

# PRÁCTICA 3

Caso de estudio: Pronóstico del fenómeno del niño en Puerto Ayora, Galápagos



Minería de Datos, Universidad de Alicante  
En colaboración con la Fundación Charles Darwin  
Àlex Macià, Manuel Peiró, Francesc Bellido y Juan Carlos Valera

# Índice

El Parque natural de las Islas Galápagos	2
El fenómeno del Niño y la Niña	2
El fenómeno del Niño y la Niña en la zona de las Islas Galápagos	3
Preprocesamiento de datos	4
Datos duplicados y valores ausentes	4
Transformaciones	6
Valores anómalos	7
Imputación de valores ausentes	8
Selección y extracción de características	9
Visualización y estadísticas de los datos	10
Clasificación	15
Pasos previos a la clasificación	15
Clasificación con el algoritmo SVC ( <i>Support Vector Classifier</i> )	16
Clasificación con árboles de decisión.	18
Clasificación con el algoritmo <i>LogisticRegression</i>	20
Predicción, dado un rango temporal, para pronosticar el fenómeno en el año 2023	23
Resultados de la clasificación	24
Conclusiones	25
Referencias	25

## **1. El Parque natural de las Islas Galápagos**

El Parque Nacional Galápagos es un parque natural ubicado en las islas Galápagos. Forma parte de un archipiélago situado en la costa de Ecuador en Sudamérica. Está formado por 13 islas y se considera una de las zonas volcánicas con más actividad en el mundo. Fue el primer parque nacional de Ecuador siendo establecido en 1959. En 1978, las islas Galápagos fueron declaradas Patrimonio de la Humanidad por la UNESCO y, desde entonces, se han convertido en un importante destino turístico y de investigación científica.

El parque protege una gran cantidad de flora y fauna única, incluyendo especies de animales como los icónicos pinzones de Darwin y las tortugas gigantes de las Galápagos. Desde su establecimiento, el parque ha trabajado en la protección de estas especies y en la preservación de su hábitat natural, pero debido a la afluencia turística y el cambio climático, se ha visto afectado en las últimas décadas. Actualmente, varias asociaciones trabajan en proyectos de conservación y restauración ambiental en el archipiélago. Entre otras muchas actividades, se está trabajando en proyectos que incluyen la restauración de hábitats naturales, la eliminación de especies invasoras y la educación y sensibilización sobre la importancia de la conservación ambiental. Además, en el Parque se trabaja en estrecha colaboración con la comunidad local para fomentar la sostenibilidad y reducir el impacto negativo del turismo en las islas. [1]

## **2. El fenómeno del Niño y la Niña**

El Niño y la Niña representan un fenómeno meteorológico que condiciona el clima a través de cambios en la temperatura de los océanos.

El fenómeno del Niño (FEN), conocido también por sus siglas en inglés ENSO (El Niño Southern Oscillation), es la mitad cálida y húmeda de este ciclo. En este episodio las superficies de los océanos sufren un calentamiento más alto de lo normal lo que provoca una humedad más alta teniendo como consecuencias lluvias abundantes en las costas e interior, y cambios en los patrones de viento y precipitación. Esto se acentúa en la zona ecuatoriana del Pacífico, así como en la costa Central y del Sur del continente americano. El Niño puede provocar inundaciones en algunas regiones y sequías en otras, ya que se pueden observar patrones de lluvia intensa y tormentas tropicales en el Pacífico Central y Este, mientras que en otras partes del mundo se experimentan condiciones más secas y cálidas de lo normal. [2]

Por otro lado, durante la fase de la Niña, las aguas del Pacífico ecuatorial se enfrian, lo que tiene un efecto contrario en el clima, ya que se produce menor humedad de ambiente y una disminución en las precipitaciones. En el caso del continente americano y más concretamente la zona ecuatorial y sur, se observan los patrones de lluvia más intensos en el Pacífico occidental y partes de Asia. Los fenómenos de la Niña suelen presentarse durante un mayor tiempo de duración, lo que puede llegar a causar largos períodos de sequía en las zonas afectadas. [2]

## 2.1. El fenómeno del Niño y la Niña en la zona de las Islas Galápagos

Para realizar el estudio de las consecuencias que tiene el fenómeno del Niño y La Niña en la zona ecuatorial, y más concretamente en las Islas Galápagos, hemos obtenido datos de dos centros meteorológicos que estudian este fenómeno. El primero es el **Climate.gov**, sitio web perteneciente al departamento del gobierno de Estados Unidos, y que ofrece información del país enfocándose, además, en los fenómenos del resto del mundo. El sitio web está dirigido por la Administración Nacional Oceánica y Atmosférica (NOAA) y fue diseñado para proporcionar información accesible y fácil de entender sobre el clima a una amplia variedad de usuarios, incluyendo científicos, educadores, periodistas, responsables políticos y el público en general, además de trabajar de otros departamentos similares a nivel mundial. Se centra en presentar información actualizada sobre los patrones climáticos, el cambio climático, los eventos climáticos extremos y los impactos del clima extremo en la sociedad y el medio ambiente [3].

El segundo departamento al que nos hemos dirigido para la búsqueda de información ha sido el **NIWA Taihoro Nukurangi**, equipo de investigación de Nueva Zelanda y que pertenece a la empresa National Institute of Water and Atmospheric Research (NIWA). Este equipo se centra en el desarrollo de sistemas de predicción climática a largo plazo, así como en la realización de investigaciones para comprender mejor los patrones climáticos del Pacífico Sur. Trabajan en colaboración con otras organizaciones y comunidades para adaptarse y responder a los impactos del cambio climático en Nueva Zelanda y en la región del Pacífico Central y Sur [4].

A través de estos dos departamentos, hemos obtenido la serie de acción del Niño y La Niña desde el año 1972 hasta el 2022. Estos períodos vienen dados en las estaciones meteorológicas de Nueva Zelanda que es la que nos ofrece la información, siendo adaptados a nuestra documentación. Estos períodos son los siguientes:

- Del 1 de diciembre al 28 de febrero se considera verano.
- Del 1 de marzo al 31 de mayo se considera otoño.
- Del 1 de junio al 31 de agosto se considera invierno.
- Del 1 de septiembre al 30 de noviembre se considera primavera.

Estos períodos se han incluido en el Excel [climate\\_puerto-ayora-1964-2023.xlsx](#), creando la columna **fenomeno**. Esta columna será la que se utilice para obtener modelos y predicciones sobre los efectos del Niño y la Niña. En un principio dicha columna tiene tres valores:

- El Niño, el cual se sitúa en los días que estuvo presente este fenómeno.
- La Niña, la cual está situada en los días que estuvo presente este fenómeno.
- (-). Se sitúa en los días que no se consideran de ninguna de las dos anteriores opciones, momentos de transición entre los dos fenómenos o en situaciones del Niño o la Niña que transcurrieron de forma muy leve.

Hemos realizado la adaptación de nuestros datos a la información que ofrecen estas dos organizaciones ya que, según sus investigaciones, sitúan las repercusiones del Niño y la Niña de una forma general en el Océano Pacífico, extendiéndose a la zona Central, Ecuatorial y Pacífico Sur, y de forma más general al resto del mundo [4].

Así mismo, debido a que no ha sido posible encontrar un histórico del fenómeno meteorológico del Niño y la Niña en ninguna publicación o estamento gubernamental del país de

Ecuador, para poder verificar que la acción de dicho fenómeno en algunos períodos del tiempo que se han encontrado se corresponde con los datos recopilados por el equipo que realiza esta investigación, hemos recurrido a investigar en varias de las publicaciones de información meteorológica del país en organizaciones como la página web de la Cámara Nacional de Pescadería (CNP), del Instituto Oceanográfico y Antártico de la Armada del Ecuador y del CIIFEN (Centro Internacional para la Investigación del Fenómeno El Niño), perteneciente al Centro Regional del Clima para el Oeste de Sudamérica (CRC OSA) [5][6][7].

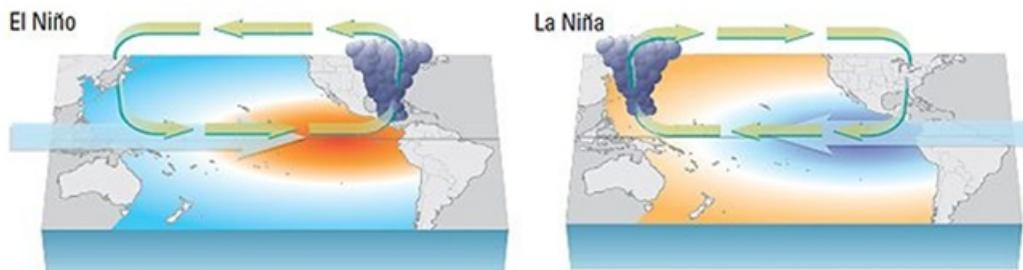


Figura 2.1. Gráfico de El Niño y La Niña en el Océano Pacífico. Fuente: [www.enso.info](http://www.enso.info)

Por último, queremos destacar que los últimos 10 meses de datos de la presencia del fenómeno han sido obtenidos a través de la publicación mensual que realiza CIIFEN sobre el fenómeno del Niño, desde su página web [ciifen.org](http://ciifen.org).

### 3. Preprocesamiento de datos

Para la limpieza y tratamiento de los datos, el grupo de análisis ha decidido trabajar de forma colaborativa en *Kaggle*, utilizando esta plataforma y el lenguaje de programación Python para el tratamiento y visualización de los datos. Se realizará este análisis de manera privada y posteriormente se podrá hacer público una vez que esté terminado, utilizando *GitHub* como una de las plataformas para visualizar los resultados obtenidos.

#### 3.1. Datos duplicados y valores ausentes

Para agregar valor a los datos, en primer lugar, vamos a especificar el significado de las variables de cada columna del conjunto de datos:

- **fenomeno**: Fenómeno del Niño. Sus posibles valores son 'El Niño', 'La Niña' y '-'. Este último posible valor indica que es un día de transición entre ambos fenómenos o una situación donde uno de los valores ocurre de forma muy leve.
- **observation\_date**: Fecha de observación.
- **min\_air\_temp**: Temperatura del aire mínima.
- **max\_air\_temp**: Temperatura del aire máxima.
- **mean\_air\_temp**: Temperatura del aire promedio.
- **sea\_temp**: Temperatura del mar.

- **humidity**: Humedad.
- **precipitation**: Precipitación.
- **sunshine\_hours**: Horas de sol.
- **clouds**: Cantidad de nubes.

