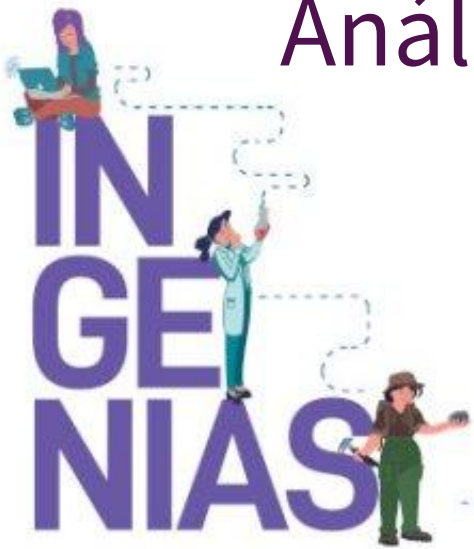
FUNDACIÓN YPF

Trabajo Final Ingenias + Fundación YPF: Análisis del Cambio Climático en Ambientes Marinos.



Juliana Mattei - Azul Senn

Contenidos

1) Introducción:
Presentación del problema, del dataset a utilizar y del objetivo del proyecto.

2) Recolección de datos:
Análisis del dataset “Shifting Seas: Ocean Climate & Marine Life Dataset.”

3) Exploración y Procesamiento:
Correlación entre variables y tratamiento de outliers.

4) Modelo supervisado:
Algoritmos para Clasificación de Bleaching Severity.

5) Modelo no Supervisado:
Clustering con K-means y DBSCAN

Introducción

A medida que el cambio climático se acelera, los océanos del mundo experimentan transformaciones significativas. El dataset **“Shifting Seas: Ocean Climate & Marine Life Dataset”** recopila mediciones sintéticas, pero realistas, de la temperatura superficial del mar (TSM), los niveles de pH, la gravedad del blanqueamiento de corales y observaciones de especies en zonas marinas ecológicamente críticas. Abarca el período de **2015 a 2023** y simula cómo responden los entornos marinos al calentamiento global, la acidificación y las olas de calor.

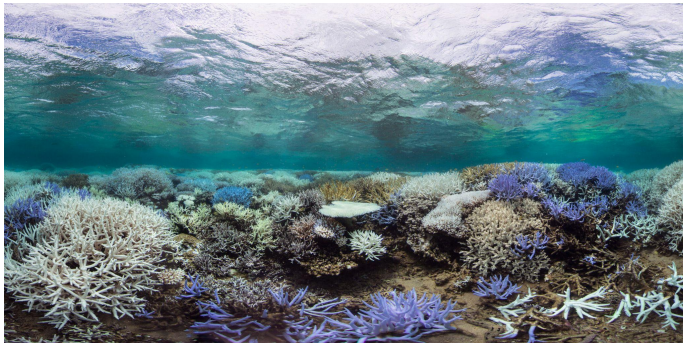
El **objetivo** de este trabajo es analizar cuáles de estas variables climáticas han tenido mayor variabilidad a lo largo del tiempo y si el cambio en las condiciones del medio tiene o no una correlación con los cambios en los registros biológicos, dependiendo del lugar en que se realizan las simulaciones.

Bleaching

El **bleaching** es un fenómeno en el que los corales expulsan zooxantelas, algas microscópicas que viven en sus tejidos y que les proporcionan color y alimento a través de la fotosíntesis.

Cuando los corales están estresados, ya sea por altas temperaturas, contaminación o cambios en la química del agua, sucede este fenómeno, pasando a estar blancos, lo que los hace más vulnerables a la mortalidad.

Arrecife de Coral

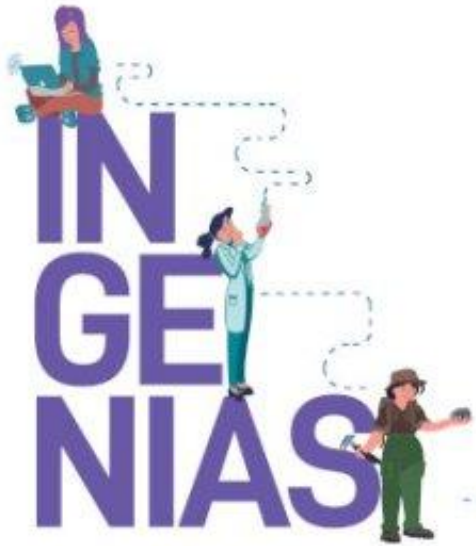


Blanqueamiento de un Arrecife de Coral

Recolección de datos

FUNDACIÓN YPF

¿Qué tenemos en las columnas de nuestro Dataset?



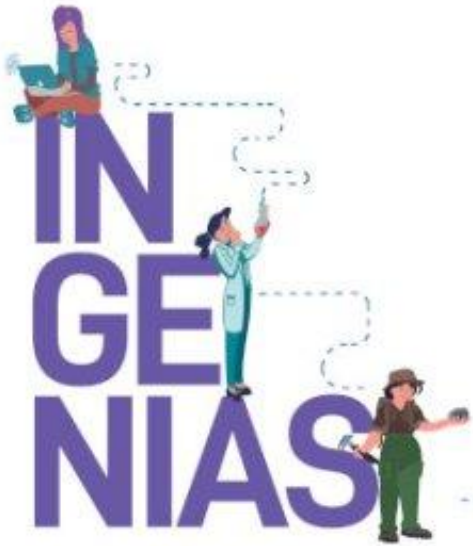

```
df.head(10)
```

	Date	Location	Latitude	Longitude	SST (°C)	pH Level	Bleaching Severity	Species Observed	Marine Heatwave
0	2015-01-01	Red Sea	20.0248	38.4931	29.47	8.107	NaN	106	False
1	2015-01-07	Great Barrier Reef	-18.2988	147.7782	29.65	8.004	High	116	False
2	2015-01-14	Caribbean Sea	14.9768	-75.0233	28.86	7.947	High	90	False
3	2015-01-20	Great Barrier Reef	-18.3152	147.6486	28.97	7.995	Medium	94	False
4	2015-01-27	Galápagos	-0.8805	-90.9769	28.60	7.977	NaN	110	False
5	2015-02-02	Red Sea	20.0055	38.4425	29.06	8.009	Low	109	False
6	2015-02-09	South China Sea	9.9699	115.0926	28.48	7.998	NaN	132	False

