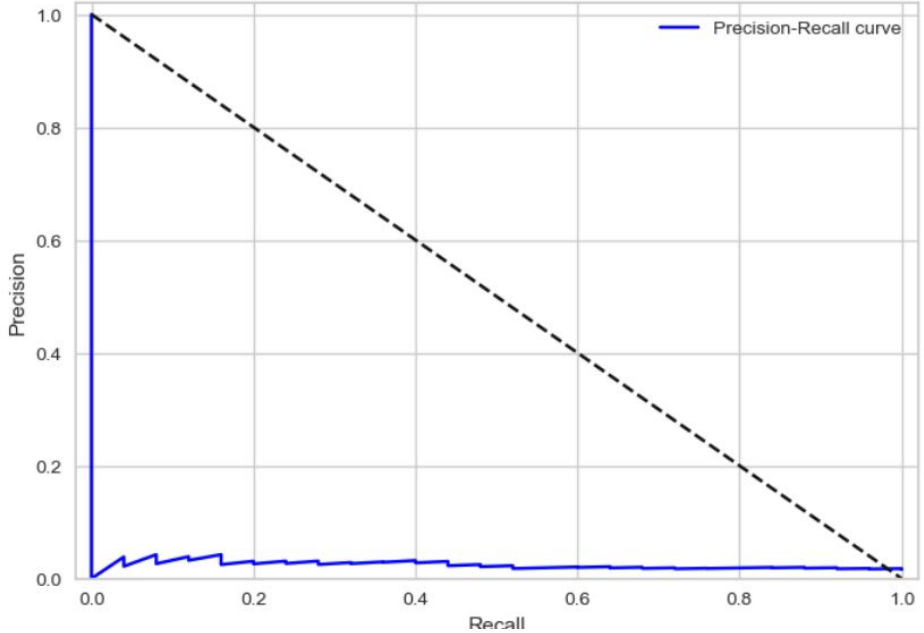
Curva ROC



Área bajo la curva: 0.524

El modelo no tiene capacidad discriminatoria.

Curva Recall

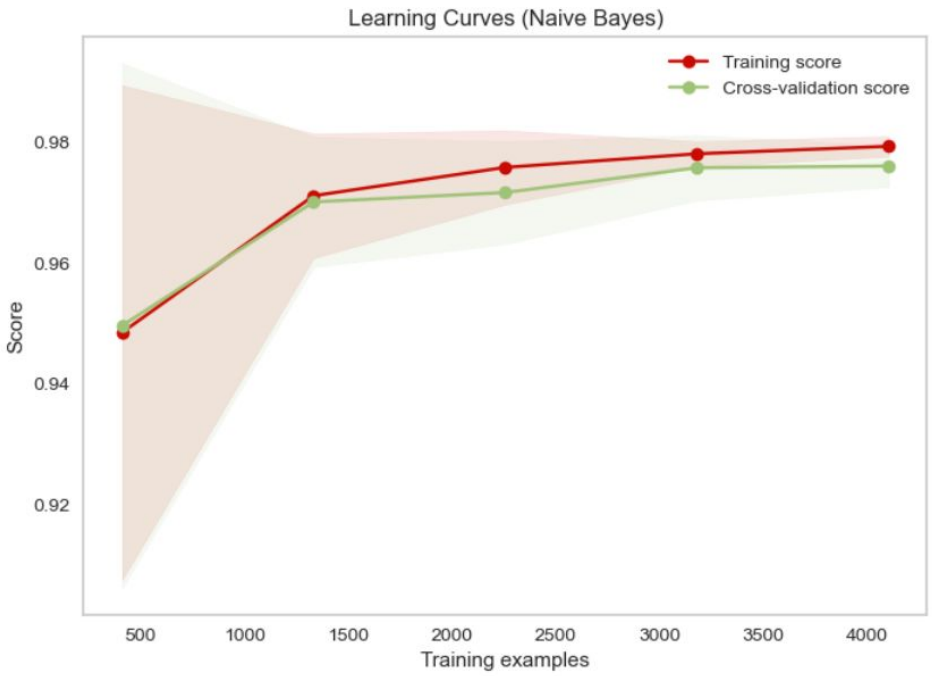


Área bajo la curva: 0.023

El modelo no mejora tras el entrenamiento.

