

NAHUEL GHILARDI

CODERHOUSE

GRANIZO

<https://github.com/Nahuelito22>

TABLA DE CONTENIDO

01

Introducción

Contexto, objetivo
y público objetivo

02

Preguntas

Hipótesis y preguntas
de investigación

03

Metadata

Características del
Dataset utilizado.

04

Análisis exploratorio

Exploración y
visualización de datos

05

Modelado

Selección y entrenamiento
del mejor modelo.

06

Conclusiones

Resultados y
recomendaciones

BIENVENIDOS

¿Qué buscamos lograr?

Este proyecto se enfoca en la predicción de tormentas de granizo, un fenómeno climático que puede tener un impacto significativo en la agricultura, la infraestructura y la seguridad.

Nuestra motivación

El proyecto nace de la necesidad de contar con herramientas más precisas para predecir tormentas de granizo, especialmente en regiones como Mendoza, Argentina, donde este fenómeno puede afectar gravemente las cosechas y, por ende, la economía agrícola.



OBJETIVOS

El principal objetivo de este proyecto es desarrollar un modelo de clasificación que sea capaz de predecir la probabilidad de granizo (granizo = 1 / sin granizo = 0), basándose en un conjunto de variables climáticas.



1



1



0



HIPOTESIS

Se plantea que variables como la latitud, longitud, temperatura, precipitación, altitud y presión atmosférica pueden ser claves en la predicción de la ocurrencia de granizo.

Específicamente, se espera que ciertas combinaciones de estas variables puedan permitir una predicción precisa, dada la correlación histórica entre estos factores y los eventos de granizo en diversas regiones.

DATASET

¿Qué se utilizo?

Para este proyecto utilizamos un Dataset que contiene la frecuencia del granizo por estación meteorológica y el año de la medición.

Además del uso de una api para conseguir mas datos meteorológicos.

Estructura



Id_Estación



Nombre
de la estación



Latitud



Longitud



Año del
registro



Frecuencia
de granizo



Temperatura
Media Anual



Temperatura
Máxima Anual



Temperatura
Mínima Anual



Viento
en Km/h



Presión
en Hpa



Precipitación
en mm



Altitud



Granizo_Si_No

CONSIDERACIONES

Cabe aclarar que el dataset conseguido es de [Mazher, Romina Nahir](#). Pero este solo tenía datos sobre las estaciones(Id_Estacion, Latitud y longitud de las estaciones, Año de registro). Y lo mas importante la frecuencia de granizo por año

Lo que se hizo para conseguir el resto de información fue utilizar la api de [Meteostat](#), mediante código se unifico todo y se creo la variable objetivo respecto a la frecuencia.

A continuación vamos a ver un ejemplo:

Id_Estacion	Latitud	Longitud	Año	Frecuencia	Granizo_Si_No
10196	-36.75	-59.83	1934	3	1
10196	-36.75	-59.83	1935	0	0
10196	-36.75	-59.83	1936	1	1





GRAFICOS

Cuando analizamos el Dataset primero utilizamos un grafico para ver que estaciones tenían mas nulos.

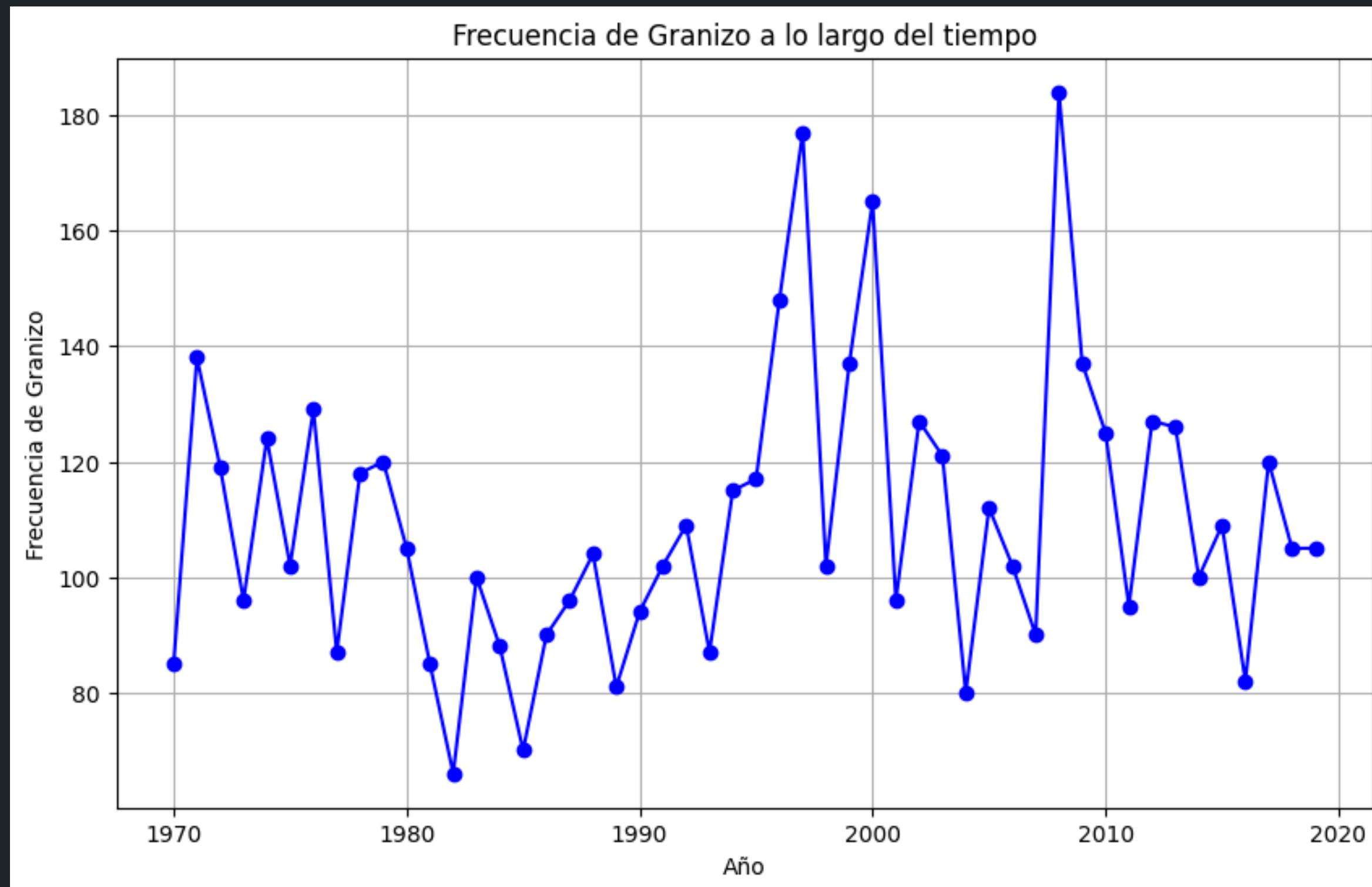
Y si su ubicación era determinante o se podía remplazar con otra cercana.