Recolección de datos

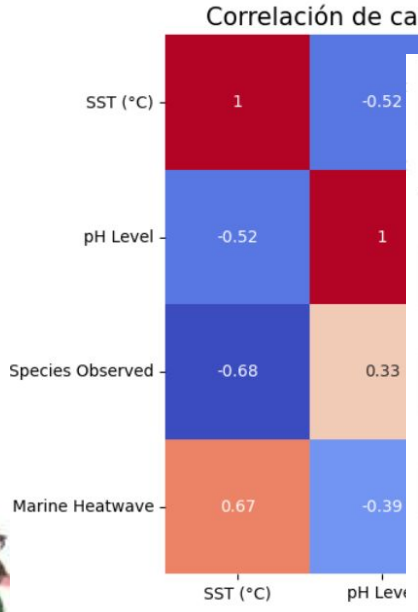
FUNDACIÓN YPF

Mapa de las locaciones presentes en el Dataset



Exploración y Procesamiento

Trabajamos primero con las variables numéricas para ver cómo se correlacionan entre sí y destacar algunas características de cada una.




```
df.describe()
```

	Latitude	Longitude	SST (°C)	pH Level	Species Observed
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	6.728440	11.930638	28.537200	8.049872	120.472000
std	13.534648	108.174778	1.425076	0.055707	20.475441
min	-18.383000	-155.717600	23.640000	7.872000	54.000000
25%	-0.932925	-90.926400	27.530000	8.011000	107.750000
50%	10.018750	38.520000	28.515000	8.052000	120.000000
75%	19.891000	114.990975	29.452500	8.085250	133.250000
max	20.122000	147.827100	33.210000	8.195000	171.000000

Outliers

FUNDACIÓN YPF

```
# Vemos si los outliers por encima del boxplot y su relación con las olas de calor
high_sst = df[df['SST (°C)'] > df['SST (°C)'].quantile(0.95)]
print(high_sst[['SST (°C)', 'Marine Heatwave']].head())
```

```
     SST (°C)  Marine Heatwave
33       31.35            True
50       31.68            True
77       31.31            True
124      31.32            True
135      31.13            True
```

Verificamos que los outliers de ten

```
Q1_SST = df['SST (°C)'].quantile(0.25)
Q3_SST = df['SST (°C)'].quantile(0.75)
IQR_SST = Q3_SST - Q1_SST
lower_SST = Q1_SST - 1.5 * IQR_SST  # Outliers de pH por debajo de este valor

print(f"Outliers de SST (°C): valores < {lower_SST:.2f}")
```

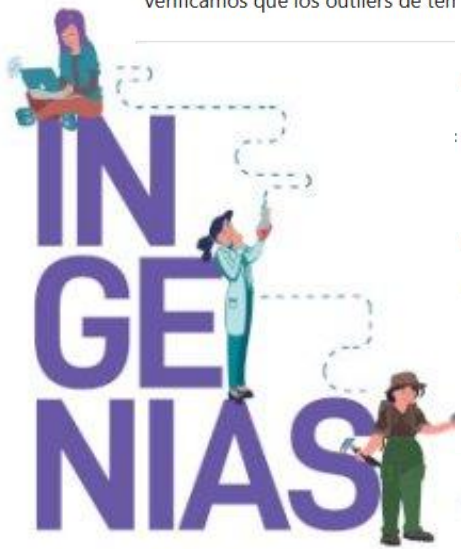
```
Outliers de SST (°C): valores < 24.65
```

```
outliers_SST = df[df['SST (°C)'] < lower_SST]
print("Outliers en SST:", len(outliers_SST))
print(outliers_SST[['Date','Location']])
```

```
Outliers en SST: 1
          Date          Location
52  2015-12-09  Hawaiian Islands
```