En la **Figura 3.1** podemos ver una muestra de los datos obtenida del *dataset* sobre un grupo de 10 filas:

fenomeno	observation_date	min_air_temp	max_air_temp	mean_air_temp	sea_temp	humidity	precipitation	sunshine_hours	clouds
El Niño	01/01/1972	21,20	25,50	23,70	22,50	85,00	0,00		3,00
El Niño	02/01/1972	20,90	26,60	24,00	22,80	82,00	0,00		5,00
El Niño	03/01/1972	21,20	27,20	24,20	22,00	82,00	0,00		5,00
El Niño	04/01/1972	20,70	26,80	24,00	22,30	84,00	0,00		4,00
El Niño	05/01/1972	21,30	26,80	24,50	22,90	85,00	0,00		4,00
El Niño	06/01/1972	22,00		25,10	23,20	86,00	0,00		4,00
El Niño	07/01/1972	21,50		25,10	22,80	84,00	0,00		3,00
El Niño	08/01/1972	22,00		24,30	22,50	89,00	0,00		2,00
El Niño	09/01/1972	20,00	27,00		23,20		4,40		

Figura 3.1. Muestra de 10 filas de los datos proporcionados, donde se incluye la columna fenómeno.

Como primera observación, vemos fechas donde el valor de *sunshine\_hours* se presenta ausente. Este punto será uno de los temas a estudiar en el preprocessado de los datos. En primer lugar, se va a comprobar, si existen filas duplicadas o valores ausentes y posteriormente se aplicarán las transformaciones y filtros que puedan llegar a ser necesarios. No es el caso de duplicación de filas, pero sí que existen valores nulos. Podemos ver en la **Figura 3.2** el resultado obtenido de la comprobación.

fenomeno	0
observation_date	0
min_air_temp	362
max_air_temp	936
mean_air_temp	856
sea_temp	254
humidity	865
precipitation	0
sunshine_hours	11869
clouds	812
dtype: int64	

Figura 3.2. Valores nulos en los datos en el primer momento.

Como podemos ver, tenemos múltiples columnas con valores nulos. La cantidad de valores nulos de ***min\_air\_temp***, ***max\_air\_temp***, ***mean\_air\_temp***, ***sea\_temp***, ***humidity*** y ***clouds*** es menor de 1.000, lo que en números absolutos podríamos considerarlo anecdótico. El principal problema lo encontramos en la columna ***sunshine\_hours*** que presenta más de 10.000 valores nulos. Esta columna empezó a medirse en el año 2004 por lo que no tenemos información en algo más de la mitad de los años que recopila el conjunto de datos. Por otro lado, vemos como las columnas ***fenomeno***, ***observation\_date*** y ***precipitation*** no contienen valores

nulos. Para obtener una idea del número de valores nulos sobre el total de muestras se obtiene el porcentaje de cada columna que podemos ver en la siguiente tabla:

sunshine_hours	63.3723
max_air_temp	4.9976
humidity	4.6185
mean_air_temp	4.5705
clouds	4.3355
min_air_temp	1.9328
sea_temp	1.3562

Figura 3.3. Porcentaje de valores nulos en los datos en el primer momento.

Con estos valores confirmamos que la columna ***sunshine\_hours*** tiene la mayoría de sus filas con valores nulos. A continuación, se comprueba si los valores nulos de las columnas sobre la temperatura del aire coinciden en las mismas filas, obteniendo que los valores nulos registrados para las temperaturas mínimas y máximas del aire coinciden en cerca del 10% de los casos, por lo que en la mayoría de las filas donde la temperatura mínima o máxima del aire contiene valores nulos al menos uno de los dos campos sí contiene información. Para solucionar esta cantidad de nulos se decide realizar más comprobaciones y correcciones de estos valores teniendo en cuenta la correlación de este valor entre las variables ***max\_air\_temp***, ***min\_air\_temp*** y ***mean\_air\_temp***.

Tras las correcciones se logra reducir tanto el porcentaje nulos de ***max\_air\_temp*** como el de ***min\_air\_temp*** sobre el total. En primera instancia ***max\_air\_temp*** presentaba cerca del 5% de valores nulos y ***min\_air\_temp*** cerca del 2% de valores nulos. Ahora tienen el 1.3% y el 0.9% respectivamente. Por lo tanto, si recapitulamos, hemos reducido los porcentajes de nulos en las siguientes cantidades:

- ***mean\_air\_temp***: Se reduce del 4.5% al 1.5%. Una diferencia del 3%.
- ***max\_air\_temp***: Se reduce del 5% al 1.3%. Una diferencia del 3.7%.
- ***min\_air\_temp***: Se reduce del 2% al 0.9%. Una diferencia del 1.1%.

## 3.2. Transformaciones

Para realizar el tratamiento correcto de los datos se han de utilizar los tipos adecuados en cada columna. Para ello, en primer lugar, se comprueban los tipos de datos.

fenomeno	object
observation_date	datetime64[ns]
min_air_temp	float64
max_air_temp	float64
mean_air_temp	float64
sea_temp	float64
humidity	float64
precipitation	float64
sunshine_hours	float64
clouds	float64

Figura 3.4. Tipos de datos.

Como se puede observar en la **Figura 3.4**, únicamente se ha necesitado transformar los datos de la columna ***clouds*** ya que la cantidad de nubes se debe medir por números enteros. Las columnas ***fenomeno*** y ***observation\_date*** contienen el formato adecuado con ***object*** ya que guarda el texto que indica el fenómeno ocurrido y ***datetime64[ns]*** ya que almacena las fechas observadas respectivamente. Las columnas ***min\_air\_temp***, ***max\_air\_temp***, ***mean\_air\_temp***, ***sea\_temp***, ***humidity***, ***precipitation*** y ***sunshine\_hours*** tienen el formato correcto de ***float64*** ya que sus medidas se registran con decimales, con lo que se decide transformar ***clouds*** a ***Int64***.

### 3.3. Valores anómalos

En este subapartado del tratamiento de datos, se eliminan los valores anómalos o erróneos que contiene el conjunto de datos, y para ello, realizan comprobaciones como, por ejemplo, que las temperaturas de cada tipo registradas tengan valores lógicos, que el total de horas de sol no superen las 24 o que todas las fechas observadas sean correctas y se encuentren entre los años 1972 y 2023. Se realiza un estudio para comprobar que todas las temperaturas registradas como máximas son mayores que las medias y las mínimas y que, todas las temperaturas mínimas son menores que las medias y las máximas. Se comprueba que hay una pequeña minoría de filas donde las cantidades indicadas son erróneas. Hemos comprobado que estos datos erróneos se encuentran en el conjunto de datos original y no es un error producido en las operaciones realizadas durante la limpieza y tratamiento. Para solucionar este problema, se decide intercambiar los valores mínimos y máximos de temperatura de aquellas filas donde el mínimo es mayor que el máximo o viceversa.

<b>Temperatura mínima del aire</b>
Valor mínimo: 0.5
Valor máximo: 27.5
Valor medio: 21.1
<b>Temperatura máxima del aire</b>
Valor mínimo: 18.5
Valor máximo: 38.6
Valor medio: 27.0
<b>Temperatura media del aire</b>
Valor mínimo: 18.2
Valor máximo: 29.9
Valor medio: 24.1
<b>Temperatura del mar</b>
Valor mínimo: 15.9
Valor máximo: 29.8
Valor medio: 23.5

Figura 3.5. Rango de los datos de temperatura después del tratamiento.

Como podemos ver en la **Figura 3.5**, el rango de valores tanto de las temperaturas mínimas, máximas y medias del aire como de las temperaturas del mar parece razonable y no presentan valores anómalos después del tratamiento.

También se comprueba que el rango de horas de sol y cantidad de nubes se encuentren dentro de valores razonables. Se obtiene que el valor mínimo de horas de sol es 0 y el máximo es próximo a 12 y el mínimo de cantidad de nubes es 0 y el máximo es 8.

### 3.4. Imputación de valores ausentes

Para el tratamiento de los valores ausentes en las columnas se van a realizar imputaciones de tipo KNN (vecinos más cercanos), con un valor  $K$  igual a 2 vecinos. Como hemos visto en el apartado anterior, las columnas ***sunshine\_hours***, ***humidity***, ***clouds***, ***mean\_air\_temp***, ***sea\_temp***, ***max\_air\_temp*** y ***min\_air\_temp*** son las que presentan valores nulos.

En primer lugar, se imputa en ***sunshine\_hours***, ***min\_air\_temp***, ***max\_air\_temp***, ***mean\_air\_temp*** y ***sea\_temp*** conjuntamente teniendo en cuenta correlaciones entre las variables de horas de luz y temperaturas, y, tras ello, ***humidity*** y ***clouds***, teniendo en cuenta las horas de sol (***sunshine\_hours***) y la temperatura del mar (***sea\_temp***). Hay que tener en cuenta que en la mayoría de las filas las horas de sol no están registradas ya que se empezaron a recopilar en el año 2004 y en nuestro conjunto de datos las fechas empiezan en 1972. En el contexto de este estudio, se opta por imputar las horas de sol en las filas anteriores a 2004 aunque en el proceso de predicción final se tendrá en cuenta que los datos sobre las horas de sol en el rango de años de 1972 a 2004 están en entredicho por haber sido imputadas y, por lo tanto, carecer de una fiabilidad total. Tras la imputación efectuada volvemos a calcular el total de valores nulos:

min_air_temp	0
max_air_temp	0
mean_air_temp	0
sea_temp	0
humidity	0
sunshine_hours	0
clouds	0
dtype:	int64

Figura 3.6. Valores nulos en los datos tras la imputación.

Como resultado, vemos que los datos ya no contienen valores nulos. La cantidad de datos que se han imputado en las columnas de ***min\_air\_temp***, ***max\_air\_temp***, ***mean\_air\_temp***, ***sea\_temp***, ***humidity*** y ***clouds*** es un porcentaje pequeño sobre el total de muestras. Por otro lado, tal como hemos indicado con anterioridad, se imputan los datos de ***sunshine\_hours*** desde 1972 hasta 2004, año donde empezaron a recopilarse las horas de sol diarias, por lo que es una cantidad de filas suficientemente grande como para tener en cuenta su fiabilidad en ese rango de años para el entrenamiento del modelo para la predicción del pronóstico final del fenómeno del Niño.

### 3.5. Selección y extracción de características

En este apartado se centra en la selección y extracción de las características del conjunto de datos. Para ello se aplican diferentes métodos de selección y se extraen finalmente las características que nos sean más útiles para el estudio. En otros trabajos realizados es muy frecuente seleccionar características en base al número de valores nulos presentes en las columnas y, en caso de tener un gran número de características en el conjunto de datos, prescindir de, por ejemplo, las columnas que superen el umbral del 50% de valores nulos. En este caso, únicamente la columna **sunshine\_hours** habría superado ese umbral, pero al tomar la decisión de imputar todos sus valores nulos, hemos corregido ese déficit de información registrada y vamos a mantener la columna para el posterior estudio.

En primer lugar, se comprueba que la varianza de las variables numéricas no supere el valor 1. Ninguna de las columnas numéricas presenta una varianza menor a 1 por lo que no prescindiremos de ninguna columna por razones de varianza mínima.

A continuación, se realiza la selección por correlación para analizar si las columnas numéricas **min\_air\_temp**, **max\_air\_temp**, **mean\_air\_temp**, **sea\_temp**, **humidity**, **precipitation**, **sunshine\_hours** y **clouds** están fuertemente relacionadas.