7. Bias y varianza

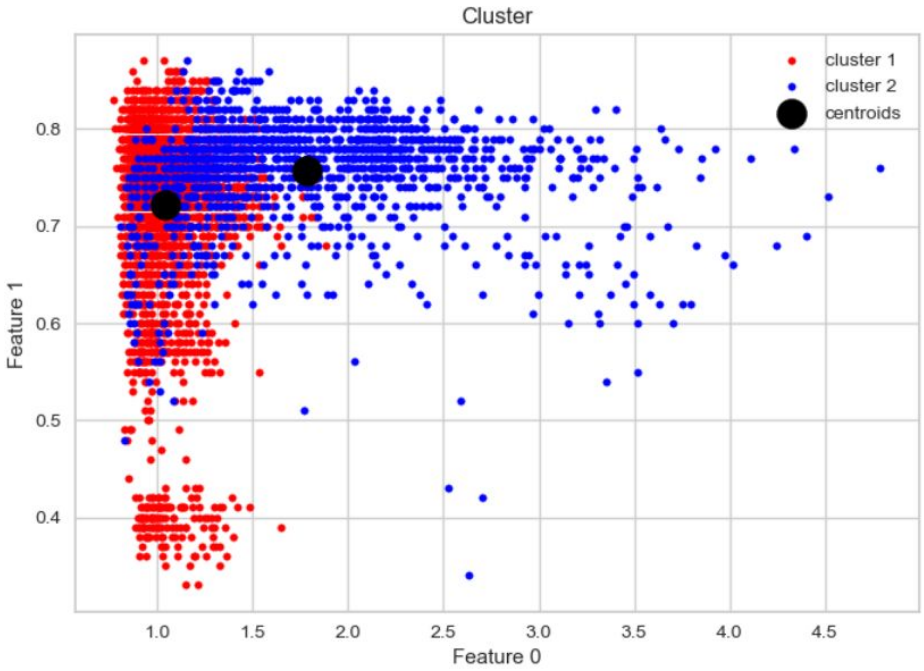
Curva de aprendizaje



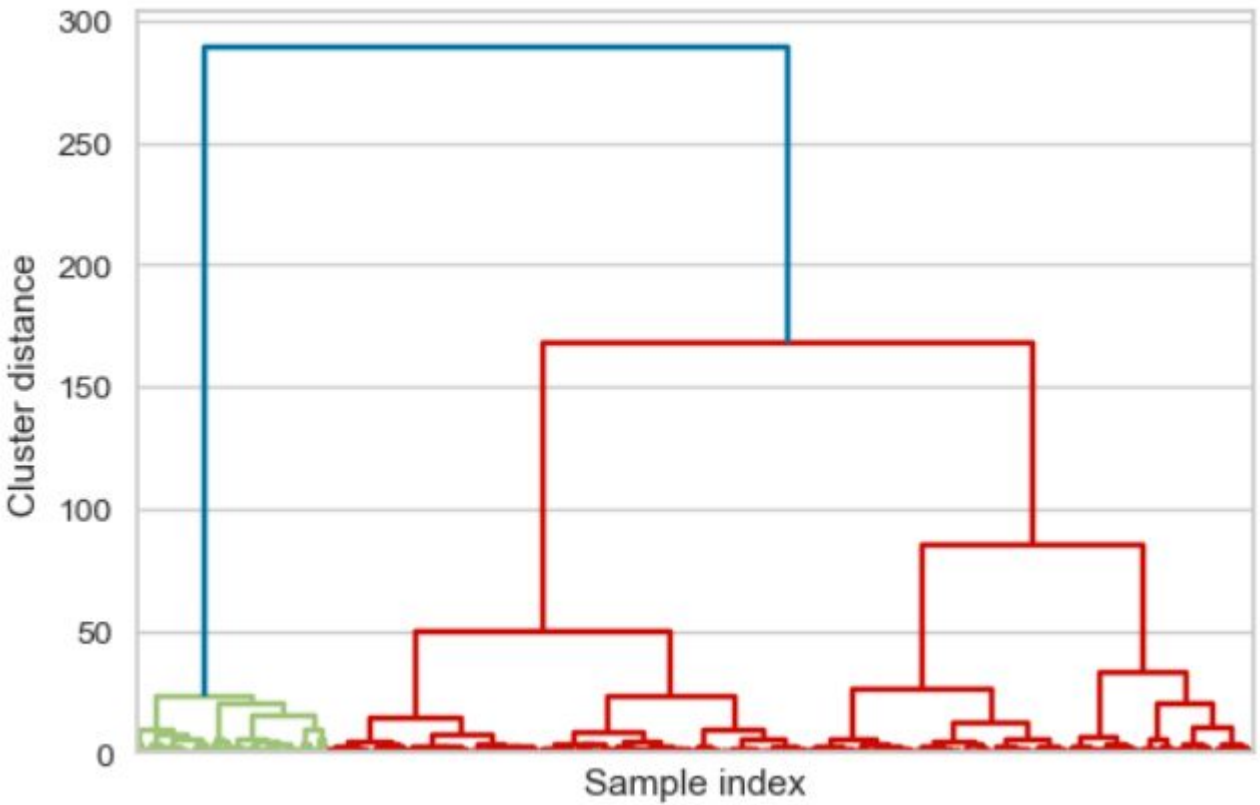
Sobre ajuste

7. Clustering