Como vemos obtuvimos:

- Estación sin datos nulos
- Estación con datos nulos

La decisión que se tomo fue eliminar las estaciones que perjudicaban el Dataset ya que eran pocas y tenían un remplazo relativamente cerca

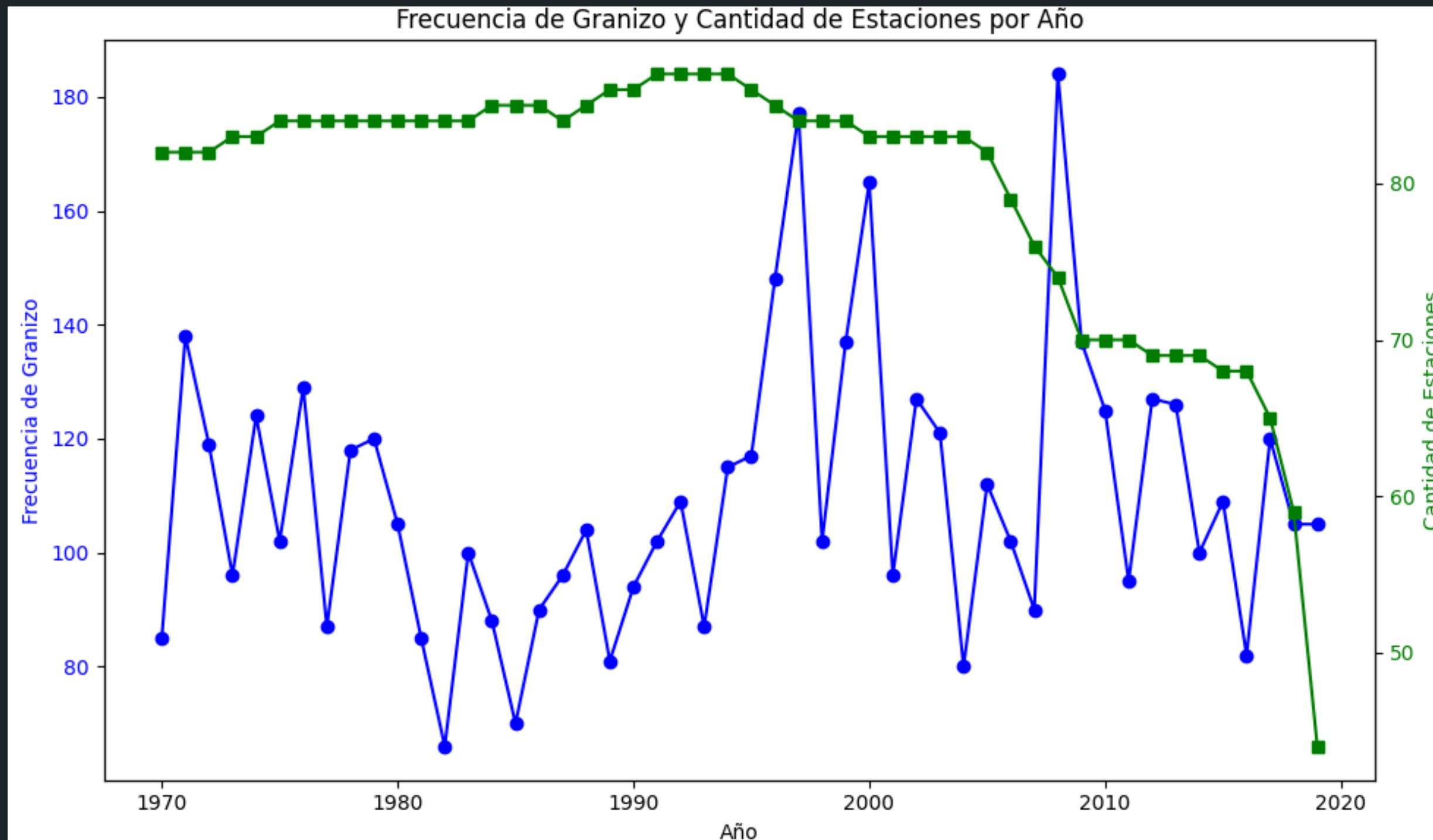
FRECUENCIA EN EL TIEMPO



En este grafico vemos como fue evolucionando la frecuencia del granizo a lo largo del tiempo, esto nos puede dar una idea de si existen patrones temporales y que hacer con ellos.

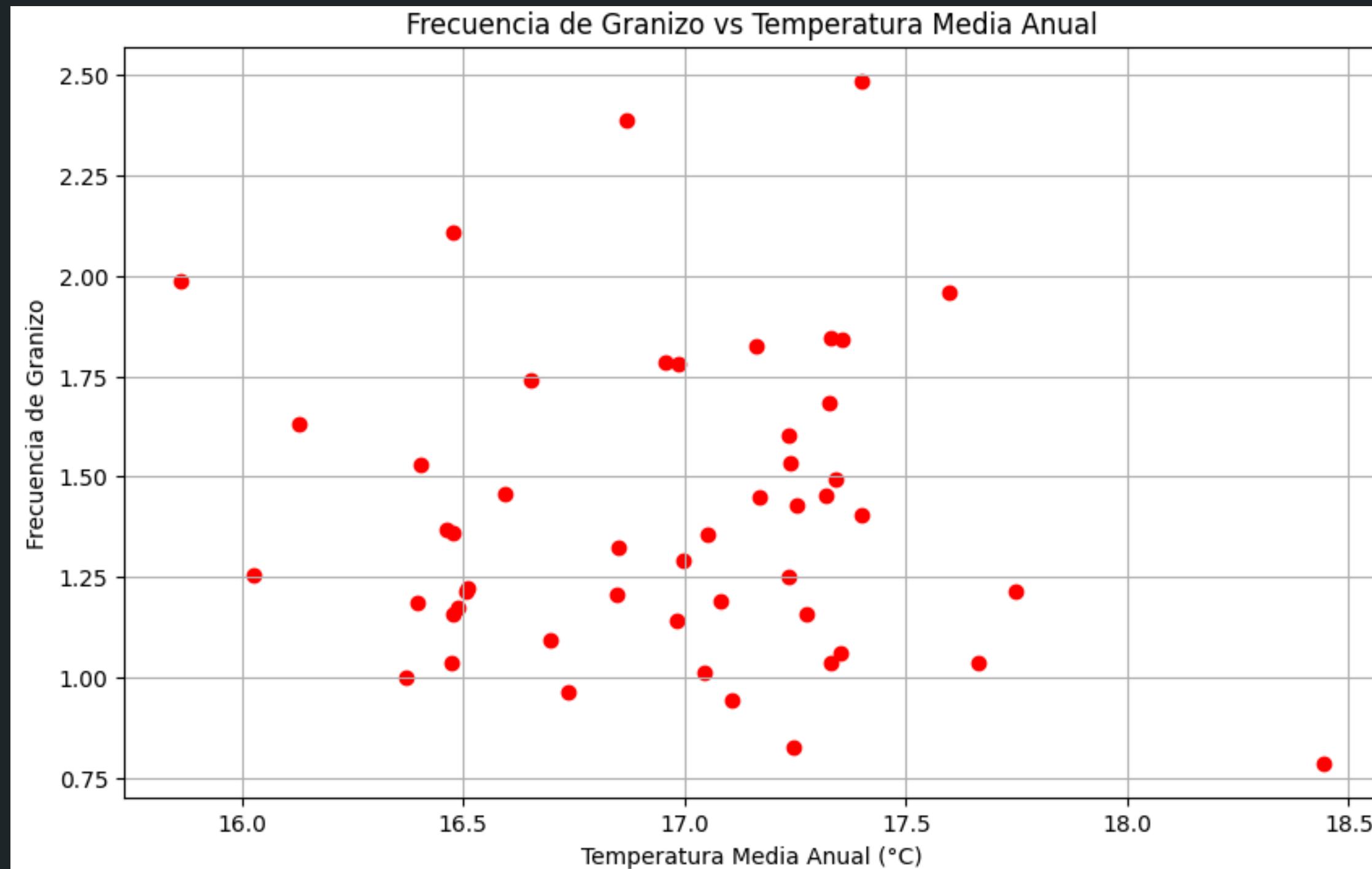
FRECUENCIA VS ESTACIONES

En este grafico vemos como la fluctuación de frecuencias no se debe a un aumento o reducción de la cantidad de estaciones que registran las tormentas, incluso la cantidad de estaciones se redujo.