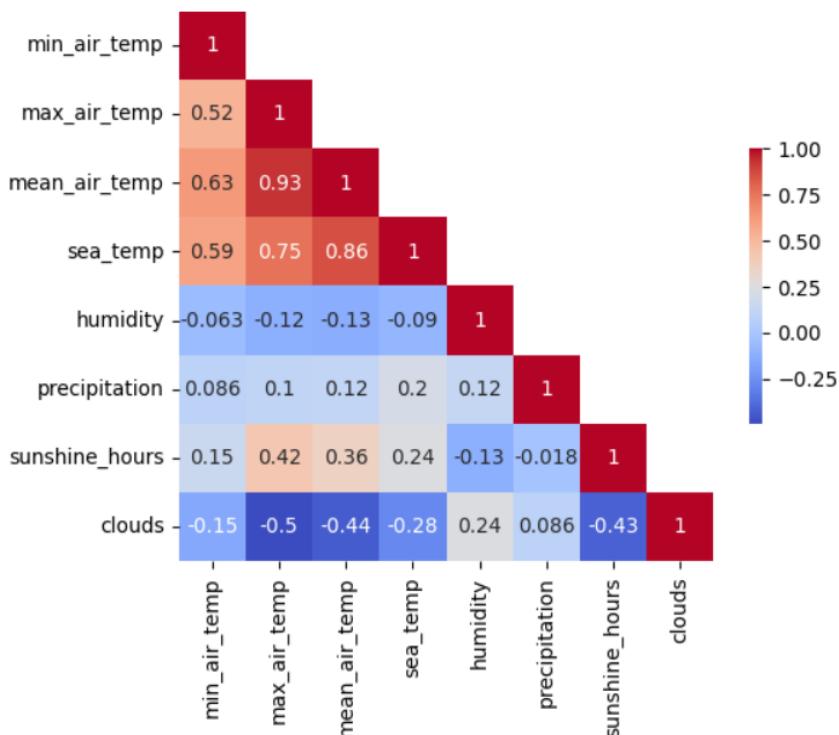


Figura 3.7. Correlación entre las variables numéricas.

En la **Figura 3.7** se puede observar una fuerte correlación entre la temperatura media del aire (**mean\_air\_temp**) y la temperatura máxima del aire (**max\_air\_temp**). A pesar de la fuerte correlación, se decide mantener estas columnas en el conjunto de datos ya que es

información de gran sensibilidad para predecir el fenómeno del Niño. Por otro lado, también existe una correlación fuerte, aunque menor que la anterior, entre la temperatura del mar (**sea\_temp**) y la temperatura media del aire (**mean\_air\_temp**). Por último, aunque no sea una correlación excesivamente fuerte, es destacable la correlación de 0.75 entre la temperatura del mar (**sea\_temp**) y la temperatura máxima del aire (**max\_air\_temp**). Por lo tanto, se toma la resolución de mantener todas las columnas incluidas en el conjunto de datos ya que todas son interesantes para el objeto de estudio de la predicción del fenómeno del Niño y no presenciamos una falta de varianza o una correlación total entre dos columnas como para tomar la decisión de prescindir de alguna de ellas.

## 4. Visualización y estadísticas de los datos

Para identificar tendencias y patrones relacionados con los distintos fenómenos que influyen en “El Niño” y “La Niña”, se han realizado visualizaciones gráficas de series temporales y líneas de tendencia. Como paso previo, la columna **observation\_date** se establecerá como índice del dataset.

Como ejemplo del trabajo realizado, vamos a trasladar a este documento algunas de las visualizaciones efectuadas. En primer lugar, vemos en la **Figura 4.1** la serie temporal de la temperatura media en el ambiente durante los tres fenómenos: “El Niño”, “La Niña” y el periodo de transición.

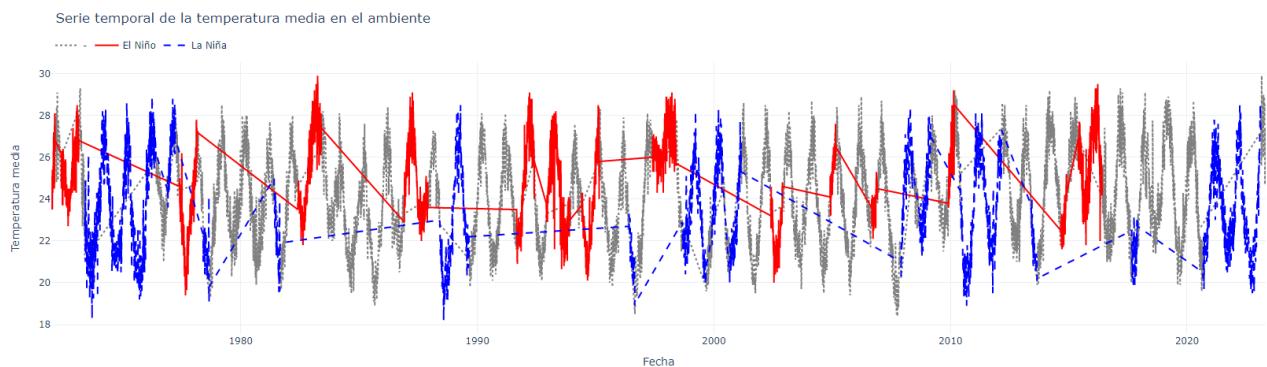


Figura 4.1. Serie temporal de la temperatura media en el ambiente.

En este gráfico, podemos observar la temperatura media en el ambiente durante los distintos fenómenos. Vemos que durante los días en los que se produjo el Niño la temperatura media rondaba entre los 24 y 30 grados, una temperatura superior a la de la Niña que presenta una mayor variabilidad rondando entre los 20 y 28 grados.

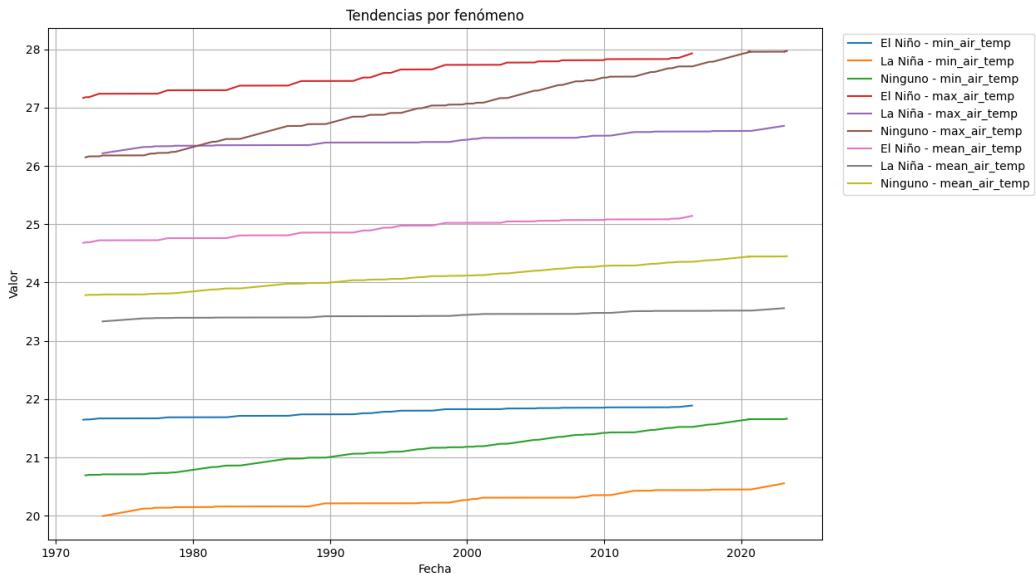


Figura 4.2. Tendencias de temperatura agrupadas por fenómeno.

En la **Figura 4.2** se analizan las líneas de tendencia de la temperatura del ambiente agrupado por fenómenos. Podemos ver como en todas las variables el valor más elevado de la temperatura se da cuando sucede “El Niño” y el más bajo cuando sucede “La Niña”, quedando en un valor intermedio cuando ocurre el periodo de transición.

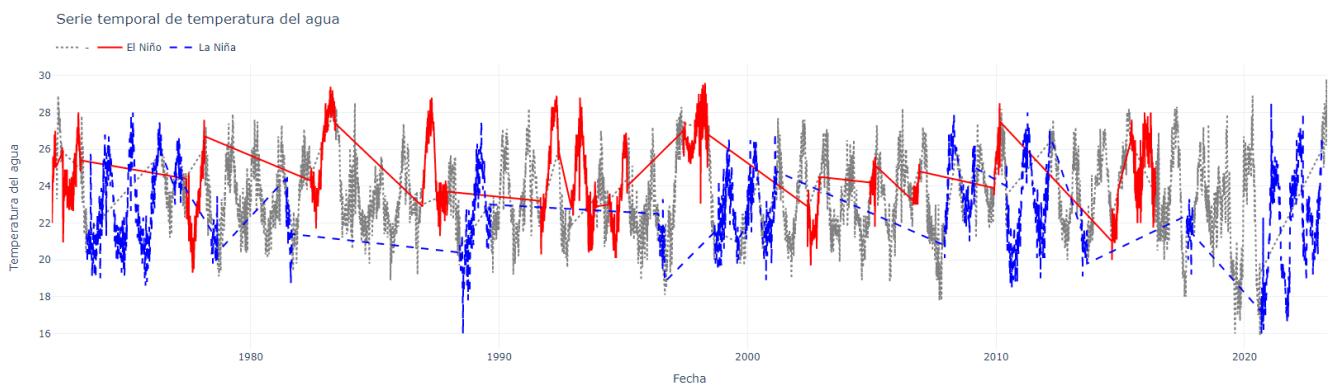


Figura 4.3. Serie temporal de la temperatura del agua del mar.

La **Figura 4.3** representa la temperatura del agua del mar y, como se puede observar, durante los días en los que se produce “El Niño” la temperatura del agua se eleva, al contrario de lo que sucede durante “La Niña”. Como se ha explicado anteriormente, no siempre se producen los fenómenos durante las mismas épocas del año, aunque sea más común que durante diciembre y febrero (época cálida) se produzca el fenómeno de “El Niño” y entre junio y agosto (época más fría) se produzca “La Niña”, por lo tanto, como existe de la posibilidad de que se produzca “El Niño” en la época más fría o “La Niña” en la época más cálida podemos ver temperaturas lejos de la media en cada uno de los fenómenos.