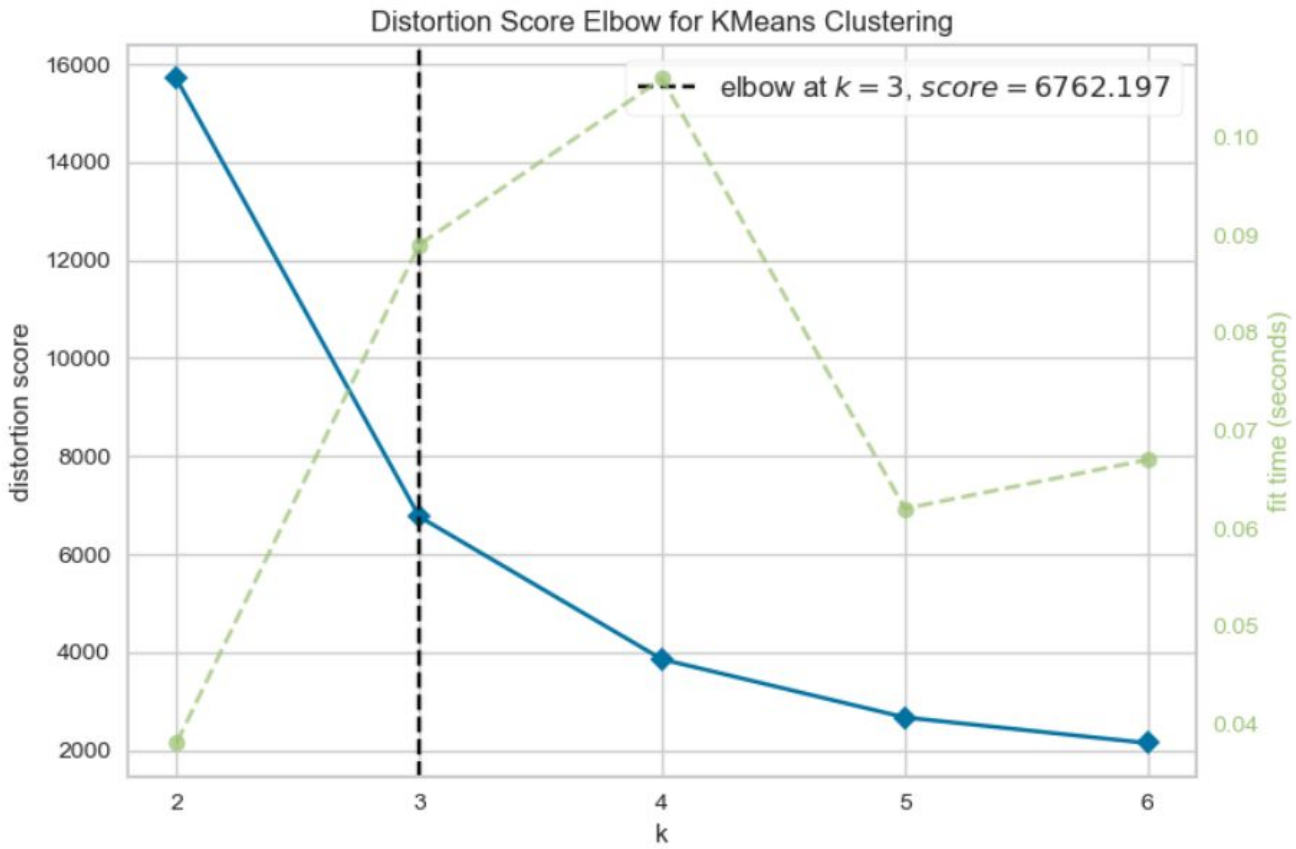
K means


Cluster
cluster 1
cluster 2
centroids
0.8
0.7
0.6
0.5
0.4
Feature 1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Feature 0