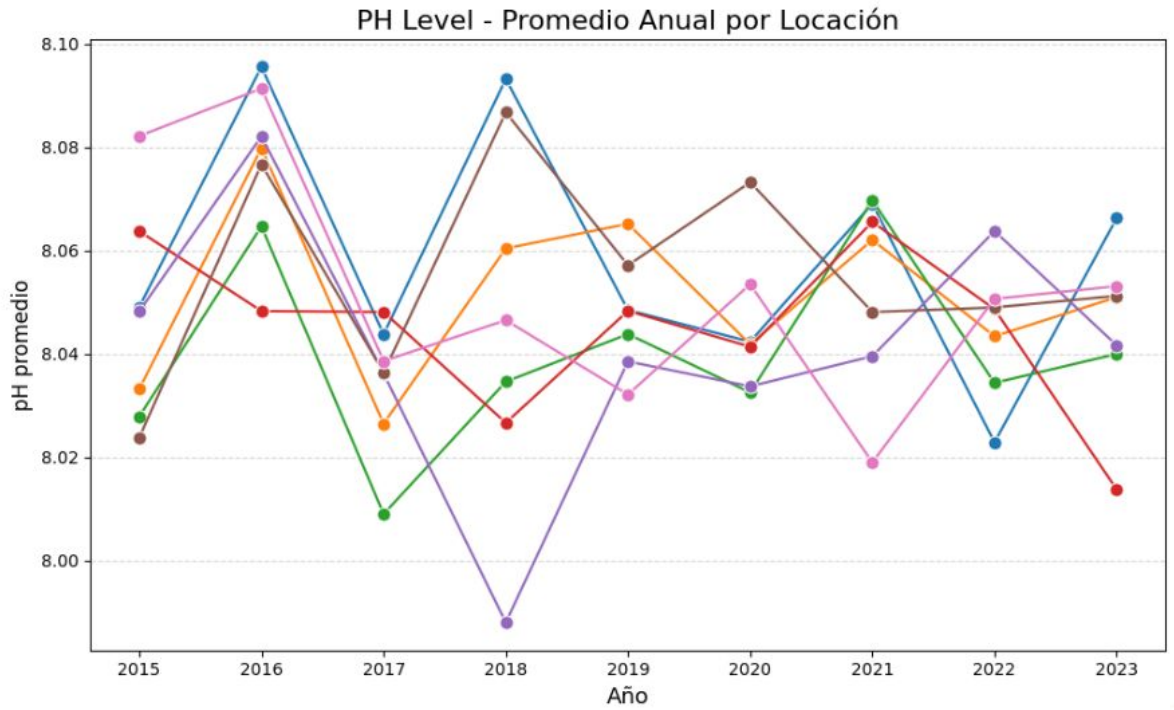
Con respecto al outlier por debajo del boxplot, tenemos un registro en Hawaii correspondiente al invierno en la zona. La temperatura entra dentro del rango normal de la locación.

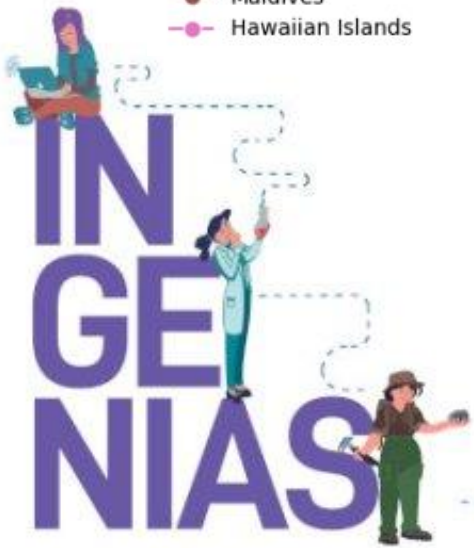
Un pH moderadamente bajo (ácido) podría no ser mortal para especies adaptadas, pero combinado con alta temperatura, sí.



FUNDACIÓN YPF

Otro análisis interesante es considerando la ubicación espacial y temporal de los eventos, ya que el cambio climático no afecta a las regiones de igual manera y se necesita una escala temporal grande para observar las consecuencias.



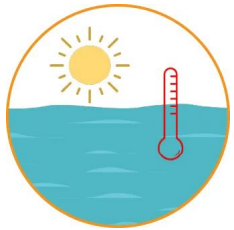


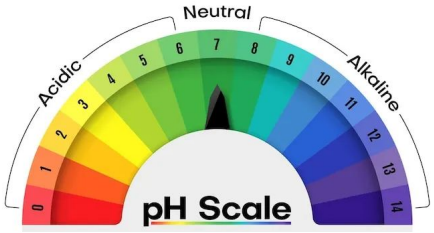
Modelo Supervisado

Para el **Modelo Supervisado** buscamos predecir la pertenencia de nuevas observaciones a las clases de **Bleaching**. Al ser esta una variable categórica utilizamos la metodología de **Clasificación**.

Las **variables predictoras** seleccionadas fueron:

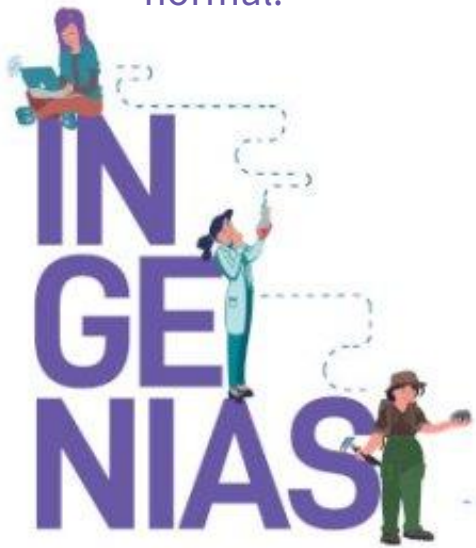
- SST (°C)
- pH Level
- Species Observed
- Marine Heatwave





Modelo Supervisado

Como preparación para el entrenamiento de los modelos, usamos **LabelEncoder** para codificar las clases de la columna Bleaching Severity a valores numéricos y convertimos la columna booleana Marine Heatwave a valores enteros (1-0). También usamos **StandardScaler** sobre estas variables para llevarlas a una distribución normal.



Para la evaluación de desempeño de los modelos se usó la separación 80-20

Modelo Supervisado

Utilizamos los cuatro algoritmos más comunes:

- Random forest
- Árbol de decisión
- k-Nearest Neighbors (kNN)
- Regresión logística

Con cada modelo evaluamos la **matriz de confusión** y los valores de **Accuracy**, **Recall**, **Precision** y **F-1 Score**.

Todos los modelos presentaron matrices de confusión que favorecen notablemente la clase “None” y valores casi idénticos en las restantes métricas, lo que sugiere un comportamiento homogéneo **donde ningún algoritmo logra captar patrones significativos en los datos.**

Modelo Supervisado

Como la cantidad de datos de la muestra es relativamente limitada (500 datos), probamos la validación cruzada para reevaluar los modelos. Utilizamos **StratifiedKFold** ya que encontramos que es mejor que **KFold** puro cuando hay clases desbalanceadas, porque mantiene la proporción de clases en cada fold.