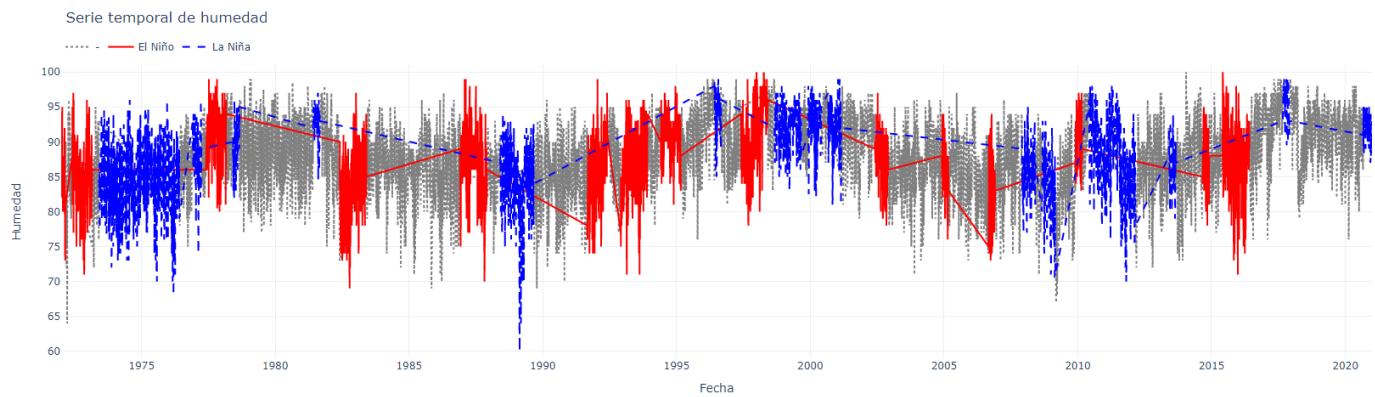


Figura 4.4. Serie temporal de la humedad relativa del aire durante los fenómenos.

En la **Figura 4.4** podemos observar cómo aumenta la humedad relativa del aire durante el fenómeno de El Niño, se mantiene en los períodos de transición y disminuye durante La Niña.

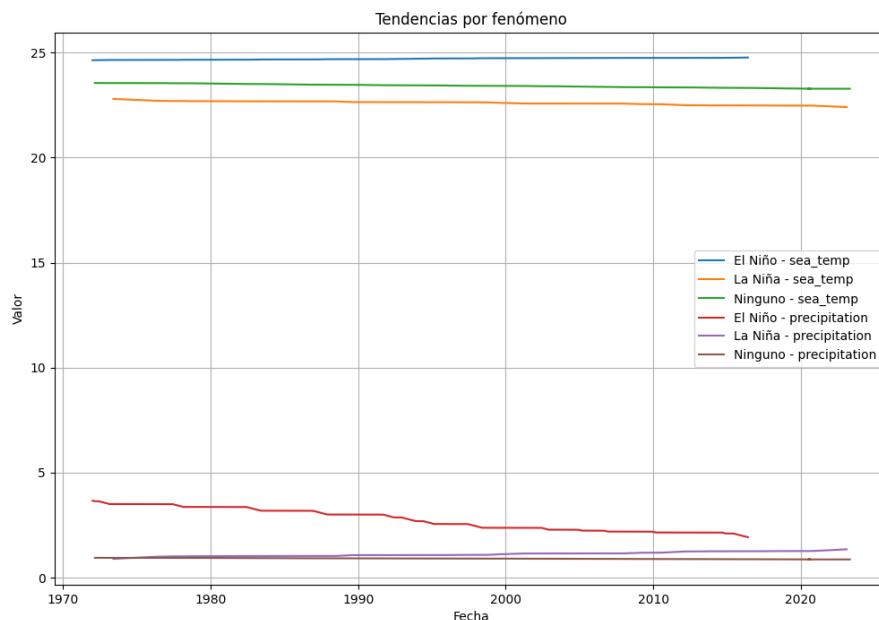


Figura 4.5. Tendencias de precipitación y temperaturas del mar dura

En la **Figura 4.5** podemos observar que, al igual que sucedía con la temperatura del ambiente, la mayor temperatura en el agua se observa durante “El Niño” y la menor durante “La Niña”. A su vez, se aprecia una clara diferencia en las líneas de tendencia de las precipitaciones, siendo superiores cuando sucede “El Niño”, y en este caso, se observa mayor cantidad de precipitación durante “La Niña” que en los períodos de transición.

Por último, podemos ver en la **Figura 4.6**, la serie temporal de las horas de sol para cada uno de los tres fenómenos.

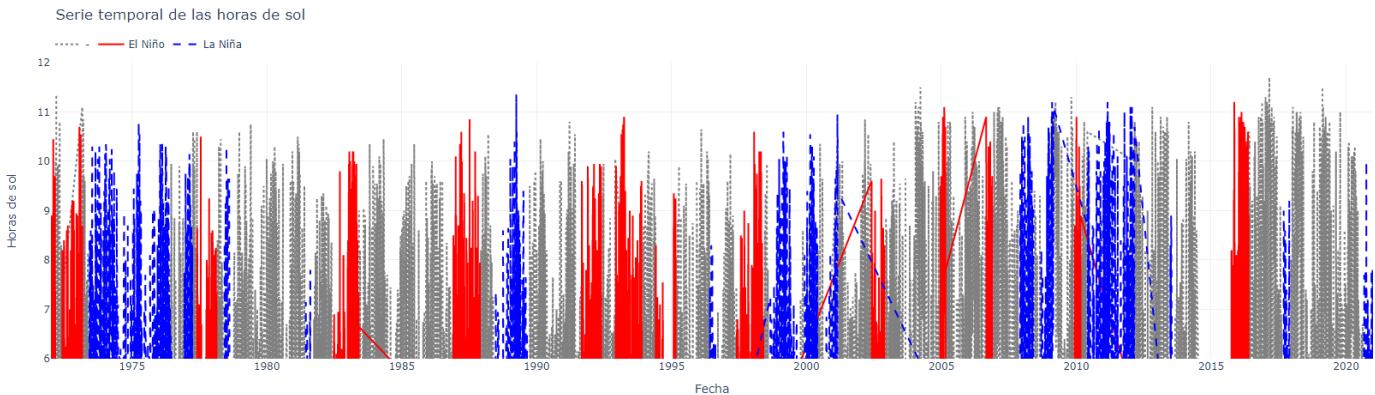


Figura 4.6. Serie temporal de las horas de sol durante los fenómenos.

Como conclusión de esta gráfica, no podemos concluir patrones claros, ya que, normalmente las horas de sol son bastante constantes y, además, nos falta información sobre los datos ya que no sabemos cómo se captan estas horas de sol, es decir, si son tomadas desde que sale el sol hasta que se va o si la aparición de intervalos nubosos hace que el número de horas de sol cambie.

Por otro lado, se ha realizado un análisis complementario de varianza (ANOVA). Este análisis permite determinar si hay diferencias significativas en las variables entre los diferentes fenómenos, lo cual nos ayudará a evaluar si hay efectos estadísticamente significativos asociados con cada fenómeno. Las métricas estadísticas que obtenemos como medidas de resultados son las siguientes:

- **F-statistic** es una medida de la variabilidad entre los grupos en comparación con la variabilidad dentro de los grupos. Un valor *F-statistic* grande indica una mayor diferencia entre los grupos.
- **p-value** es la probabilidad de obtener un *F-statistic* igual o más extremo que el observado si la hipótesis nula es verdadera. Un *p-value* pequeño ( $< 0.05$ ), indica que las diferencias observadas son estadísticamente significativas.

Los valores de las métricas estadísticas obtenidos son los siguientes:

Variable: min_air_temp	Variable: humidity
F-statistic: 307.0190017727969	F-statistic: 14.565450213659323
P-value: 6.350336673324992e-132	P-value: 4.777717845910451e-07
Variable: max_air_temp	Variable: precipitation
F-statistic: 167.47134489171393	F-statistic: 93.64606825619589
P-value: 8.145633978118553e-73	P-value: 3.404615065691186e-41
Variable: mean_air_temp	Variable: sunshine_hours
F-statistic: 460.46932321634085	F-statistic: 30.421651505635438
P-value: 6.062730621131601e-196	P-value: 6.448548269889805e-14
Variable: sea_temp	Variable: clouds
F-statistic: 990.0318807080221	F-statistic: 44.85729723391167
P-value: 0.0	P-value: 3.67486466939192e-20

Figura 4.7. Análisis de la varianza ANOVA de las variables numéricas.

En los resultados de la **Figura 4.7**, podemos observar que los *p-value* son extremadamente pequeños (cercaos a cero, incluso cero en el caso de la temperatura del mar), lo que indica que hay diferencias significativas entre los fenómenos para estas variables. En cuanto a los valores de *F-statistic*, vemos que los valores más grandes se encuentran en la temperatura del mar y en la temperatura ambiente, lo que indica que es donde existen las mayores diferencias.

## 5. Clasificación

En este apartado se van a utilizar distintos modelos de clasificación para predecir la ocurrencia del fenómeno del Niño y de la Niña, con el fin de comparar dichos modelos y determinar cuál de ellos presenta un desempeño superior. Para ello, realizaremos entrenamiento de los modelos con un porcentaje alto de los datos y realizaremos una *Cross-Validation* con 5 *splits* para comprobar la fiabilidad de cada modelo en la clasificación de datos. Finalmente, realizaremos una validación de los modelos utilizando como conjunto de *testing* los datos de 2023. Los algoritmos de clasificación que han sido utilizados son los siguientes:

- **SVC** (*Support Vector Machine*), utilizando tres tipos de kernel (*rbf*, *poly* y *sigmoid*) y ajustando los hiper-parámetros *C*, *degree* y *coef0* con dependencia del *kernel* utilizado.
- **Clasificación con árboles de decisión**, donde probaremos diferentes ajustes para los hiper-parámetros *max\_depth* y *max\_leaf\_nodes*, que indican el número máximo de profundidad del árbol y el número máximo de hojas por nodo respectivamente.
- **Regresión logística**, ajustando los hiper-parámetros *C*, *penalty*, *solver* y *class\_weight*.

### 5.1. Pasos previos a la clasificación

Para obtener unos resultados óptimos se han realizado varios de ajustes en los datos para su correcta aplicación a los distintos algoritmos de clasificación. Además de probar estos algoritmos, nos ha interesado hacer pruebas con distintos parámetros de entrada. Un primer paso ha sido convertir las fechas a un formato con el que podamos trabajar sin ningún problema. Como la mayoría de los algoritmos a probar trabajan con distancias, el formato *Posix* es una buena apuesta ya que mantiene la secuencialidad del tiempo sin dejar de ser numérico.

En primer lugar, para utilizar los diferentes algoritmos, sepáramos los datos en la columna de salida y las de entrada. La variable de salida, a la que llamaremos *y*, representa la columna **fenomeno** convirtiendo los valores “-”, “El Niño” y “La Niña” a los valores numéricos 0, 1 y 2 respectivamente. El segundo paso es definir los diferentes *dataset* con los que entrenaremos los algoritmos para poder realizar comparaciones en los resultados. Los conjuntos de datos sobre los que trabajaremos son los siguientes:

- **data\_x\_no\_sun**: conjunto de datos ignorando la columna de horas de sol.
- **data\_x\_sun\_2004**: *dataset* ignorando las filas previas a 2004, ya que son las que tienen horas de sol imputadas artificialmente.
- **data\_x\_no\_clouds**: *dataset* ignorando la columna de cantidad de nubes.
- **data\_x\_balanced**: *dataset* equilibrando datos quitando gran parte de los “-” (clase mayoritaria).