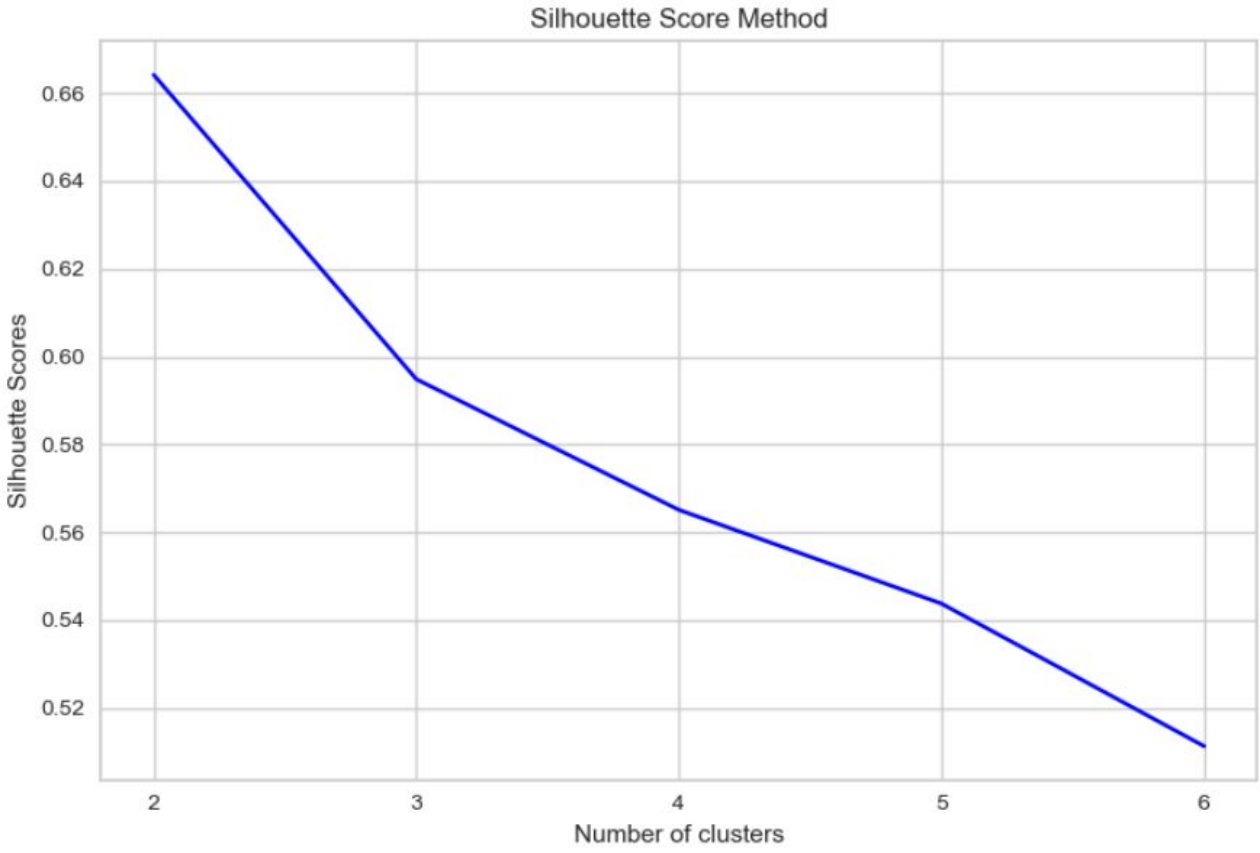
Dendograma



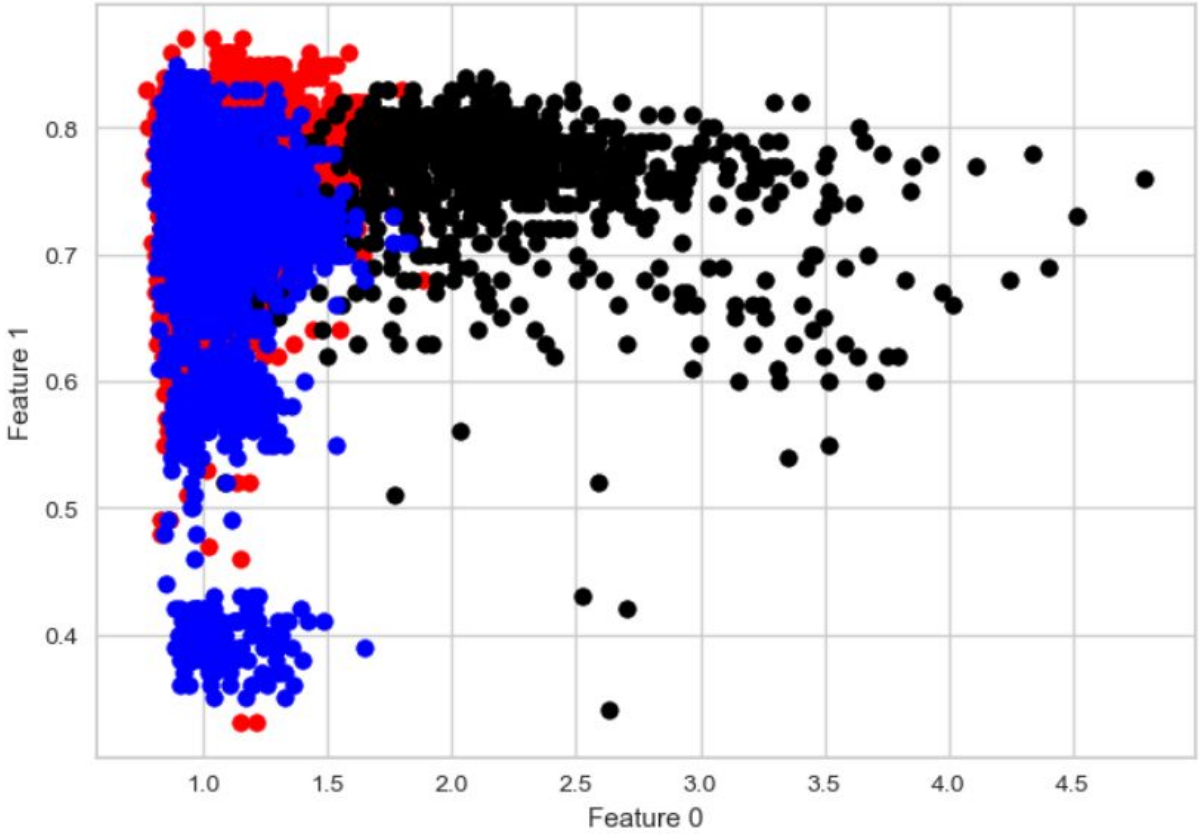
Elbow


Distortion Score Elbow for KMeans Clustering
elbow at $k = 3$, $score = 6762.197$
distortion score
16000
14000
12000
10000
8000
6000
4000
2000
fit time (seconds)
0.10
0.09
0.08
0.07
0.06
0.05
0.04
2
3
4
5
6
k