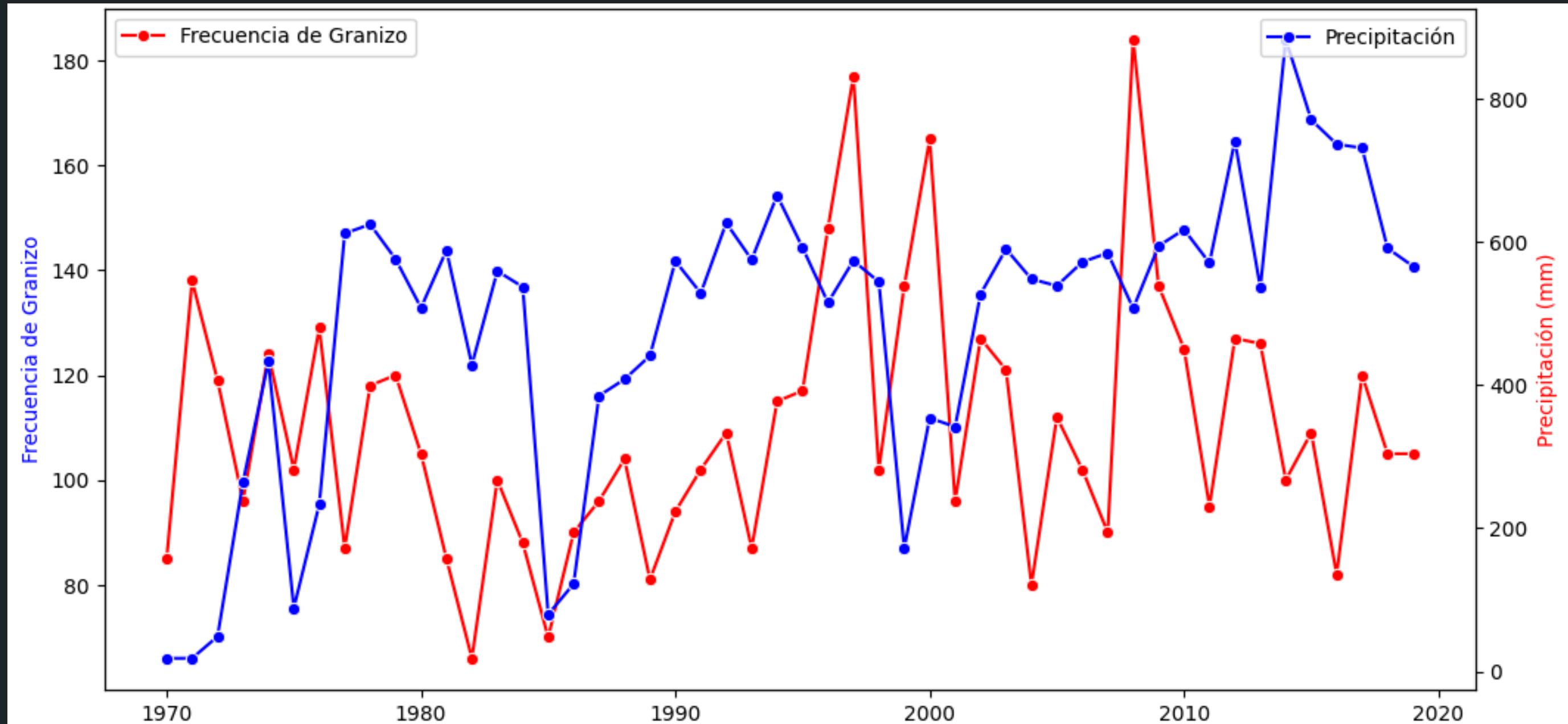


TEMPERATURA Y FRECUENCIA

Como vemos a continuación, no detectamos un patrón climático que defina en gran porcentaje a un conjunto de datos de frecuencia, pero si se puede observar un leve agrupamiento entre los 16.6 ° y los 17°

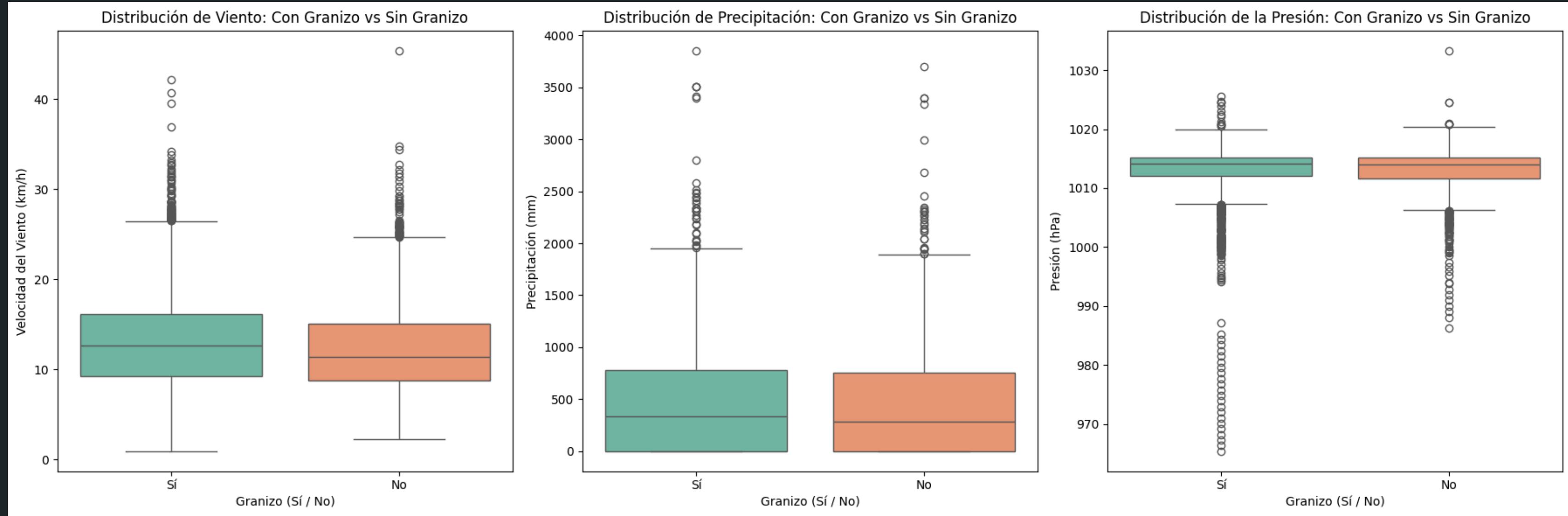


¿LA PRECIPITACIÓN AFECTA?