El tercer paso será construir varias funciones auxiliares para poder aplicar los algoritmos de clasificación:

- Función auxiliar que realiza el entrenamiento con *Cross-Validation* para un modelo con unos parámetros concretos, un *dataset* concreto y un escalador concreto.
- Función auxiliar que utiliza la anterior función con todos los escaladores definidos para un *dataset* y algoritmo concreto recibidos por parámetro.
- Función auxiliar que utiliza una función para probar un algoritmo concreto por parámetro y la utiliza con todos los *dataset*. La función parámetro de cada algoritmo se definirá en su respectiva sección y probará dicho algoritmo con un *dataset* concreto y con cada uno de los escaladores definidos, que son el *StandardScaler* y el *MinMaxScaler*, además de la opción sin escalador.

## 5.2. Clasificación con el algoritmo SVC (*Support Vector Classifier*)

El **SVC** es un algoritmo que se basa en construir un hiperplano óptimo que pueda separar de manera efectiva las muestras de diferentes clases en un espacio de características de alta dimensión. Una de las características destacadas de SVC es su capacidad para manejar tanto datos linealmente separables como no separables, gracias al uso de funciones de *kernel* que permiten mapear los datos y obtener espacios de mayor dimensionalidad. Es muy utilizado por las distintas ventajas que aporta, como su eficacia en conjuntos de datos pequeños y medianos, y su capacidad para lidiar con problemas de alta dimensionalidad. Además, suele aprender mejores generalizaciones. Una desventaja que tiene es que puede ser computacionalmente costoso en conjuntos de datos muy grandes, ya que el tiempo de entrenamiento y la complejidad computacional aumenta enormemente con el número de muestras, como es nuestro caso. A pesar de esta limitación, como es uno de los algoritmos más utilizados hemos creído que es conveniente probar el algoritmo con nuestros datos.

En este algoritmo hemos tenido que elegir automáticamente los mejores valores de hiper-parámetros utilizando los *kernels* ‘rbf’, ‘poly’ y ‘sigmoid’ por separado. Para ello, vamos a analizar, mediante la librería de *Python* *GridSearchCV*, los siguientes hiper-parámetros:

- **C**: penalización por los errores de clasificación en el modelo. A mayor valor de *C* más se penalizará por los errores de clasificación.
- **degree**: controla el grado del polinomio utilizado para calcular la función de decisión. Por ejemplo, un valor de 3 significa que se va a utilizar un polinomio de tercer grado. Este hiper-parámetro solo es utilizado para el *kernel* de tipo polinomial.
- **coef0**: término independiente en la función de decisión polinómica. Por ejemplo, un valor de 1 significa que se va a utilizar un término independiente igual a 1 en la función de decisión, lo que afectaría al sesgo del modelo y podría ayudar a ajustar la función de decisión para que se ajuste mejor a los datos. Este hiper-parámetro solo aplica para los *kernel* de tipo polinomial y sigmoidal.

En la **Figura 5.1**, podemos ver el mejor valor calculado para cada hiper-parámetro y la *Mean Accuracy* obtenida tras su ajuste.

Kernel	Hiper-parámetro	Mejor valor	Mean Accuracy
RBF	$C$	0.1	0.5
Poly (Polinomial)	$C$	0.2	0.56
	degree	2	0.56
	coef0	0.0	0.56
Sigmoid (Sigmoidal)	$C$	0.1	0.34
	coef0	10	0.56

Figura 5.1. Resultados de la obtención de los hiper-parámetros para la clasificación con SVM.

Destacamos el aumento de la *Mean Accuracy* obtenida para el *kernel* sigmoidal tras ajustar el parámetro *coef0*. Tras el ajuste de hiper-parámetros, hemos puesto a prueba el modelo SVC con todos los conjuntos de datos, con los diferentes *kernel* y los escaladores definidos.

```

Mean Accuracy: 0.58 for dataset 'Full dataset', Model: SVC(kernel=rbf), Scaler: No scaler
Mean Accuracy: 0.62 for dataset 'Full dataset', Model: SVC(kernel=rbf), Scaler: StandardScaler
Mean Accuracy: 0.6 for dataset 'Full dataset', Model: SVC(kernel=rbf), Scaler: MinMaxScaler
Mean Accuracy: 0.56 for dataset 'Full dataset', Model: SVC(kernel=poly), Scaler: No scaler
Mean Accuracy: 0.58 for dataset 'Full dataset', Model: SVC(kernel=poly), Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'Full dataset', Model: SVC(kernel=poly), Scaler: MinMaxScaler
Mean Accuracy: 0.56 for dataset 'Full dataset', Model: SVC(kernel=sigmoid), Scaler: No scaler
Mean Accuracy: 0.56 for dataset 'Full dataset', Model: SVC(kernel=sigmoid), Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'Full dataset', Model: SVC(kernel=sigmoid), Scaler: MinMaxScaler
Mean Accuracy: 0.58 for dataset 'No sunshine hours', Model: SVC(kernel=rbf), Scaler: No scaler
Mean Accuracy: 0.62 for dataset 'No sunshine hours', Model: SVC(kernel=rbf), Scaler: StandardScaler
Mean Accuracy: 0.61 for dataset 'No sunshine hours', Model: SVC(kernel=rbf), Scaler: MinMaxScaler
Mean Accuracy: 0.56 for dataset 'No sunshine hours', Model: SVC(kernel=poly), Scaler: No scaler
Mean Accuracy: 0.58 for dataset 'No sunshine hours', Model: SVC(kernel=poly), Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'No sunshine hours', Model: SVC(kernel=poly), Scaler: MinMaxScaler
Mean Accuracy: 0.56 for dataset 'No sunshine hours', Model: SVC(kernel=sigmoid), Scaler: No scaler
Mean Accuracy: 0.56 for dataset 'No sunshine hours', Model: SVC(kernel=sigmoid), Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'No sunshine hours', Model: SVC(kernel=sigmoid), Scaler: MinMaxScaler
Mean Accuracy: 0.6 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=rbf), Scaler: No scaler
Mean Accuracy: 0.67 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=rbf), Scaler: StandardScaler
Mean Accuracy: 0.66 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=rbf), Scaler: MinMaxScaler
Mean Accuracy: 0.6 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=poly), Scaler: No scaler
Mean Accuracy: 0.62 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=poly), Scaler: StandardScaler
Mean Accuracy: 0.66 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=poly), Scaler: MinMaxScaler
Mean Accuracy: 0.6 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=sigmoid), Scaler: No scaler
Mean Accuracy: 0.6 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=sigmoid), Scaler: StandardScaler
Mean Accuracy: 0.6 for dataset 'Sunshine hours >= 2004', Model: SVC(kernel=sigmoid), Scaler: MinMaxScaler
Mean Accuracy: 0.58 for dataset 'No cloud count', Model: SVC(kernel=rbf), Scaler: No scaler
Mean Accuracy: 0.62 for dataset 'No cloud count', Model: SVC(kernel=rbf), Scaler: StandardScaler
Mean Accuracy: 0.6 for dataset 'No cloud count', Model: SVC(kernel=rbf), Scaler: MinMaxScaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: SVC(kernel=poly), Scaler: No scaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: SVC(kernel=poly), Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: SVC(kernel=poly), Scaler: MinMaxScaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: SVC(kernel=sigmoid), Scaler: No scaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: SVC(kernel=sigmoid), Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: SVC(kernel=sigmoid), Scaler: MinMaxScaler
Mean Accuracy: 0.47 for dataset 'Balanced', Model: SVC(kernel=rbf), Scaler: No scaler
Mean Accuracy: 0.57 for dataset 'Balanced', Model: SVC(kernel=rbf), Scaler: StandardScaler
Mean Accuracy: 0.55 for dataset 'Balanced', Model: SVC(kernel=rbf), Scaler: MinMaxScaler
Mean Accuracy: 0.43 for dataset 'Balanced', Model: SVC(kernel=poly), Scaler: No scaler
Mean Accuracy: 0.51 for dataset 'Balanced', Model: SVC(kernel=poly), Scaler: StandardScaler
Mean Accuracy: 0.55 for dataset 'Balanced', Model: SVC(kernel=poly), Scaler: MinMaxScaler
Mean Accuracy: 0.39 for dataset 'Balanced', Model: SVC(kernel=sigmoid), Scaler: No scaler
Mean Accuracy: 0.39 for dataset 'Balanced', Model: SVC(kernel=sigmoid), Scaler: StandardScaler
Mean Accuracy: 0.39 for dataset 'Balanced', Model: SVC(kernel=sigmoid), Scaler: MinMaxScaler

```

Figura 5.2. Resultados de la ejecución de la validación cruzada con el modelo SVC.

En este caso, hemos de destacar que tal y como se esperaba, este es el algoritmo que más ha tardado en procesarse ya que, como se ha comentado anteriormente, es un algoritmo pesado y lento en comparación con otros enfoques de clasificación. Esto se debe a su naturaleza de maximizar el margen entre clases, lo que implica un cálculo intensivo y una complejidad computacional más alta. Aunque su desempeño puede variar según el conjunto de datos, se ha observado que los mejores resultados se obtienen utilizando el *dataset* de filas posteriores a 2004, combinado con el *Kernel RBF* y el escalado de características mediante *StandardScaler*.

Para estudiar más a fondo los resultados del clasificador SVM, vemos en la **Figura 5.3** y la **Figura 5.4** la matriz de confusión y las curvas ROC.

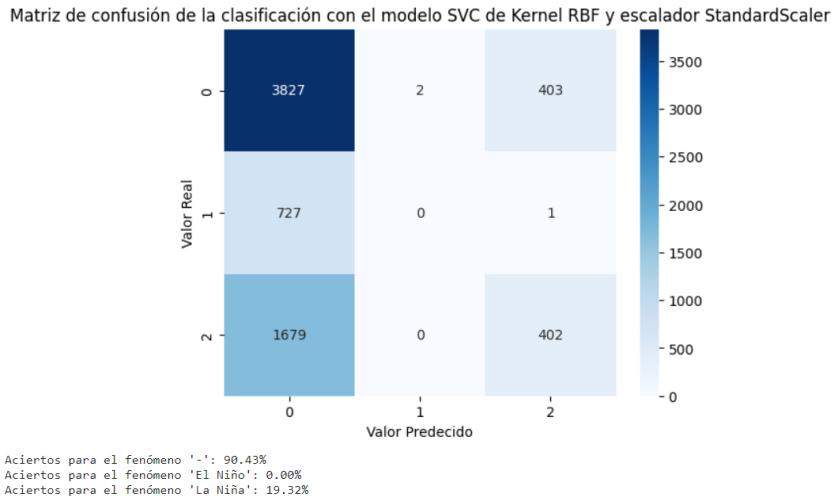


Figura 5.3. Matriz de confusión de la clasificación con el modelo SVC, *Kernel RBF* y escalador *StandartScaler*.