Utilizamos 5 folds y tomamos un promedio para evaluar las matrices de confusión y demás valores.

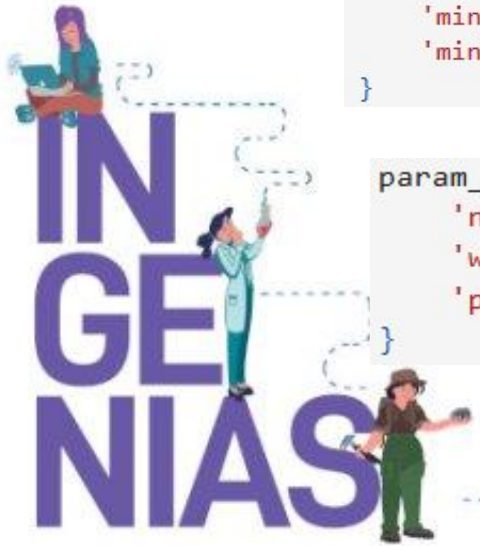
Los resultados no cambiaron. El desbalance se mantuvo con el valor de accuracy aún rondando el 30% y el mismo comportamiento de sobreajuste sobre la clase mayoritaria “None”.

Modelo Supervisado

Con la idea de mejorar el modelado, usamos **GridSearchCV** para optimizar los hiperparámetros, combinado con **StratifiedKFold**.

```
param_grid_rf = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

```
param_grid_dt = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10]
}
```




```
param_grid_knn = {
    'n_neighbors': [3, 5, 7, 9],
    'weights': ['uniform', 'distance'],
    'p': [1, 2]  # 1: Manhattan, 2: Euclidean
}
```

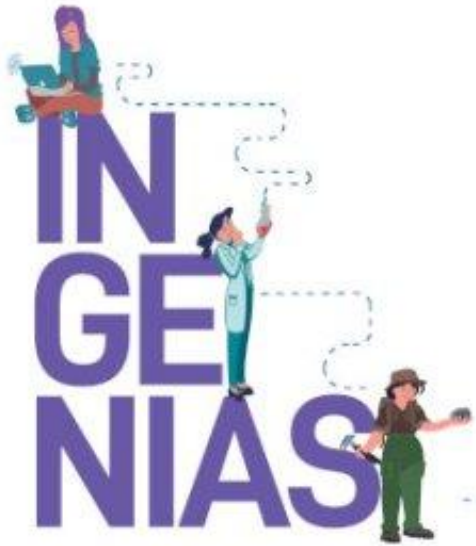
```
param_grid_lr = {
    'C': [0.01, 0.1, 1, 10],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear']
}
```

Los resultados finales no mejoraron notablemente.

Modelo Supervisado

Conclusión:

Si bien todos los modelos fueron afectados por el desbalance de clases, el **Árbol de Decisión ofreció la solución más robusta tras la optimización de hiperparámetros.**



Sin embargo, el **valor de las métricas de todos los modelos fueron muy bajos**, indicando un desempeño deficiente de los mismos.

Sabemos que, en la práctica, las causas del bleaching son la variación de la temperatura y la acidificación del agua, lo cual nos llevó a creer que el problema podría estar directamente en el dato o el preprocesamiento de los mismos.

Modelo No Supervisado

Realizamos un análisis de Clustering, específicamente **K-Means** y **DBSCAN**, para identificar patrones en variables climáticas oceánicas (como temperatura superficial, pH y biodiversidad), con el fin de explorar posibles agrupaciones naturales en los datos y su relación con eventos de blanqueo de corales.

Para esto debimos hacer una serie de transformaciones y normalizar nuestras variables además de elegir en cuántas componentes principales proyectar los datos.

Modelo No Supervisado

FUNDACIÓN YPF

Convertir columnas necesarias

```
df["Marine Heatwave"] = df["Marine Heatwave"].astype(int)
df['Bleaching Severity'] = df['Bleaching Severity'].replace({np.nan: 'None'})
```

Selección de características para clustering

```
features = ["Latitude", "Longitude", "SST (°C)", "pH Level", "Species Observed", "Marine Heatwave"]
X = df[features]
```

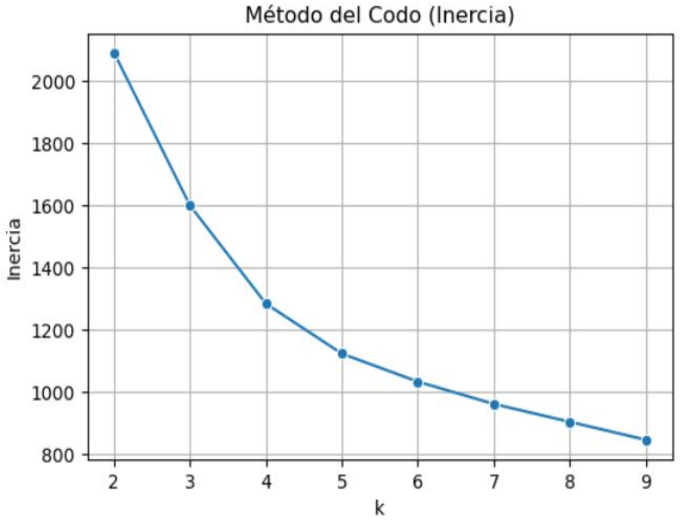
Normalización de las variables

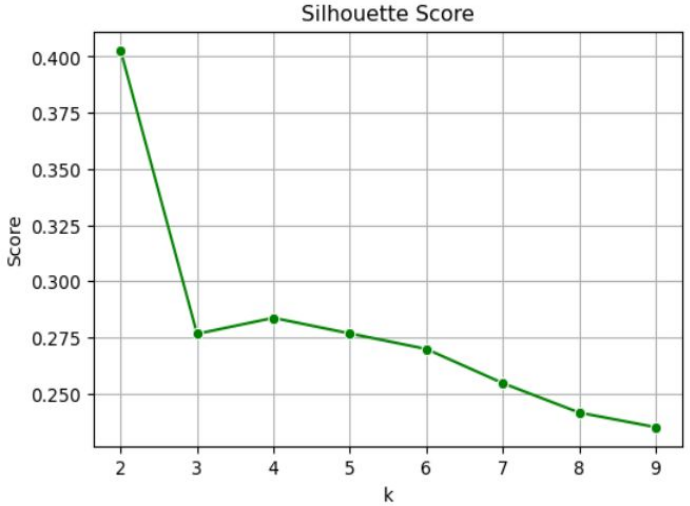
```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