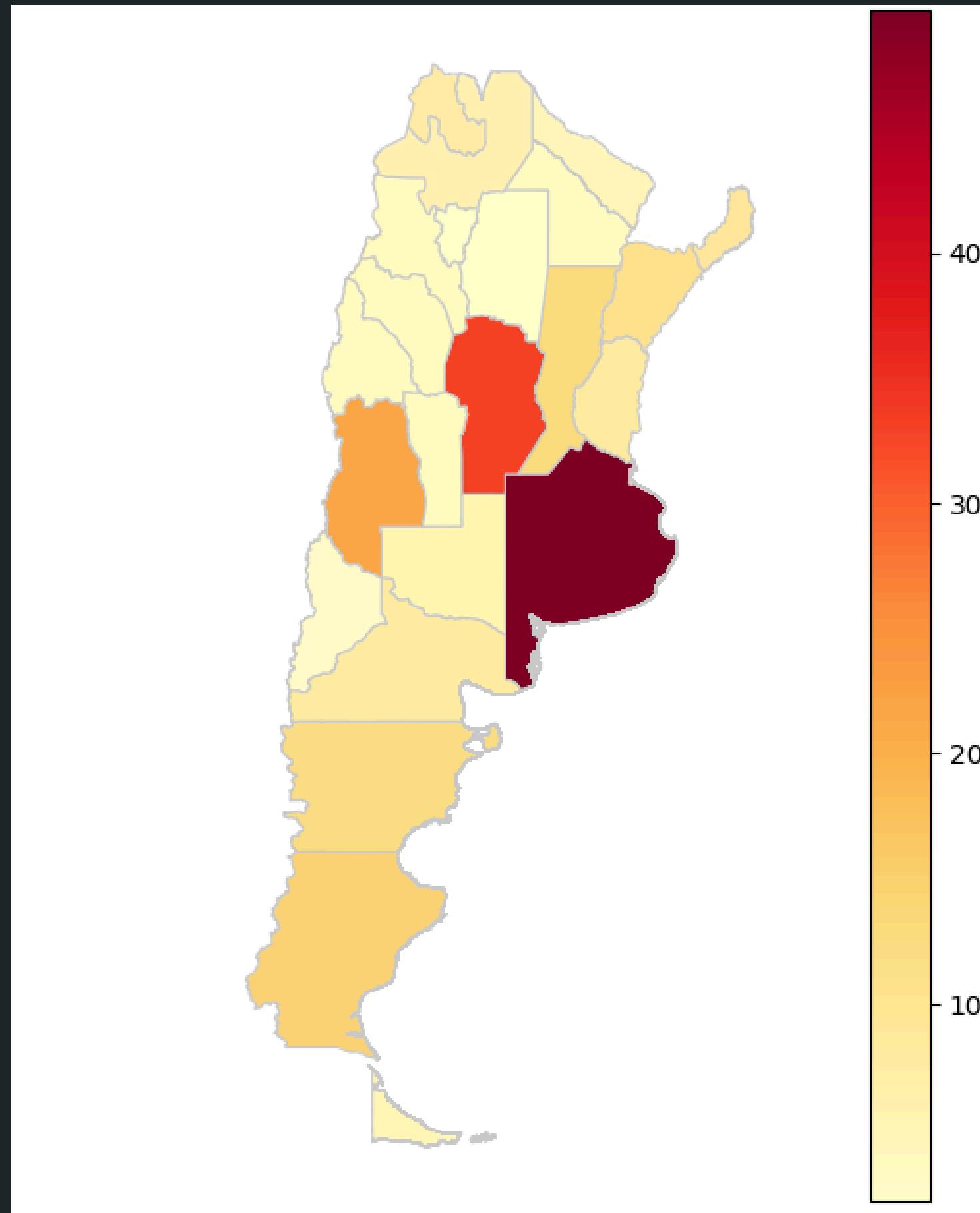


Algo que se suele pensar es que hay una fuerte relación entre la precipitación y las tormentas de granizo, pero como se observa en estos gráficos encontrar un patrón puede ser bastante complejo, para eso sería mejor ver el impacto de otras variables.

VIENTO, PRECIPITACION Y PRESION



Como se ve en estos gráficos, para el viento y la precipitación no existe un grupo tan separado de otro, pero al observar la presión esta si crea un patrón mas notorio, lo cual tiene sentido ya que las tormentas de granizo se suelen formar en sistemas de baja presión



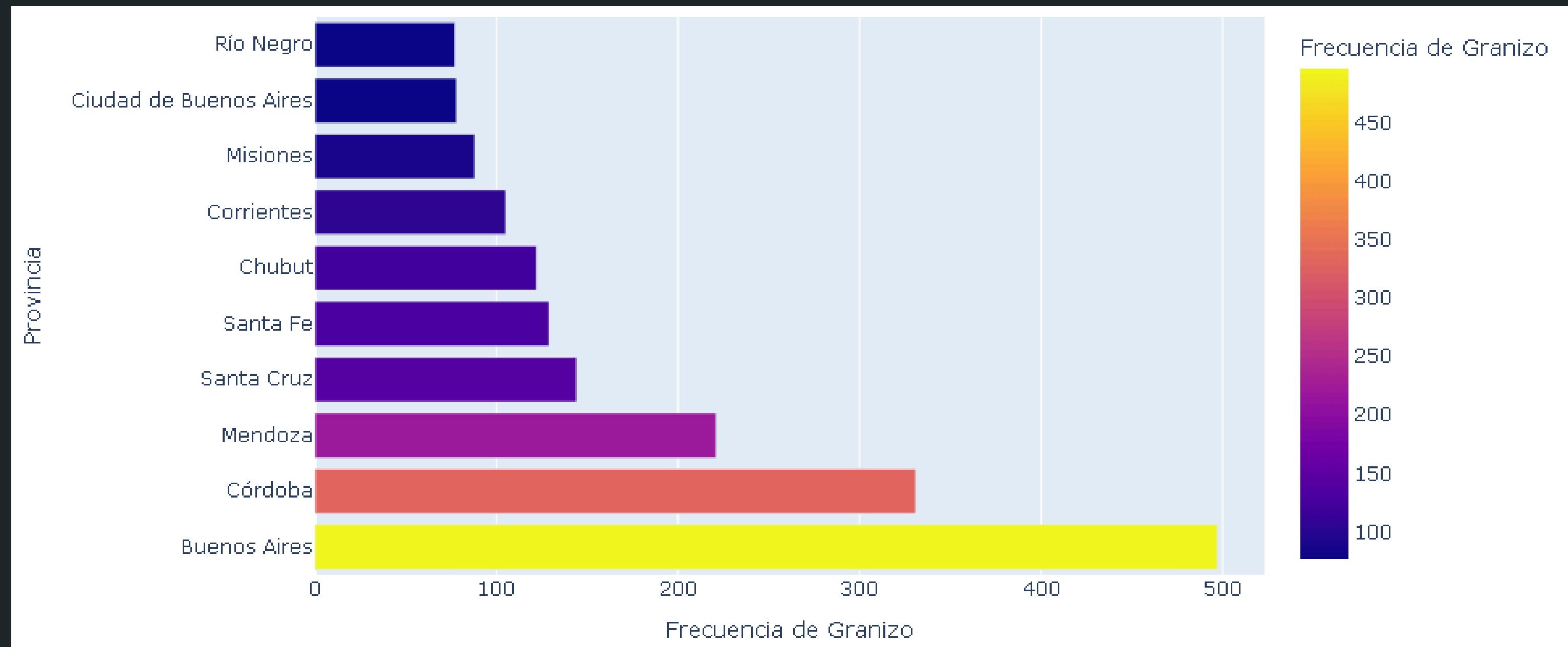
¿QUE PROVINCIAS SON MAS AFECTADAS?

Como observamos en este grafico estilo mapa de calor, las provincias mas afectadas son:

- Buenos Aires
- Córdoba
- Mendoza

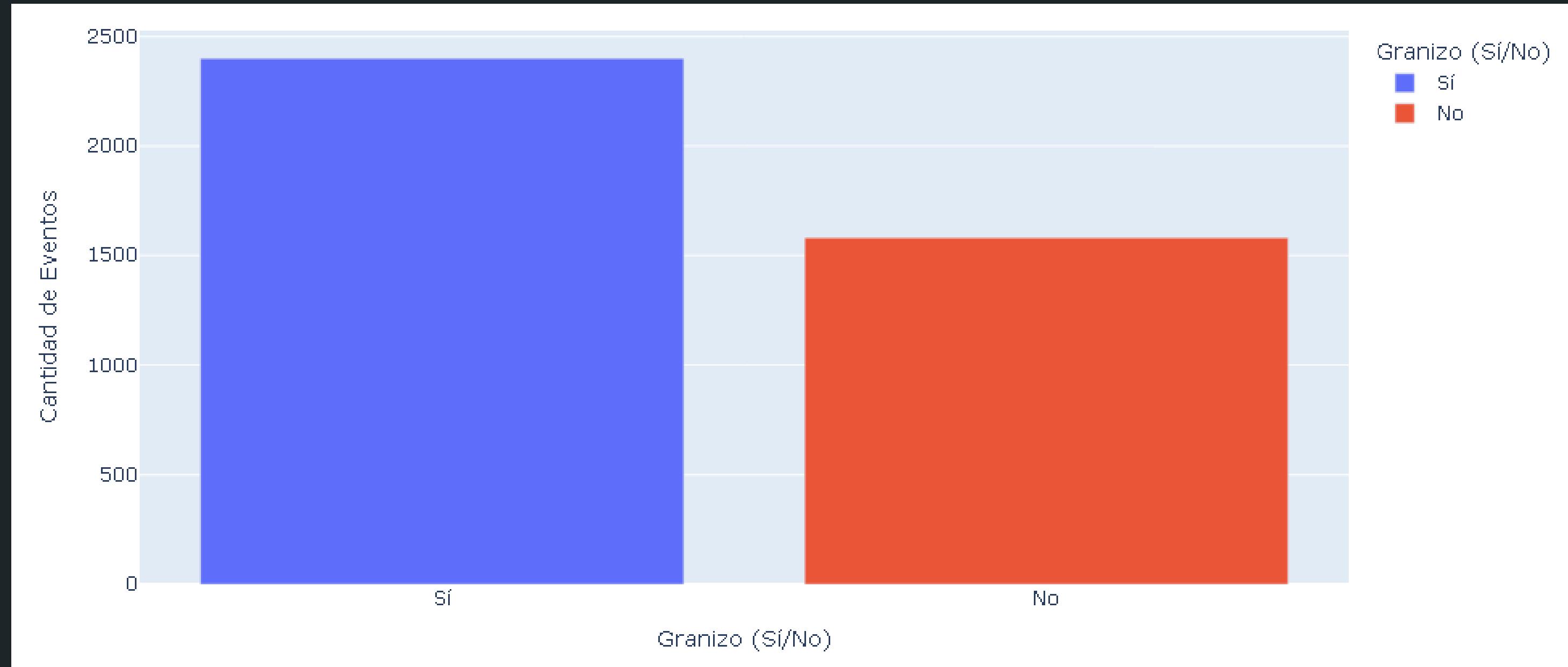
Igualmente estaría bueno ve un top 10 sobre esto.