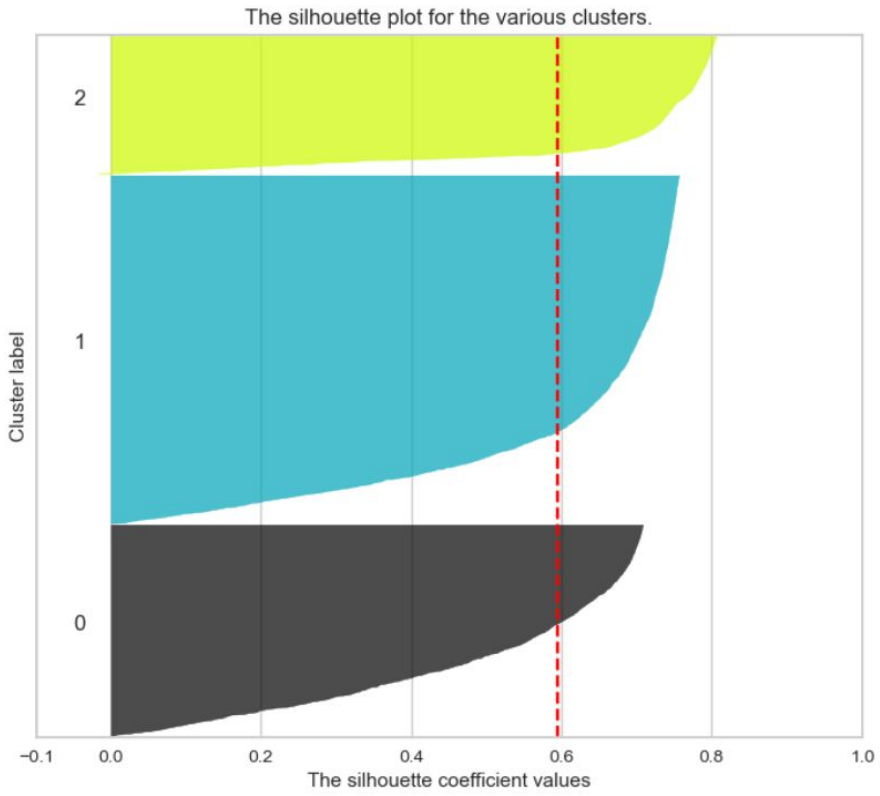
Silhouette

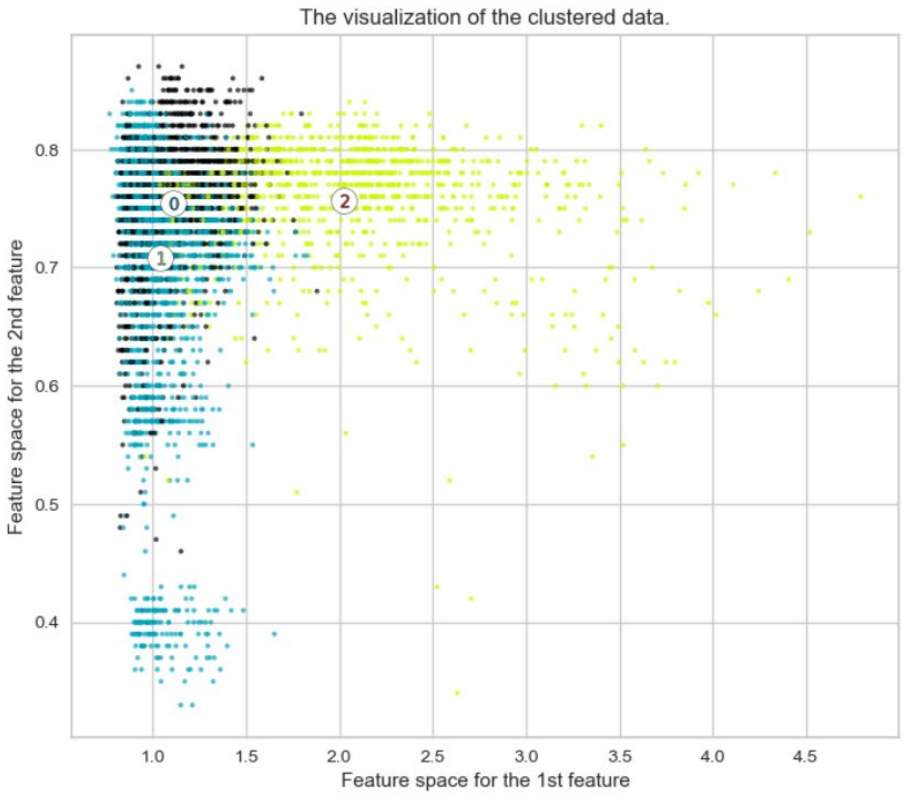

Silhouette Score Method
Silhouette Scores
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
2
3
4
5
6
Number of clusters