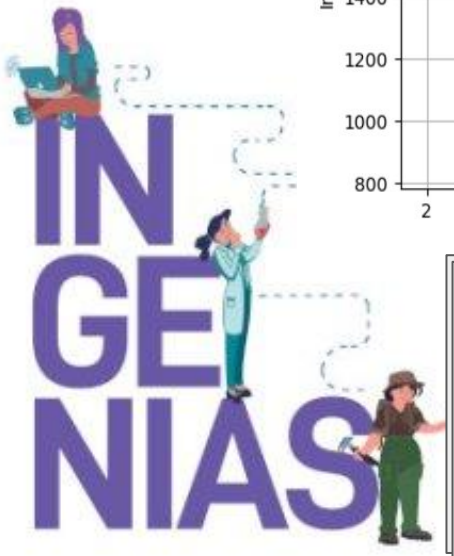
Reducción de dimensionalidad para visualización

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
```

Modelo No Supervisado - K-means



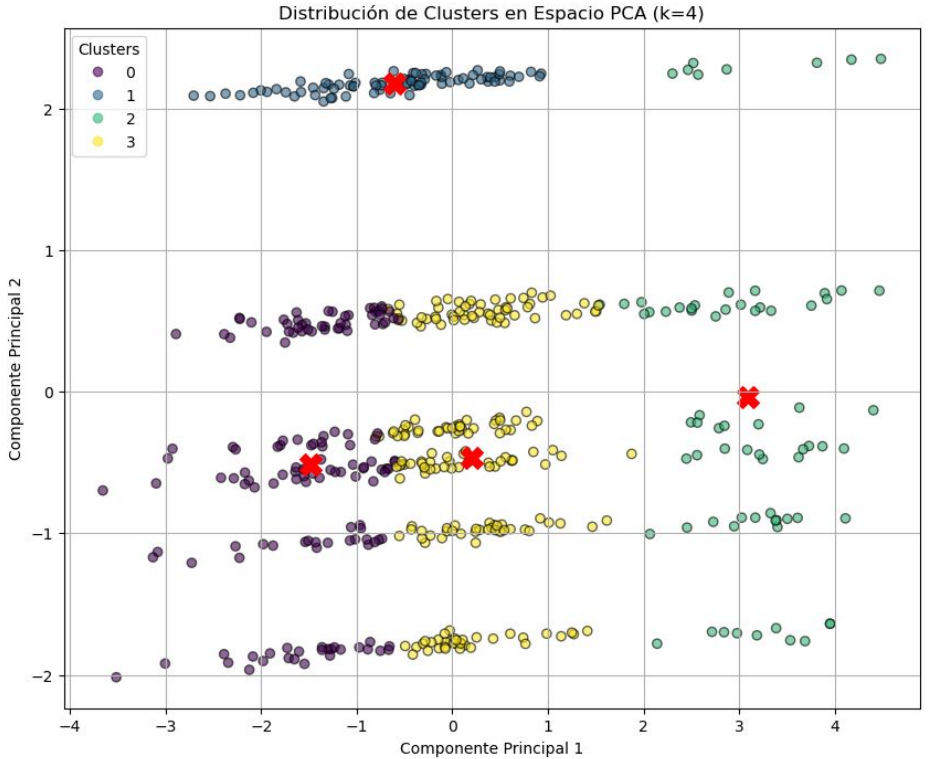




Sin embargo, como el objetivo del análisis requiere distinguir los tipos de blanqueo coralino, k = 4 es la opción seleccionada.

Modelo No Supervisado - K-means

¿Qué significa cada componente principal?



Modelo No Supervisado - K-means

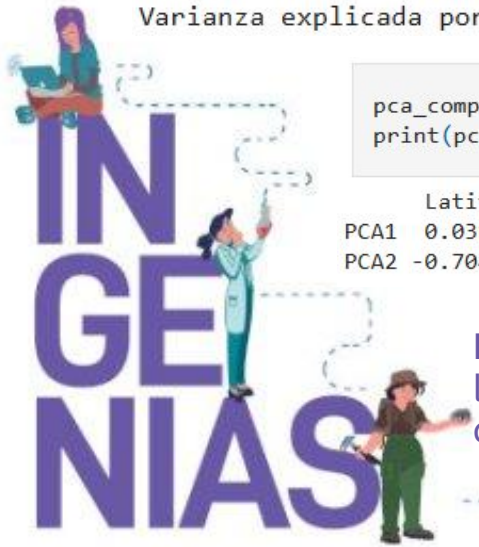
FUNDACIÓN YPF

Antes de continuar con el análisis, es importante notar cuánta información retienen los componentes PCA y además entender qué variables contribuyen a cada componente, para esto observamos lo siguiente:

```
print("Varianza explicada por PCA:", pca.explained_variance_ratio_)
```

```
Varianza explicada por PCA: [0.42480302 0.26010605]
```

Estos valores indican que PCA1 explica 42% de la varianza y PCA2, el 26%.

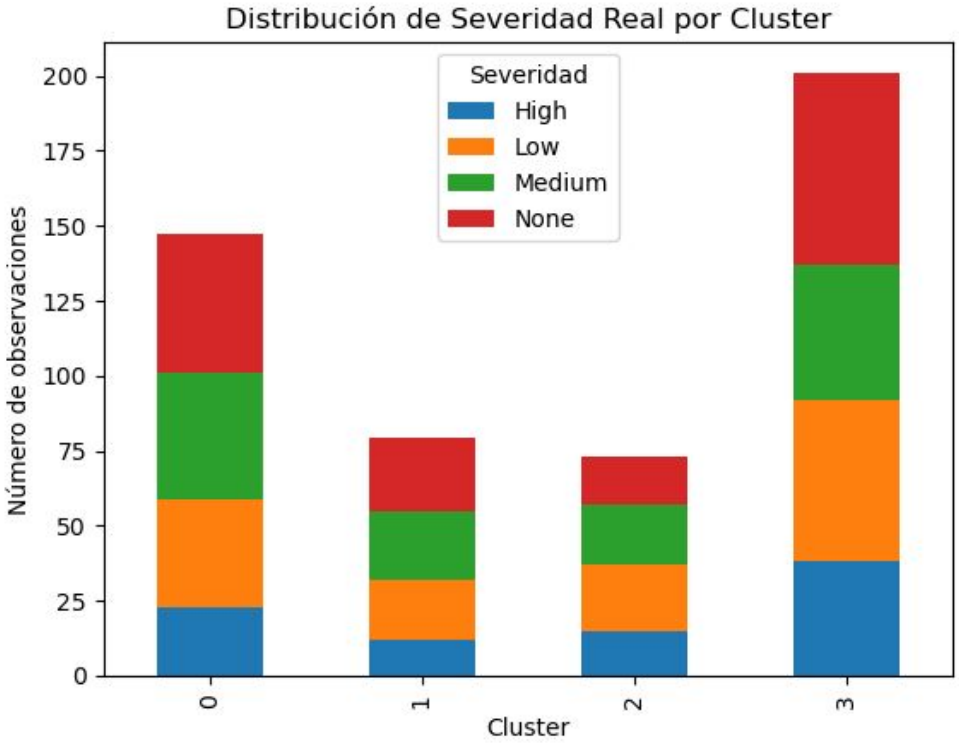