En la **Figura 5.3**, matriz de confusión se están anotando los valores reales y predecidos 0 (fenómeno '-'), 1 (fenómeno 'El Niño') y 2 (fenómeno 'La Niña'). Como primera observación, vemos que la mayoría de las muestras se clasifican como días de transición.

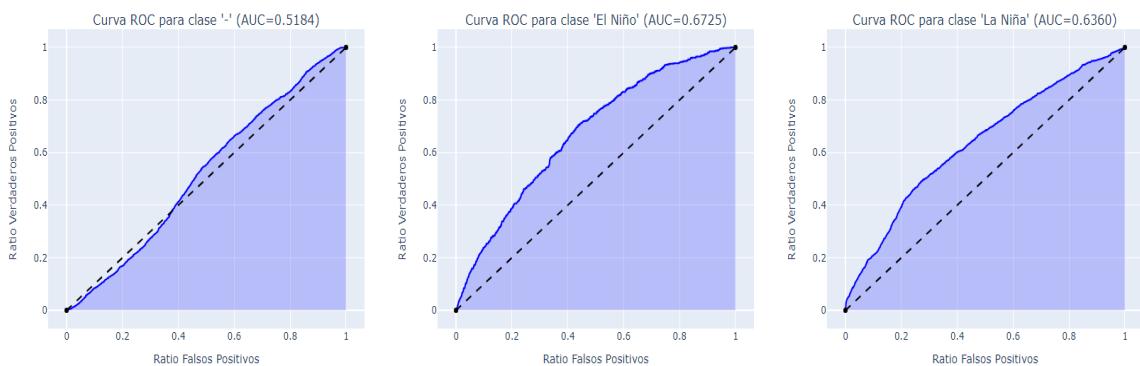


Figura 5.4. Curvas ROC para cada clase de clasificación con el modelo SVC, *Kernel RBF* y escalador *StandartScaler*.

En vista de la **Figura 5.3** y de la **Figura 5.4** (matriz de confusión y curvas ROC) la evaluación cambia drásticamente. El modelo ha aprendido a decir casi siempre que no hay fenómeno ni del Niño ni de la Niña, y al existir desbalance en los datos y encontrarse más ocurrencias de '-', el resultado de *Mean Accuracy* es engañoso y hace que parezca un mejor resultado de lo que debería. Tras ajustar varios hiper-parámetros y probar distintos

preprocesados, el algoritmo SVC no ha superado esta situación. En un principio, y llegando a una primera conclusión, SVC ha demostrado no ser el algoritmo más adecuado para esta tarea.

### 5.3. Clasificación con Árboles de Decisión

Como segundo algoritmo hemos trabajado los árboles de decisión, los cuales están basados en estructuras de árbol donde cada nodo interno representa una característica o atributo, cada rama representa una decisión basada en ese atributo y cada hoja representa una clase o resultado final. Estos árboles se construyen dividiendo repetidamente los datos en función de las características más relevantes hasta que se alcanza una pureza suficiente en las hojas. La clasificación se realiza siguiendo el camino desde la raíz hasta una hoja, tomando las decisiones basadas en los atributos en cada nodo. Sus ventajas incluyen la capacidad de manejar diferentes tipos de datos, su facilidad de interpretación y su eficiencia con grandes conjuntos de información. Sin embargo, pueden resultar que se produzca *overfitting* fácilmente, lo cual requiere ajustes manuales como la poda (limitando la profundidad máxima del árbol). A pesar de esta desventaja, los Árboles de Decisión son ampliamente utilizados debido a su simplicidad y capacidad para capturar relaciones complejas en los datos, siendo una opción atractiva en muchos casos. En nuestro caso al no estar basado en distancia no es necesario escalar los datos, pues no afectará al resultado. Por ello, solo se prueban con escalador *None*, es decir, sin escalador.

```

Mean Accuracy: 0.6 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=3)
Mean Accuracy: 0.7 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=3)
Mean Accuracy: 0.71 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=3)
Mean Accuracy: 0.73 for dataset 'Full dataset', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=3)
Mean Accuracy: 0.7 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=3)
Mean Accuracy: 0.71 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=3)
Mean Accuracy: 0.7 for dataset 'No sunshine hours', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=2)
Mean Accuracy: 0.72 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=3)
Mean Accuracy: 0.78 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=10)
Mean Accuracy: 0.72 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=2)
Mean Accuracy: 0.72 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=3)
Mean Accuracy: 0.86 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=10)
Mean Accuracy: 0.72 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=2)
Mean Accuracy: 0.72 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=3)
Mean Accuracy: 0.9 for dataset 'No sunshine hours >= 2004', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=3)
Mean Accuracy: 0.7 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=3)
Mean Accuracy: 0.71 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=10)
Mean Accuracy: 0.6 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=2)
Mean Accuracy: 0.65 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=3)
Mean Accuracy: 0.73 for dataset 'No cloud count', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=10)
Mean Accuracy: 0.45 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=2)
Mean Accuracy: 0.53 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=3)
Mean Accuracy: 0.63 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=4, max_leaf_nodes=10)
Mean Accuracy: 0.45 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=2)
Mean Accuracy: 0.53 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=3)
Mean Accuracy: 0.64 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=5, max_leaf_nodes=10)
Mean Accuracy: 0.45 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=2)
Mean Accuracy: 0.53 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=3)
Mean Accuracy: 0.7 for dataset 'Balanced', Model: DecisionTreeClassifier(max_depth=10, max_leaf_nodes=10)

```

Figura 5.5. Resultados de la ejecución de la validación cruzada con el modelo de árbol de decisión..

Este algoritmo ha sido el más rápido entrenando con una abismal diferencia, convirtiéndolo en el más eficiente que hemos probado. Esto está alineado con lo explicado anteriormente. En los resultados se pueden apreciar los valores más altos, llegando incluso a 0.9 en *Accuracy*. La poda, limitando la profundidad máxima (*max\_depth*) y el número de nodos hoja (*max\_leaf\_nodes*), ha ayudado enormemente a evitar casos de sobreajuste, ya que han permitido al árbol generalizar mejor y no abrir hojas para cada caso específico.

En la **Figura 5.6** podemos ver el árbol de decisión del modelo con el que llegamos a una *Accuracy* de 0.9. En la validación cruzada de este modelo se obtiene este resultado utilizando el dataset que contiene únicamente las muestras a partir de 2004 con un *max\_depth* de 10 y un *max\_leaf\_nodes* de 10.

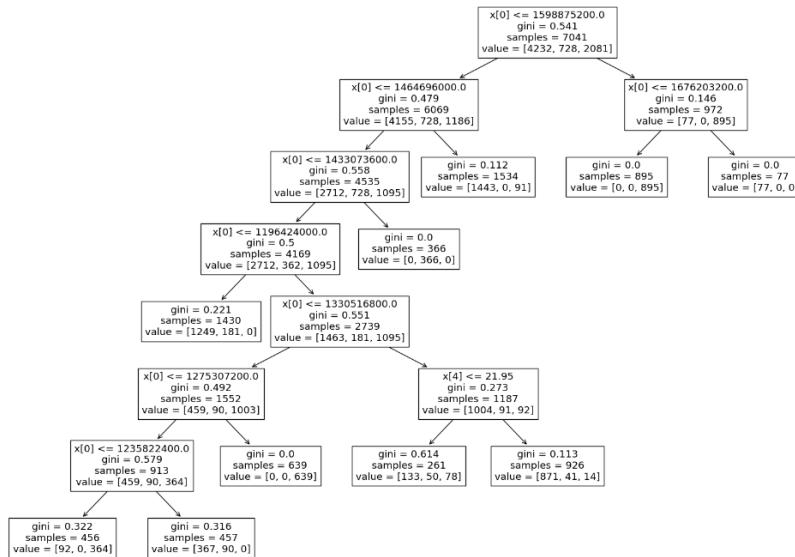


Figura 5.6. Árbol de decisión con *max\_depth* de 10 y un *max\_leaf\_nodes* de 10.

En el árbol de decisión vemos que las hojas se distribuyen a partir de los valores de *x[0]* y *x[4]* que corresponden a *min\_air\_temp* y a *humidity* respectivamente. Por lo tanto, a partir de estos dos valores el modelo decide qué fenómeno se está produciendo.

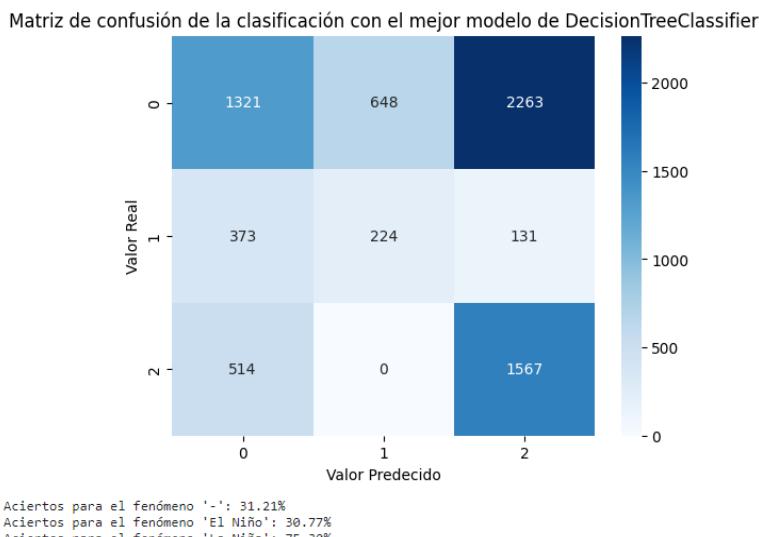


Figura 5.7. Matriz de confusión con el mejor modelo de *DecisionTreeClassifier*.

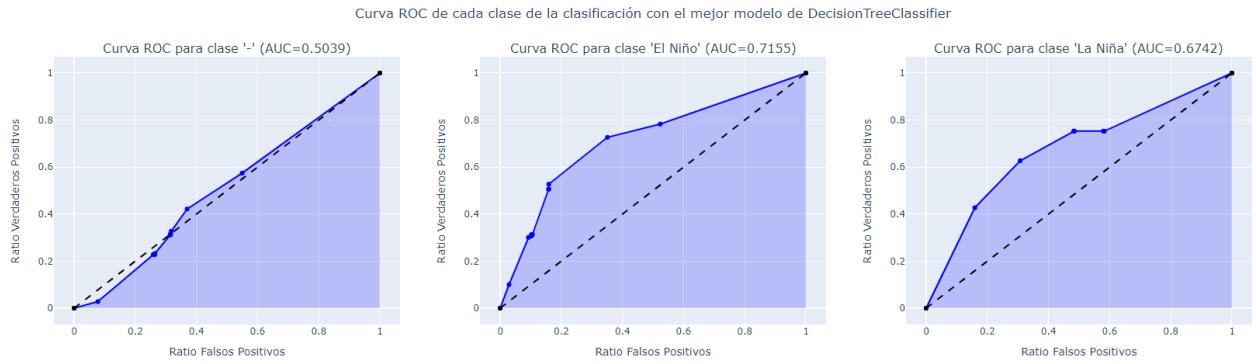


Figura 5.8. Curva ROC de cada clase de la clasificación con el mejor modelo de *DecisionTreeClassifier*.