TOP 10 PROVINCIAS MAS AFECTADAS



Este grafico nos es de gran ayuda ya que podemos ver el impacto que tendría generar un modelo capaz de ayudar a predecir futuras tormentas y lograr reducir el daño que estas causan.

¿TENEMOS CLASES DESCOMPENSADAS?



Como vemos en el grafico tenemos una clase mayoritaria, esto para los modelos no es bueno, así que vamos a optar por técnicas para equilibrar las clases

EQUILIBRAR DATOS

Como pudimos observar anteriormente, tenemos una clase descompensada, esto quiere decir que tiene mas registros que la otra. Teniendo esto en consideración optamos por utilizar 4 técnicas para equilibrar las clases, y así generar 4 Dataset independientes, con el fin de ver que técnica es mejor.

Nuestras tecnicas utilizadas y datasets resultantes fueron:

- Dataset original: Guardamos una copia del original sin modificar nada.
- Sobre muestreo: Utilizamos una técnica llamada SMOTE para equilibrar la clase minoritaria a la mayoritaria generando datos sintéticos
- Sub muestreo: Utilizamos RandomUnderSampler, esta técnica reduce aleatoriamente los registros de la clase mayoritaria.
- Equilibrado: Utilizamos SMOOT-ENN, esta técnica es una fusión de las dos anteriores para equilibrar los datos.



¿QUE SE OBTIENE?