Cluster

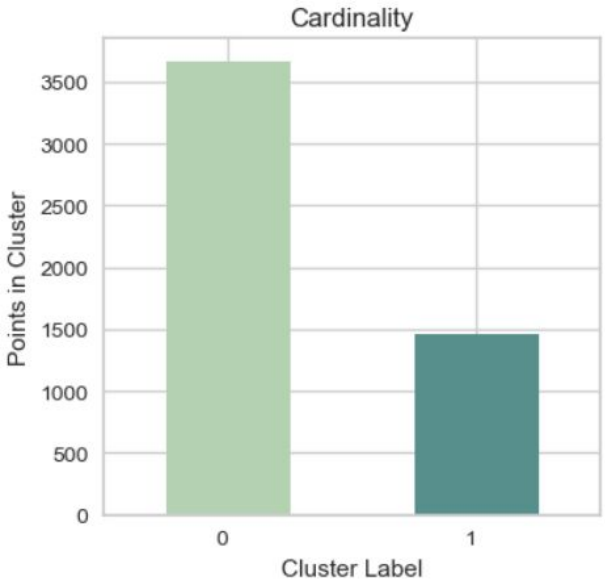


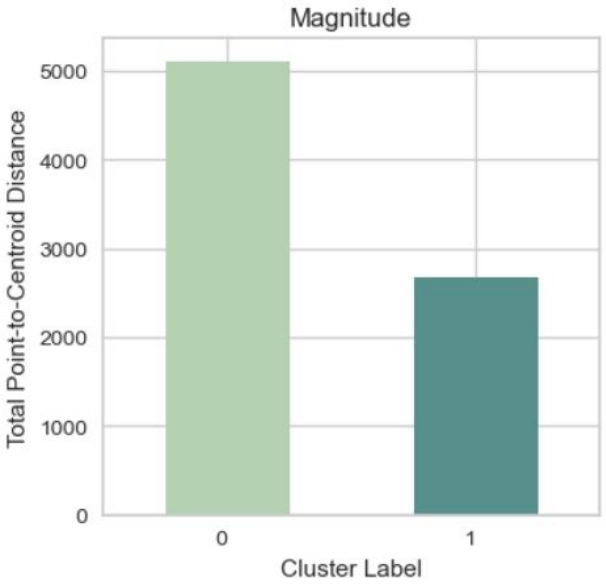
Silhouette

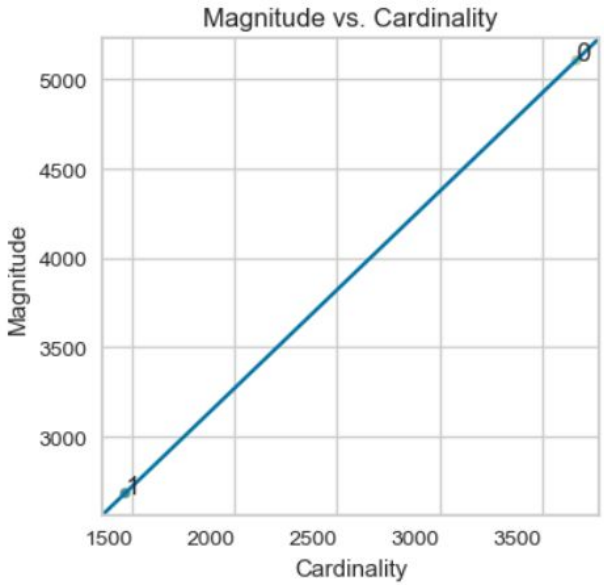

The silhouette plot for the various clusters.
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


The visualization of the clustered data.
Feature space for the 2nd feature
0.8
0.7
0.6
0.5
0.4
0
1
2