Como podemos observar en las **Figuras 5.7 y 5.8**, los resultados de este algoritmo son mucho más prometedores. Con respuestas mucho más variadas, se ha conseguido un desempeño mucho más satisfactorio. Aun así, este modelo también tiene problemas prediciendo el fenómeno del Niño. Debido a que el conjunto de datos no incluye demasiadas muestras de esta clase, esta situación es normal y la excesiva pequeña cantidad de muestras de este tipo hace que equilibrar los datos eliminando muestras de las clases mayoritarias tampoco sea lo más efectivo. Como conclusión, el árbol de decisión se presenta como una opción viable, que se podrá validar más adelante al compararla con el resto usando los datos más recientes del conjunto.

#### 5.4. Clasificación con el algoritmo *LogisticRegression*

El último algoritmo que se ha utilizado para la clasificación es el de Regresión Logística. Aunque su nombre incluye la palabra "regresión", en realidad se utiliza para la clasificación, ya que realmente utiliza una función logística para modelar la probabilidad de que una instancia pertenezca a una clase determinada. La función logística asigna los valores de entrada a un rango entre 0 y 1, lo que permite interpretar la salida como la probabilidad de pertenencia a una clase. Este algoritmo destaca por su simplicidad, eficiencia y capacidad de interpretación. Es especialmente útil en problemas de clasificación lineal, ya que estima la probabilidad de pertenencia a una clase. Aunque puede no funcionar bien en situaciones con relaciones no lineales complejas, sigue siendo una opción sólida en muchos escenarios, especialmente cuando se busca un modelo interpretable y sencillo. Una de las ventajas de este algoritmo de clasificación es su capacidad de interpretación. Este algoritmo da mucha explicabilidad a los modelos entrenados, permitiendo entender la influencia de cada característica en los resultados. La explicabilidad es una propiedad muy buscada en sistemas de IA modernos ya que permite a los seres humanos y otros sistemas informáticos distintos aprender de las relaciones que a su vez ha aprendido la máquina, habilitando así que ese conocimiento se pueda transferir a otros sistemas.

Para una óptima clasificación se procede a la elección de los mejores hiper-parámetros utilizando la librería **GridSearchCV**, obteniendo los siguientes resultados:

- $C = 0.001$ : parámetro de regularización inverso, es decir, cuanto menor sea el valor de C, mayor será la regularización.
- **penalty = l2**: norma utilizada en la regularización del modelo. Se está utilizando la norma 'l2', que penaliza la magnitud de los coeficientes al cuadrado.

También, se han definido los híper-parámetros **solver** y **class\_weight** manualmente para todos los modelos de *LogisticRegression*. Por su parte, solver especifica el algoritmo usado para la optimización, y al que se le ha dado el valor de '*liblinear*', ya que es el más eficiente en tiempo y memoria. Aunque el valor por defecto ahora es '*lbfgs*', *liblinear* ha demostrado tener una precisión similar o mejor y es mucho más rápido. Por otro lado, **class\_weight** define cómo se asignan pesos a las distintas clases. El valor que hemos usado, '*balanced*', asigna pesos inversamente proporcionales a la cantidad de muestras por clase, lo cual es de gran ayuda en conjuntos desequilibrados como este, donde hay clases muy mayoritarias. Al usarlo sufrimos sus dos desventajas: una que ralentiza el entrenamiento por los cálculos adicionales, pero no ha sido un aumento notable, y otra que, aunque mejora las predicciones de las clases minoritarias, hace que la clase mayoritaria sufra por ello. En SVC hemos visto cómo el modelo aprendía, obteniendo mayoritariamente el valor ' $-$ ', y aunque en ese caso equilibrar los pesos no fue de ayuda, aquí en *LogisticRegression* hemos visto que mejora algo en conjunto con otros híper-parámetros como **solver**.

```
Mean Accuracy: 0.56 for dataset 'Full dataset', Model: LogisticRegression, Scaler: No scaler
Mean Accuracy: 0.56 for dataset 'Full dataset', Model: LogisticRegression, Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'Full dataset', Model: LogisticRegression, Scaler: MinMaxScaler
Mean Accuracy: 0.56 for dataset 'No sunshine hours', Model: LogisticRegression, Scaler: No scaler
Mean Accuracy: 0.57 for dataset 'No sunshine hours', Model: LogisticRegression, Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'No sunshine hours', Model: LogisticRegression, Scaler: MinMaxScaler
Mean Accuracy: 0.6 for dataset 'Sunshine hours >= 2004', Model: LogisticRegression, Scaler: No scaler
Mean Accuracy: 0.61 for dataset 'Sunshine hours >= 2004', Model: LogisticRegression, Scaler: StandardScaler
Mean Accuracy: 0.6 for dataset 'Sunshine hours >= 2004', Model: LogisticRegression, Scaler: MinMaxScaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: LogisticRegression, Scaler: No scaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: LogisticRegression, Scaler: StandardScaler
Mean Accuracy: 0.56 for dataset 'No cloud count', Model: LogisticRegression, Scaler: MinMaxScaler
Mean Accuracy: 0.39 for dataset 'Balanced', Model: LogisticRegression, Scaler: No scaler
Mean Accuracy: 0.47 for dataset 'Balanced', Model: LogisticRegression, Scaler: StandardScaler
Mean Accuracy: 0.46 for dataset 'Balanced', Model: LogisticRegression, Scaler: MinMaxScaler
```

Figura 5.9. Resultados de la ejecución de la validación cruzada con el modelo de regresión logística.

La Regresión Logística ha mostrado resultados mejorables con un valor de alrededor de 0.6 en términos de precisión. Destaca por ser un punto intermedio entre diferentes aspectos de los algoritmos de clasificación, no solo en precisión, sino también en velocidad. No es tan rápido como los árboles de decisión, pero SVC ha sido mucho más lento. A pesar de sus limitaciones en casos de relaciones no lineales complejas, la Regresión Logística tiene la ventaja de la explicabilidad, gracias a la cual podemos ver qué características son las más influyentes en la decisión. Esta información puede ser muy beneficiosa para desarrollar nuevos y mejores modelos usando algoritmos que ataquen a los puntos de interés o preprocesando el conjunto de datos de otra manera. El mejor modelo en este caso ha sido un 0.61 de precisión con el *dataset* de filas posteriores a 2004 y escalador *StandardScaler*. A continuación, podemos ver algunos de sus coeficientes:

```
[-1.09968773, 0.61597127, 3.51679435, -1.45600507, -1.47513339, 2.65208339, -1.668446, 0.0551392, 0.02974742]
[-0.54375249, 0.07853107, -2.82328565, 3.56131909, 1.67824051, -1.91784259, -0.8379233, -0.88300649, 0.19700466]
[1.64344022, -0.69450234, -0.69350871, -2.10531403, -0.20310712, -0.7342408, 2.50636931, 0.82786729, -0.22675209]
```

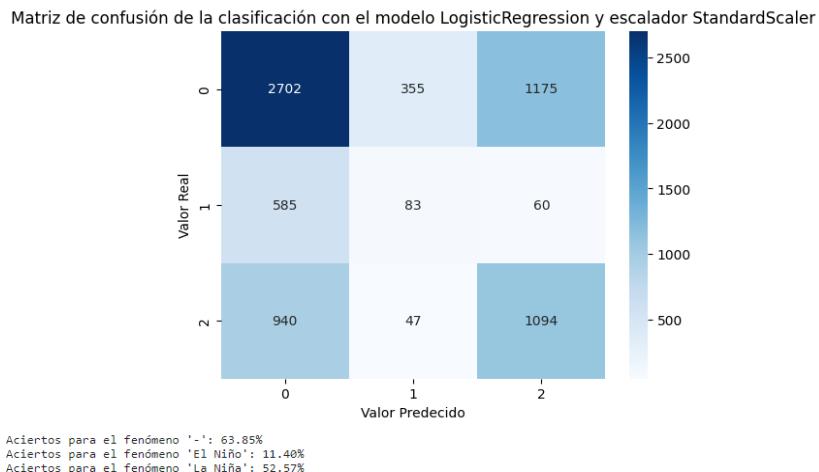
En estos coeficientes observamos que valores como el tercero de cada vector o el cuarto en el segundo vector son mayores que los demás.

#	Column	Non-Null Count	Dtype
0	observation_date	7041	non-null
1	min_air_temp	7041	non-null
2	max_air_temp	7041	non-null
3	mean_air_temp	7041	non-null
4	sea_temp	7041	non-null
5	humidity	7041	non-null
6	precipitation	7041	non-null
7	sunshine_hours	7041	non-null
8	clouds	7041	non-null

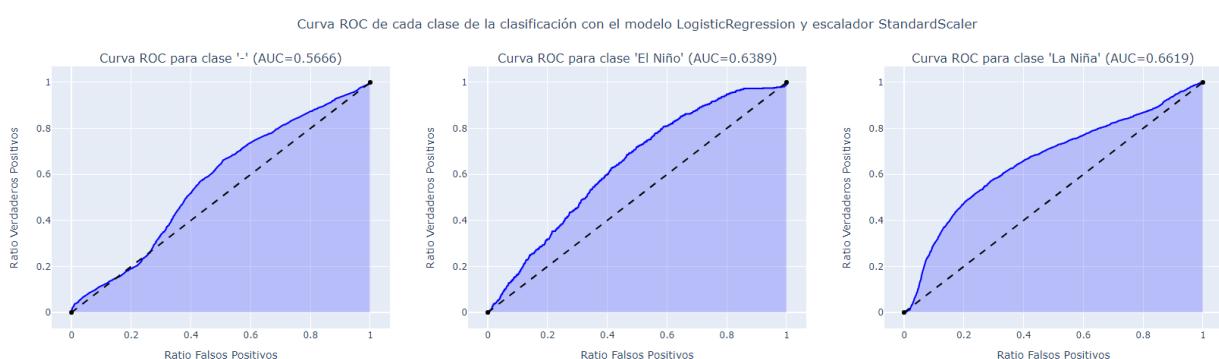
dtypes: float64(8), int64(1)  
memory usage: 550.1 KB

### 5.10 Estructura y tipos de los datos

Tal como vemos la estructura de nuestros datos en la **Figura 5.10**, y relacionándolos con los valores de los tres vectores anteriores podemos observar que los valores altos pertenecen sobre todo a *max\_air\_temp* y *mean\_air\_temp*, los valores tercero y cuarto. Dado que están relacionados (vimos una pequeña correlación en apartados anteriores), es normal que ambos tengan puntuaciones relacionadas. Este resultado indica que esos parámetros son los que más explican la ocurrencia del fenómeno del Niño o de la Niña, según el modelo aprendido con *LogisticRegression*. Los otros modelos, cuyos datos han resultado peores, con precisión del 50%, o sea fiabilidad nula, tienen valores cercanos a cero en todos los coeficientes. Esto indica que esos modelos no han encontrado ningún patrón claro en ninguna de las características, de ahí su falta de precisión.