Dataset original

47.784
registros

Dataset con sobremuestreo

57.624
registros

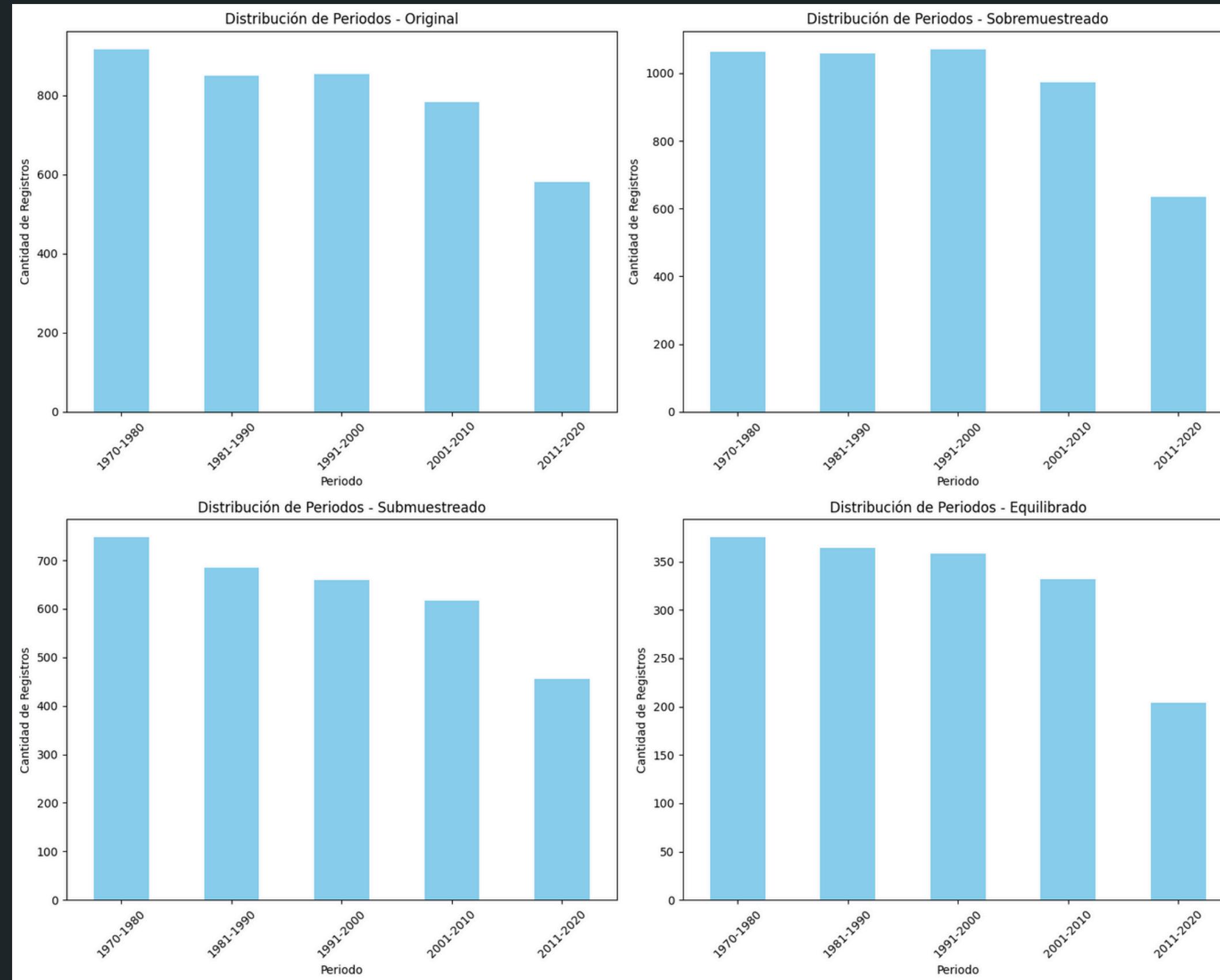
Dataset con submuestreo

37.994
registros

Dataset equilibrado

20.568
registros

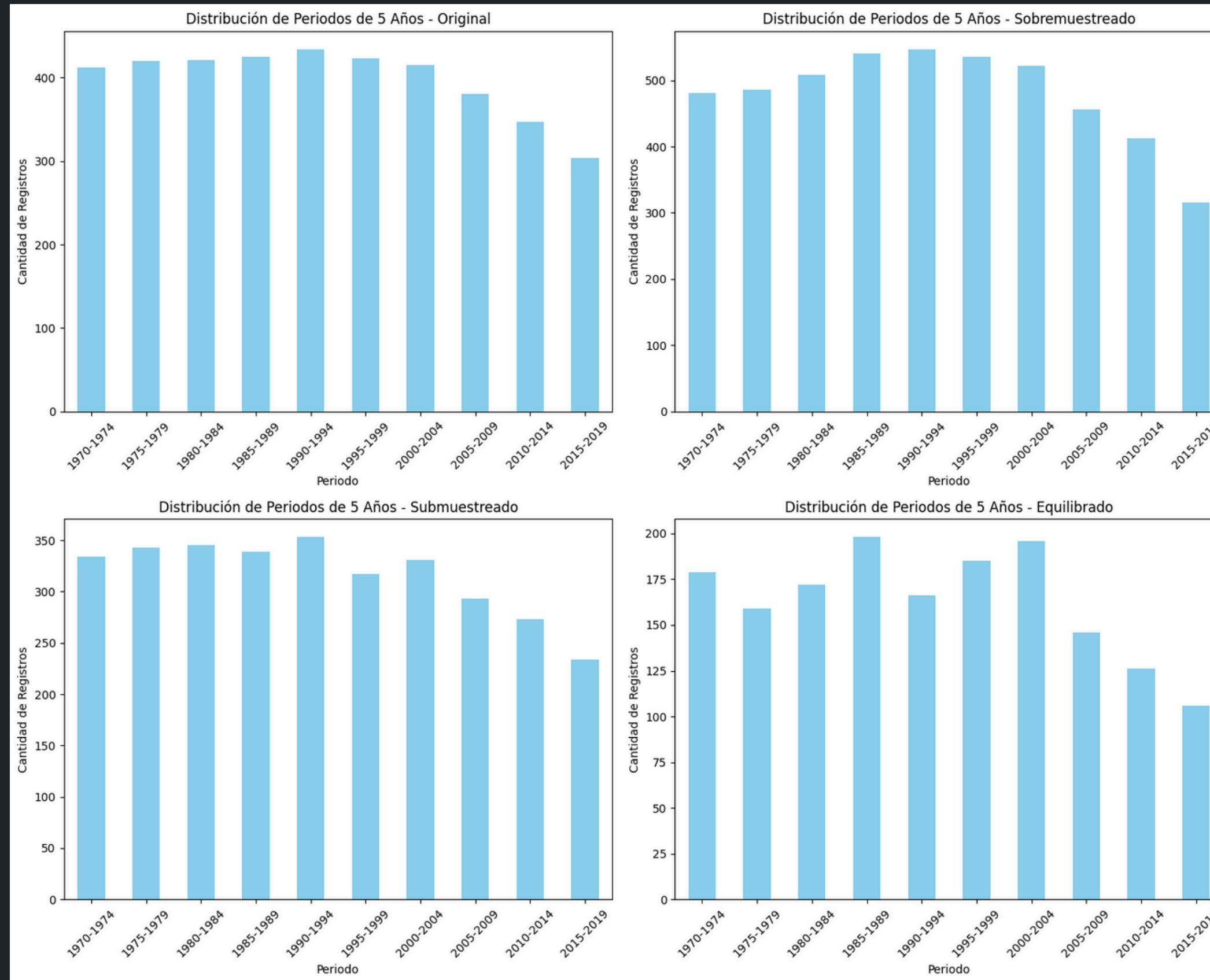
PERIODO DE 10 AÑOS



Como vimos anteriormente una vez que tenemos los 4 DataSets independientes, vamos a ver si existe alguna relación con el tiempo, sin recaer a series temporales.

Nuestro enfoque fue dividir cada Dataset independiente en periodo de 10 años, entre 1970 y 2020.

PERIODO DE 5 AÑOS



Como se observa lo bueno de dividir en periodo de 5 años es que podemos evitar caer en periodos que no reflejen mucha información y así mismo ampliar la profundidad del modelo.



ENTRENAMIENTO Y TESTEO

Ahora dividimos cada Dataset nuevo con sus respectivos periodos de tiempo en datos de entrenamiento y testo.

Los datasets que conseguimos serian:

Periodo de 10 años	Periodo de 5 años
Original(x,y)	Original(x,y)
Sobremuestreo(x,y)	Sobremuestreo(x,y)
Submuestreo(x,y)	Submuestreo(x,y)
Equilibrado(x,y)	Equilibrado(x,y)

MODELOS

Los modelos elegidos para este proyecto fueron:



Regresión Logística



Arboles de decisión



Bosque Aleatorio



XGBoost



CatBoost



LGBM

GRIDSEARCHCV

Utilizamos una gridsearch para comprobar cual es el mejor modelo sobre cada DataSets, como vemos en el ejemplo a continuación:

Lista de datasets

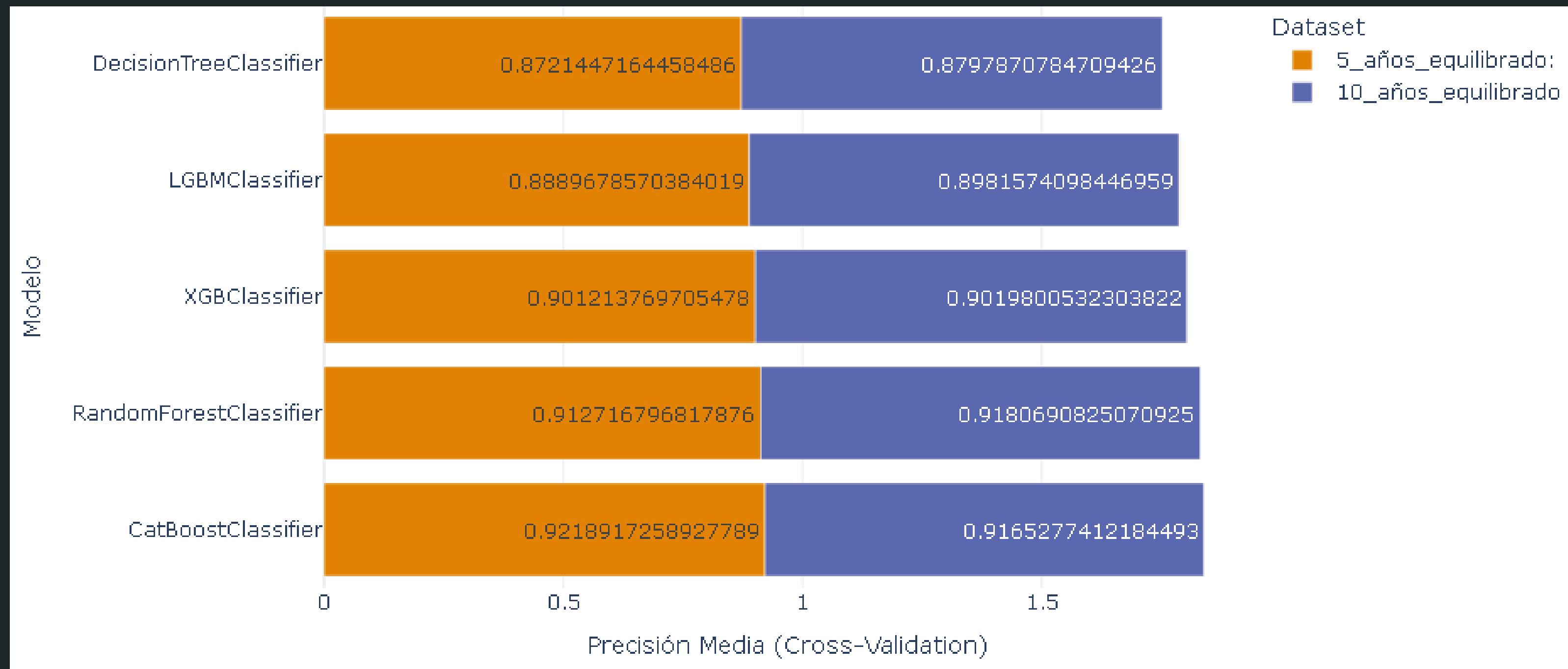


Lista de modelos y sus hiperparámetros



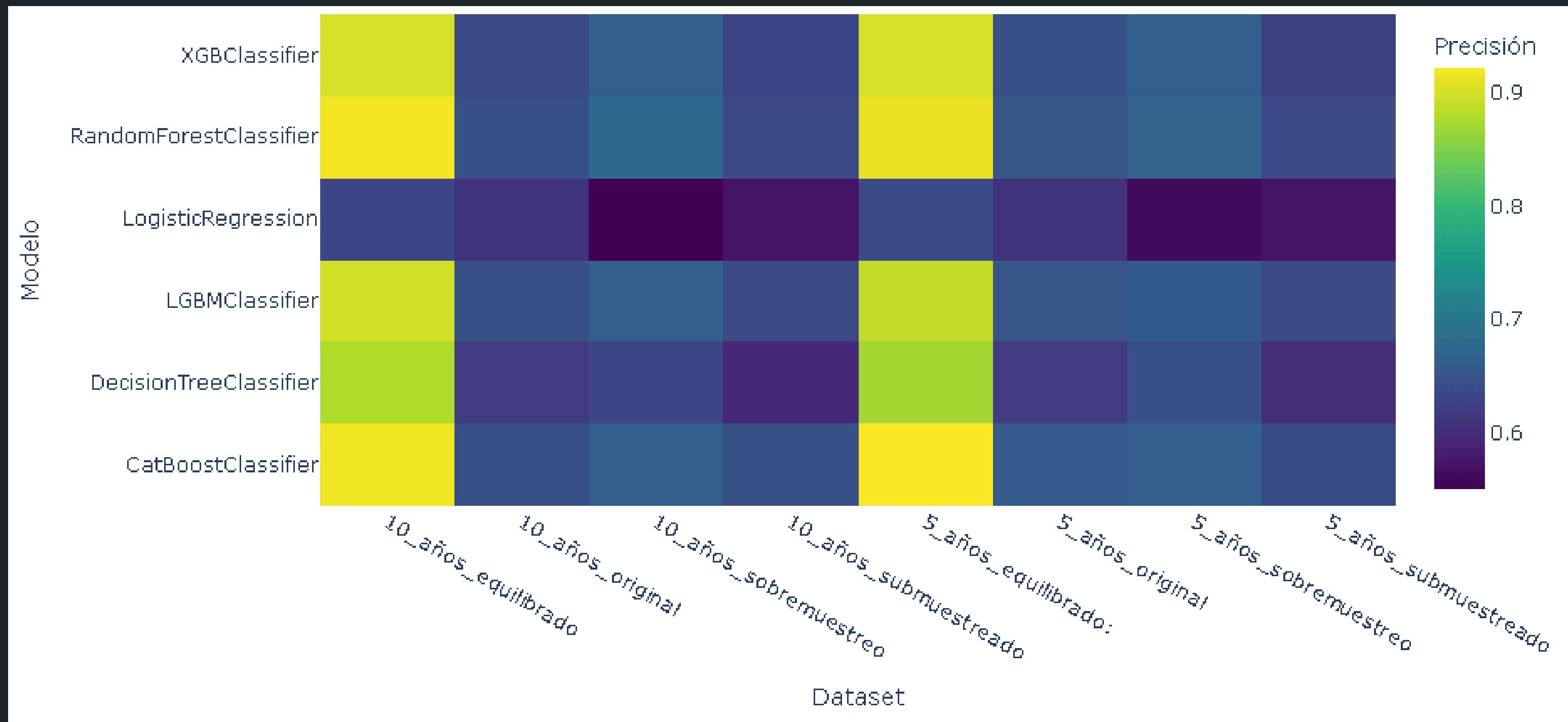
El resultado de esto es una lista con los mejores modelos sobre cada datasets y que hiperparámetros utilizar

TOP 10 MODELOS



Como podemos observar los DataSets ganadores fueron, 5 y 10 años respectivamente, y el modelo ganador es CatBoost con una performance de 0.92 en validación cruzada.

TODOS LOS MODELOS



Como podemos observar en este grafico de calor, los modelos como regresión logística para este caso no tienen buena performance, además también podemos concluir en que es importante tener en cuenta la técnica para equilibrar clases.

ENTRENAMIENTO Y EVALUACION

Casos	Precisión	Recall	F1-Score
0	0.94	0.88	0.91
1	0.89	0.94	0.91

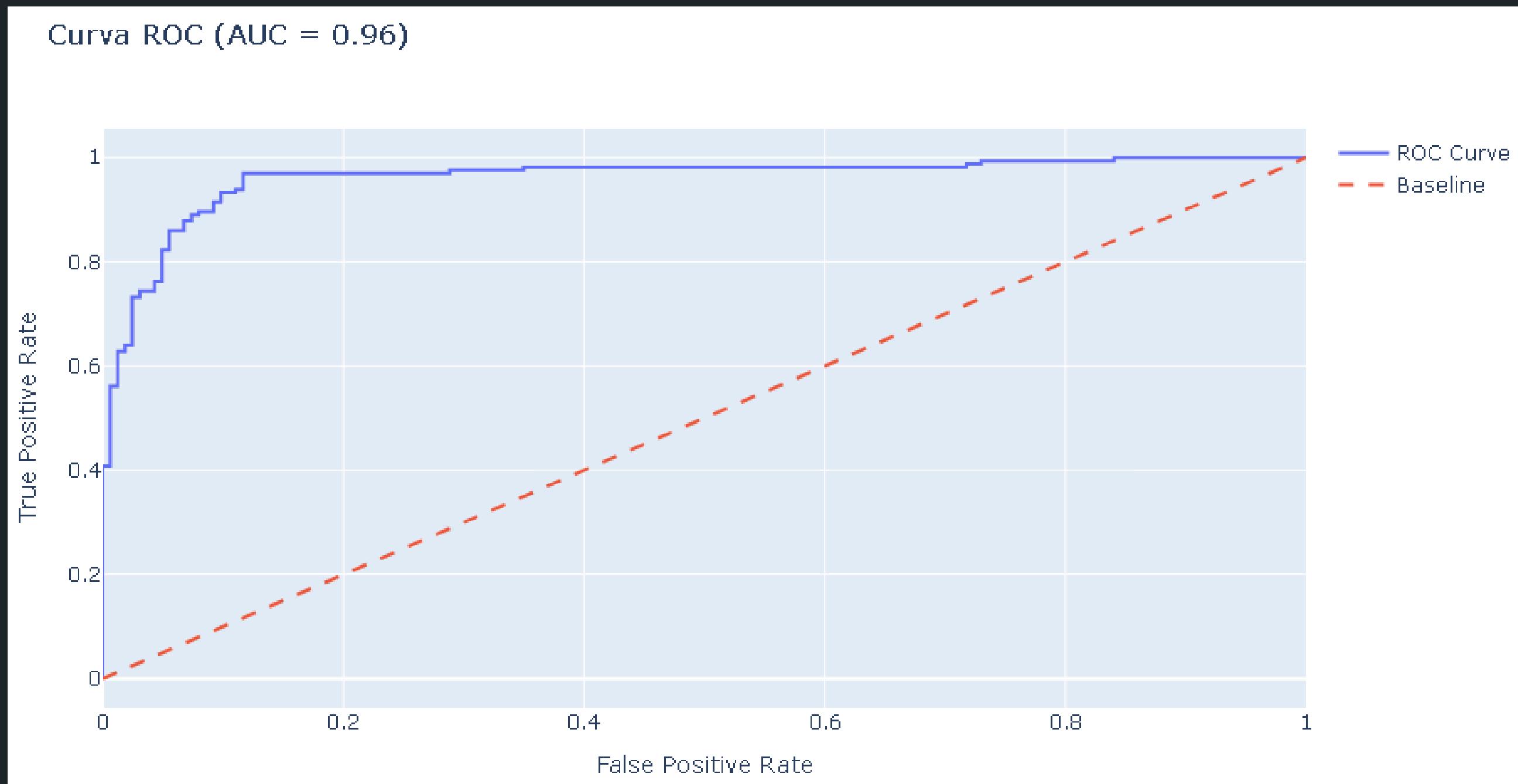
- Precisión (Precision): El modelo presenta un alto porcentaje de aciertos para la clase "0"
- Recall (Sensibilidad): El modelo demuestra un gran porcentaje de detección para los casos positivos, lo que lo hace efectivo en la reducción de falsos negativos.
- La combinación de precisión y recall no es una métrica clave en este caso, ya que el dataset ganador no presenta un desequilibrio significativo de clases.