```
pca_components = pd.DataFrame(pca.components_, columns=features, index=['PCA1', 'PCA2'])
print(pca_components)
```

	Latitude	Longitude	SST (°C)	pH Level	Species Observed	Marine Heatwave
PCA1	0.038808	-0.012592	0.573136	-0.420974	-0.490851	0.501694
PCA2	-0.704071	0.707076	0.050454	0.015212	-0.038113	-0.009955

PCA1 está mayormente influenciado por SST, el PH, las especies observadas y las olas de calor marina (refiere exclusivamente a factores climáticos) mientras que **PCA2** incluye principalmente la ubicación de los datos.

Modelo No Supervisado - K-means



Modelo No Supervisado - K-Means - Clusters

FUNDACIÓN YPF

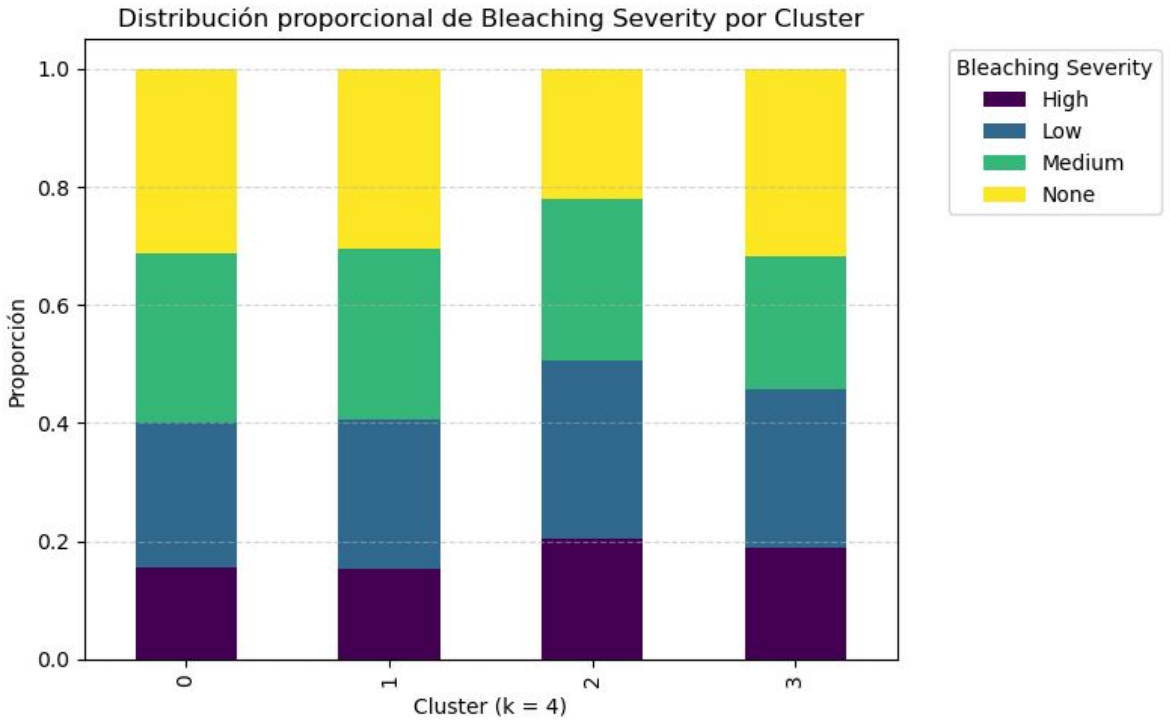
Cluster 0:
SST: 27.2°C, pH: 8.09, especies: 139, Marine Heatwave: No
Regiones relativamente estables, biodiversas y menos estresada térmicamente. Coindice con un alto % de observaciones con “None” (31%) y la menor proporción de “High” (15.6%).

Cluster 1:
SST: 28.2°C, pH: 8.06, especies: 125.5, Marine Heatwave: No
Regiones algo más cálidas, sin eventos extremos, biodiversidad media. Bleaching con proporciones similares (30% “None”).

Cluster 2:
SST: 30.85°C, pH: 7.99, especies: 97.6, Marine Heatwave: Si
Regiones con alto estrés térmico y químico, baja biodiversidad y presencia de eventos extremos.
Las proporciones de “High” (20.5%) y Low (30.1%) son las más altas. **Posiblemente sea una zona de transición o adaptabilidad, pero en riesgo.**

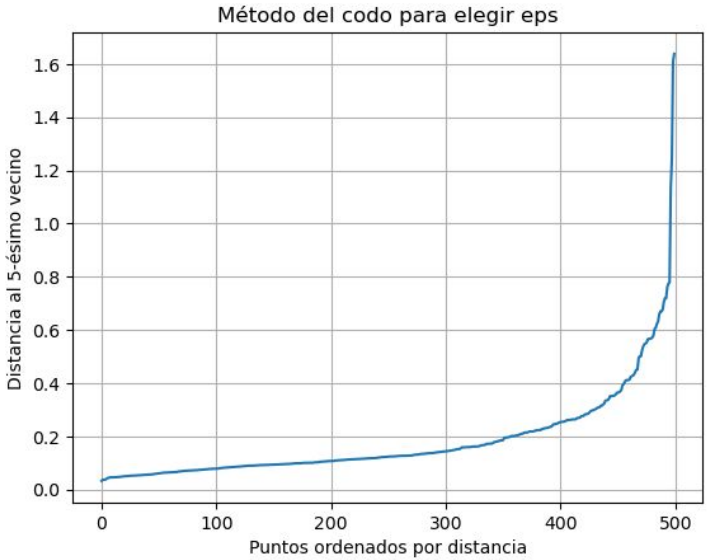
Cluster 3:
SST: 28.8°C, pH: 8.03, especies: 113, Marine Heatwave: No
Regiones con condiciones cálida y biodiversidad moderada.
Tiene el mayor número de casos “High” y “None” (31.8%).
Es un cluster intermedio y numeroso, mixto en severidad.

Modelo No Supervisado - K-means



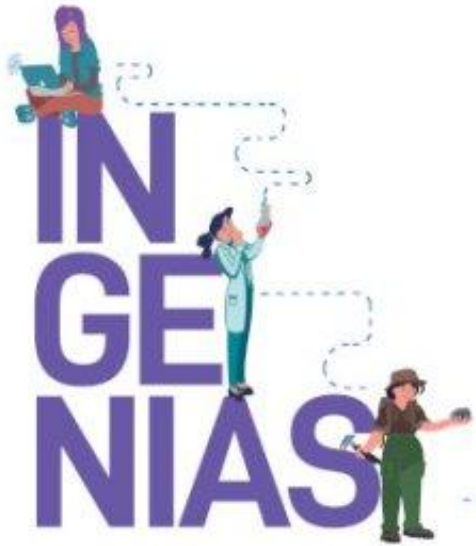
Modelo No Supervisado - DBSCAN

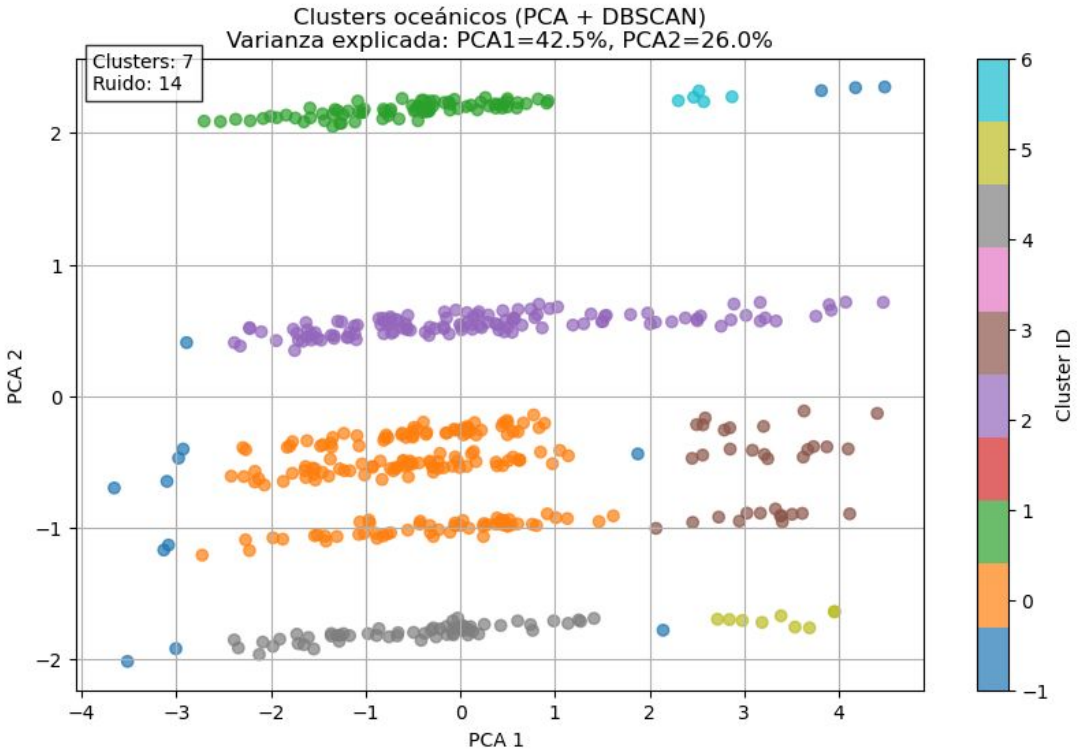