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Feature space for the 1st feature

Evaluación

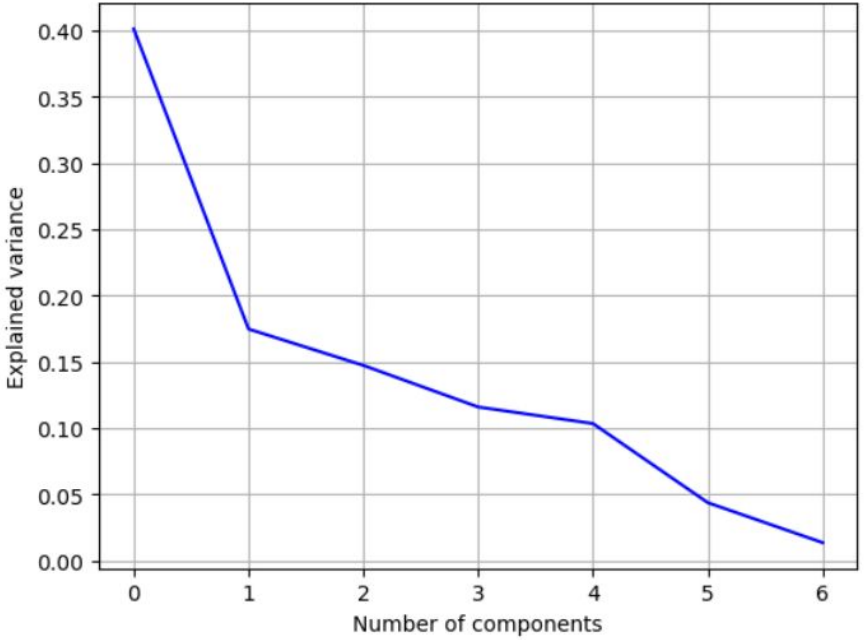

Cardinality
Points in Cluster
3500
3000
2500
2000
1500
1000
500
0
0
1
Cluster Label

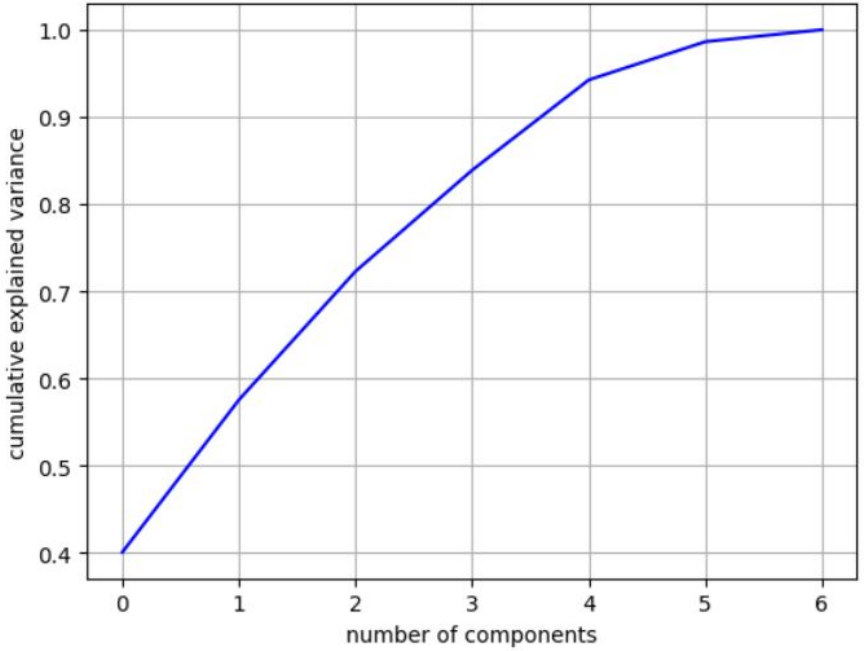

Magnitude
Total Point-to-Centroid Distance
5000
4000
3000
2000
1000
0
0
1
Cluster Label