5.11. Matriz de confusión de la clasificación con el modelo *LogisticRegression* y escalador *StandartScaler*.



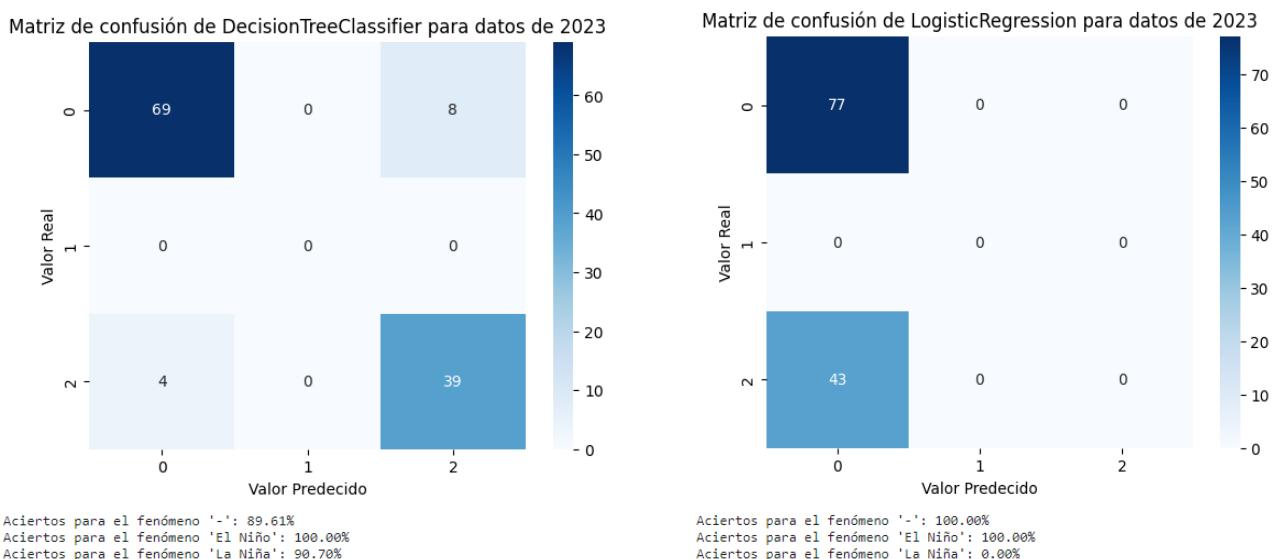
5.12. Curvas ROC de clasificación con el modelo *LogisticRegression* y escalador *StandartScaler*.

Podemos observar en la **Figura 5.11** y **Figura 5.12** que este modelo ha obtenido peores puntuaciones de Accuracy que el árbol de decisión, pero sus respuestas siguen dejando entrever una opción viable que merece la pena explorar. De forma similar al modelo anterior, el fenómeno del Niño le es difícil de predecir en comparación con la Niña o días de transición, por la misma razón del desbalance de los datos.

## 5.5. Predicción de un rango temporal del fenómeno en el año 2023

En general, cada algoritmo tiene sus fortalezas y debilidades en la clasificación del fenómeno del Niño y la Niña. Si se busca una mayor precisión y se dispone de recursos computacionales suficientes, SVC puede ser una opción viable. Si se valora la velocidad de entrenamiento y se toman precauciones para evitar el sobreajuste, los árboles de decisión pueden ser una buena elección. Por otro lado, si se busca una interpretación clara de los resultados y se necesita un modelo sencillo, o se busca indagar en las relaciones entre las características estudiadas, la Regresión Logística puede ser la opción elegida. En este caso el modelo que mejor se ha comportado es el *DecisionTreeClassifier*, ya que es el único que no ha decidido marcar casi todas las entradas con "-".

Como prueba final, nos interesa utilizar los mejores modelos de cada algoritmo para tratar de predecir los datos más recientes que podemos comprobar, los de los meses de enero a abril de 2023. Es importante remarcar que en este periodo no ocurrió el fenómeno del Niño por lo que cualquier muestra clasificada como este fenómeno será un error de clasificación. Como SVC ha mostrado un desempeño mediocre en las pruebas iniciales, no se va a añadir en esta segunda fase de pruebas. Por lo tanto, el mejor modelo de **Árbol de decisión** y el mejor modelo de **Regresión logística** han sido los elegidos para demostrar cuál es el más preciso a la hora de realizar esta tarea.



5.13. Matrices de confusión de la clasificación con árboles de decisión y regresión logística para datos del 2023.

Podemos ver en la **Figura 5.13**, al utilizar los dos modelos con los datos más recientes, observamos como el árbol de decisión consigue clasificar satisfactoriamente una alta cantidad de casos, mientras que el de Regresión Logística cae en la trampa de predecir todo como días de transición, siendo este un resultado nada correcto. Por ello, el modelo entrenado con Árboles de Decisión, usando sólo datos desde 2004 hasta la actualidad, y podando el árbol a 10 nodos hoja y 10 de profundidad máxima, ha demostrado ser un buen predictor del fenómeno del Niño y de la Niña. En su matriz de confusión, podemos ver un 89.61% de aciertos para los días de transición y un 90.70% de aciertos para el fenómeno de la Niña. Cabe recordar que, en este caso, no podemos tener en cuenta el fenómeno del Niño ya que no ocurrió en este periodo de 2023, aunque, por otra parte, el modelo no ha clasificado erróneamente como este fenómeno ninguna de las muestras por lo que puede considerarse de alguna forma que ha estado acertado también para este fenómeno.

## 5.6. Resultados de la clasificación

En el apartado de clasificación, hemos realizado diferentes pruebas y análisis con el objetivo de obtener el mejor modelo posible para predecir el fenómeno del Niño y de la Niña en el rango temporal del año 2023. Si recapitulamos, podemos ver que hemos realizado todas las siguientes cuestiones:

- Hemos probado diferentes clasificadores: SVC (*Support Vector Machine*), árboles de decisión y regresión logística.
- Hemos entrenado modelos con datos tanto sin escalar como escalados mediante *StandardScaler* y *MinMaxScaler*.
- Hemos ajustado hiper-parámetros para SVC (*kernel*, *C*, *degree* y *coef0*), árboles de decisión (*max\_depth* y *max\_leaf\_nodes*) y regresión logística (*C*, *penalty*, *solver* y *class\_weight*).
- Hemos entrenado los modelos con diferentes configuraciones: completo, sin la columna *sunshine\_hours*, únicamente con los datos a partir de 2004, sin la columna *clouds* y con las muestras balanceadas por tipo de fenómeno.
- Hemos obtenido la matriz de confusión y las curvas ROC de los modelos donde hemos obtenido mejor *Mean Accuracy* en la validación cruzada de cada clasificador.
- En el caso del árbol de decisión, hemos obtenido y analizado el árbol resultante, donde hemos ajustado el número máximo de profundidad del árbol y el número máximo de hojas de cada nodo.

En base a los resultados obtenidos en el subapartado anterior, donde hemos clasificado las muestras que van del mes de enero al mes de abril de 2023, hemos obtenido que, en este caso, el mejor modelo implementado es el de árbol de decisión entrenado con los datos de 2004 en adelante (en cuyas muestras ya se informa la columna *sunshine\_hours*) y podando el árbol de forma que la profundidad máxima es de 10 y la cantidad máxima de hojas por nodo es de 10. Su resultado medio de aciertos en la matriz de confusión es superior al 90%.

## 6. Conclusiones

Como primera conclusión de este trabajo, podemos decir que los datos son totalmente fehacientes, ya que, al observar distintos resultados, ha quedado demostrada la relación que existe entre las temperaturas, especialmente en la del agua del mar, en los períodos del Niño. En estos períodos, al subir la temperatura del agua, se produce humedad suficiente para que las precipitaciones en las zonas interiores se incrementen. Lo mismo ocurre en los períodos donde la temperatura del ambiente sufre incrementos debido a la acción del Niño.

Se han creado varios modelos óptimos para cada algoritmo utilizado, llegando a la conclusión de que el algoritmo SVC, además de ser costoso computacionalmente, no obtiene buenos resultados de clasificación con este conjunto de datos. Sin embargo, con regresión logística y árboles de clasificación, los resultados son más que aceptables. Además, se ha demostrado que este último es capaz de predecir cómo afectarán los distintos fenómenos en la serie temporal del año 2023, con una tasa de aciertos del 89.61% para los días de transición y un 90.70% para el fenómeno de la Niña. Este modelo se puede utilizar para predicciones de tiempo más amplias, pero la dificultad encontrada en los datos puede influir en el resultado a obtener.

Como opciones de ampliación de trabajo en este estudio, se pueden desplegar líneas de comprobación sobre la forma en que influyen cada uno de los fenómenos en los incrementos de temperatura del agua del mar o del ambiente, y realizar predicciones sobre las series de los años en cada período. Si se encuentra que el incremento es paulatino o si se producen variaciones temporales, se podrían adaptar los datos existentes y tomar decisiones al respecto. El problema existente es que cada una de estas investigaciones, por sí solas, representa mucho más tiempo de trabajo y una colaboración más estrecha con la Fundación. Nuestro trabajo se ha realizado de una forma más genérica, con pocas opciones de colaboración estrecha y poco tiempo disponible.

# Referencias

## Fundación Charles Darwin

<https://www.darwinfoundation.org/es/>

## Conjunto de datos

[https://www.darwinfoundation.org/images/climate/climate\\_puerto-ayora.csv](https://www.darwinfoundation.org/images/climate/climate_puerto-ayora.csv)

## Imagen de portada

<https://www.expedia.es/Puerto-Ayora.dx602686?gallery-dialog=gallery-open>

## Texto

[1] <https://galapagos.gob.ec/>

[2] 2017. Marcelo Hidalgo Proaño. Variabilidad climática interanual sobre el Ecuador asociada a ENOS. CienciAmérica (2017) Vol.6 (2). ISSN 1390-9592

[3] <https://www.climate.gov/>

[4] <https://niwa.co.nz/climate/information-and-resources/elnino>

[5] <https://ciifen.org/el-nino-la-nina-ciifen/#>

[6]

[https://www.inocar.mil.ec/web/index.php/articulos/770-el-nino-la-nina-enso-enos-el-nino-modoki-el-nino-canonico-el-nino-extraordinario-el-nino-godzilla-el-nino-costero-el-nino-oriental-en-que-consisten-realmente-y-como-affectan-al-ecuador](https://www.inocar.mil.ec/web/index.php/articulos/770-el-nino-la-nina-enso-enos-el-nino-modoki-el-nino-canonico-el-nino-extraordinario-el-nino-godzilla-el-nino-costero-el-nino-oriental-en-que-consisten-realmente-y-como-afectan-al-ecuador)

[7]

<http://www.ideam.gov.co/documents/21021/418818/An%C3%A1lisis+Impacto+La+Ni%C3%B3n/a.pdf/640a4a18-4a2a-4a25-b7d5-b3768e0a768a>