Un paso importante para utilizar **DBSCAN** es la elección del eps óptimo. Un eps demasiado pequeño clasificará todo como ruido, mientras que uno demasiado grande unirá todos los puntos en un solo cluster. Para esto podemos emplear una gráfica de distancia a los k-vecinos más cercanos (**Elbow Method**).



Modelo No Supervisado - DBSCAN

¿Qué representa cada cluster?





Modelo No Supervisado - DBSCAN

FUNDACIÓN YPF

¿Qué significa el cluster 5 (alto valor 'High')?

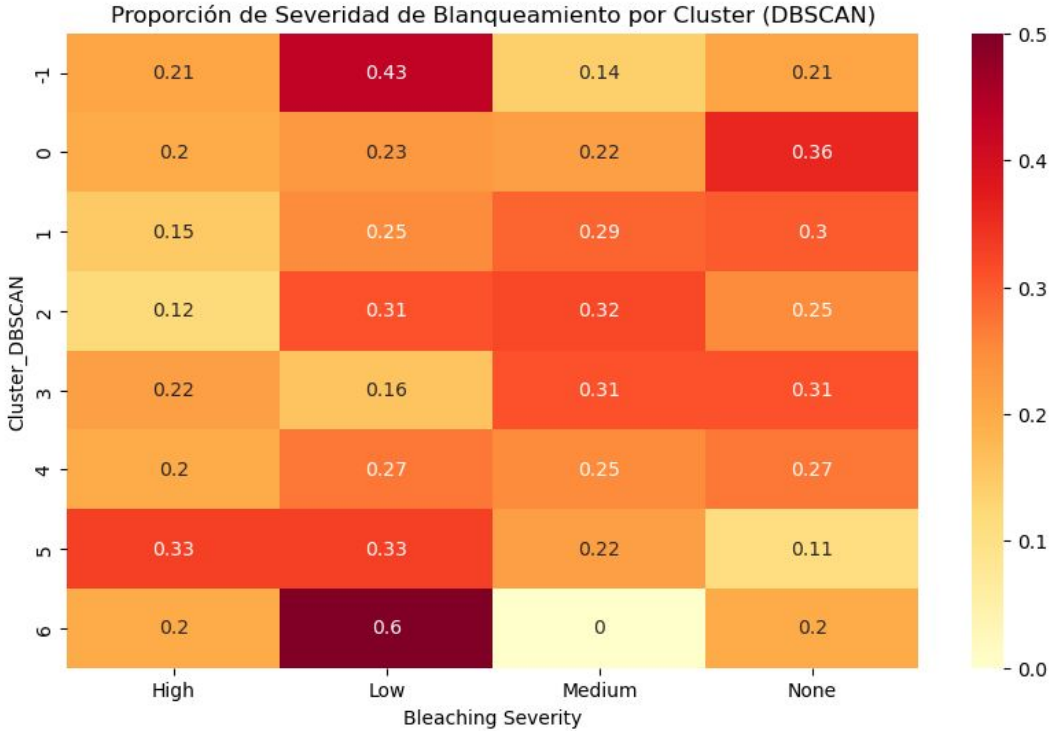


Tabla cruzada entre clusters y severidad de blanqueamiento

Modelo No Supervisado - DBSCAN

FUNDACIÓN YPF

Análisis de Resultados

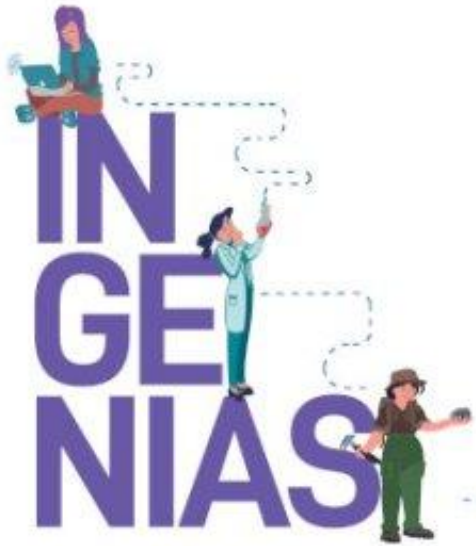
Como vemos, con **SST alta (>30°C) y pH relativamente bajo** tenemos una **alta correlación con blanqueamiento**, por lo tanto, sumado a que **Marine Heatwave = 1** (lo cual confirma que hay eventos de estrés térmico), vemos que este cluster es crucial para el análisis.

Nos indica que en particular, en las zonas ubicadas a la latitud y longitud mencionadas (**el área coincide con arrecifes costeros de Hawái**, expuestos a estrés térmico y actividad humana) hay problemas ambientales.

Posibles causas: Combinación de temperaturas extremas, acidificación, o eventos de calor. En este caso, el pH no influye tanto como sí las olas de calor extremas. Sería interesante investigar variables ambientales específicas y su ubicación geográfica para diseñar estrategias de conservación.

Conclusiones

FUNDACIÓN YPF



En resumen, si bien los modelos supervisados no lograron buenos resultados por el desbalance de clases, el análisis no supervisado reveló patrones valiosos.

Identificamos condiciones ambientales asociadas al bleaching, como la combinación de altas temperaturas y eventos extremos.

Este tipo de análisis puede ayudar a identificar zonas de riesgo y contribuir al diseño de estrategias de conservación marina.

FUNDACIÓN YPF

¡Muchas Gracias!