Magnitude vs. Cardinality
Magnitude
5000
4500
4000
3500
3000
0
1
1500
2000
2500
3000
3500
Cardinality

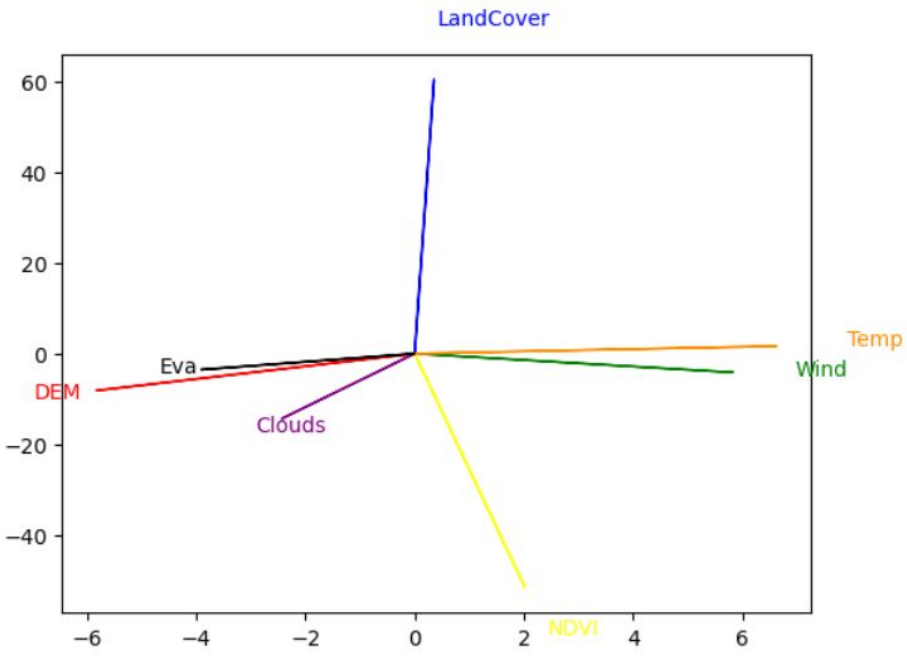
8. Componentes Principales

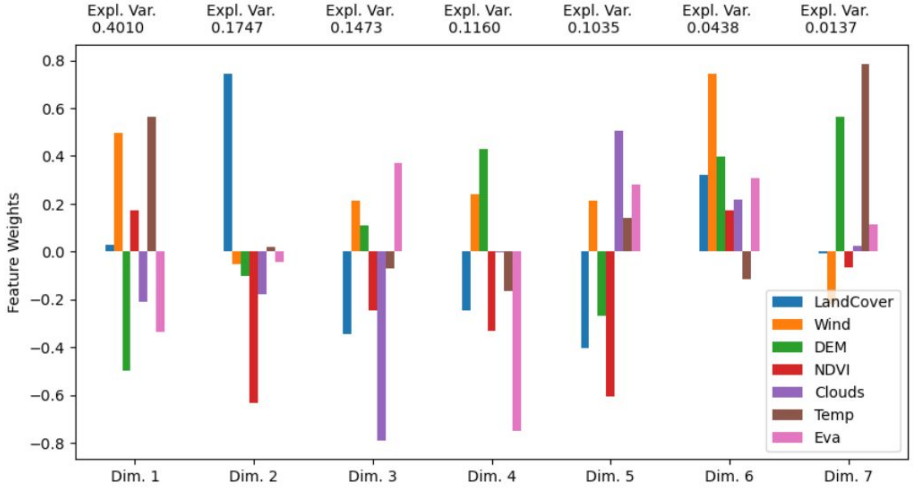
ACP


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Explained variance
0
1
2
3
4
5
6
Number of components


1.0
0.9
0.8
0.7
0.6
0.5
0.4
cumulative explained variance
0
1
2
3
4
5
6
number of components

ACP


LandCover
Temp
Wind
Eva
DEM
Clouds
NDVI
60
40
20
0
−20
−40
−6
−4
−2
0
2
4
6


Expl. Var. 0.4010
Expl. Var. 0.1747
Expl. Var. 0.1473
Expl. Var. 0.1160
Expl. Var. 0.1035
Expl. Var. 0.0438
Expl. Var. 0.0137
Feature Weights
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Dim. 1
Dim. 2
Dim. 3
Dim. 4
Dim. 5
Dim. 6
Dim. 7
LandCover
Wind
DEM
NDVI
Clouds
Temp
Eva

Muchas Gracias