ENTRENAMIENTO Y EVALUACION



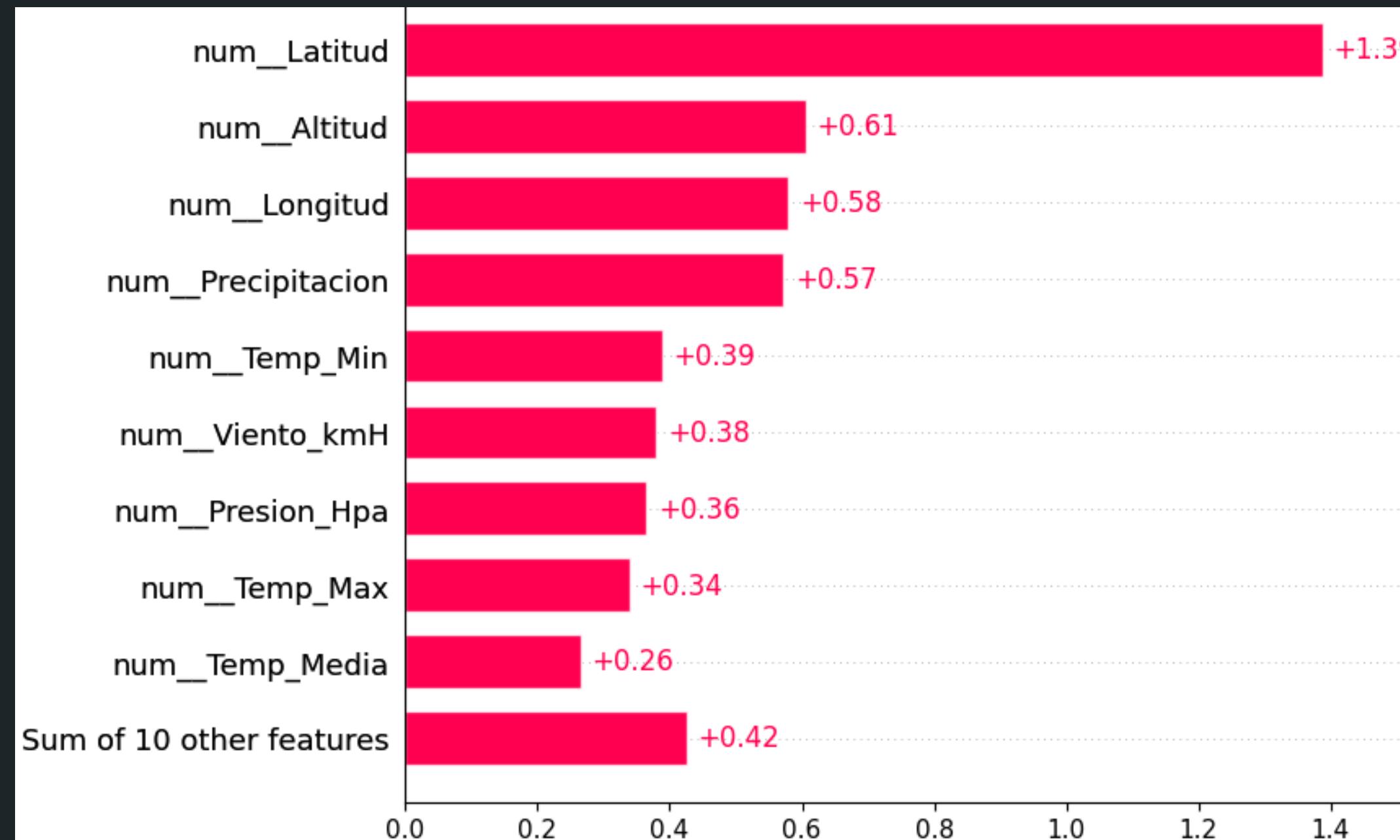
Como vemos en esta matriz de confusión el modelo per forma bastante bien y tiene pocos casos de falsos positivos y falsos negativos

ENTRENAMIENTO Y EVALUACION



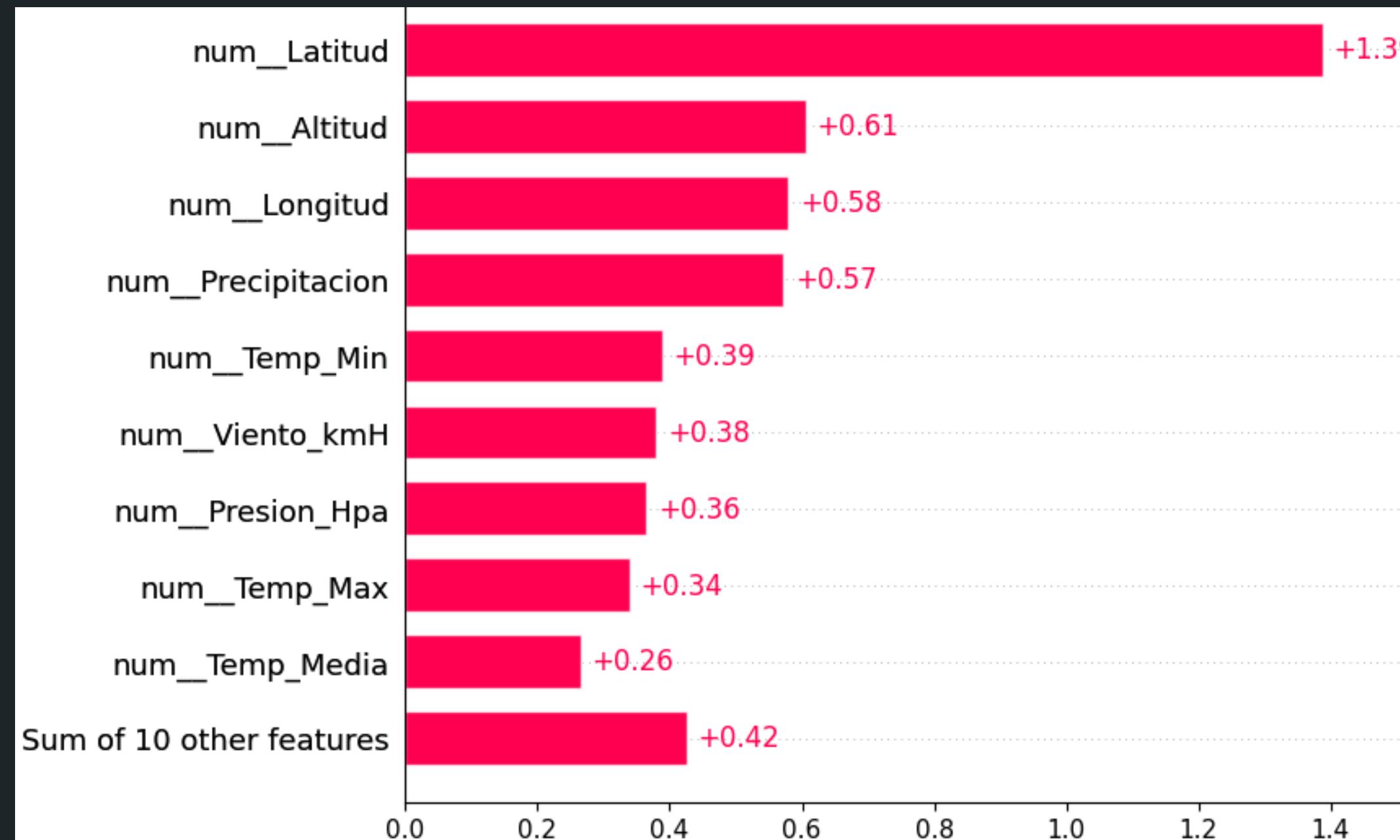
Como se ve en nuestra curva Roc el modelo tiene una excelente capacidad para separar las clases positivas y negativas, ya que su puntaje fue de 0.96

INTERPRETABILIDAD



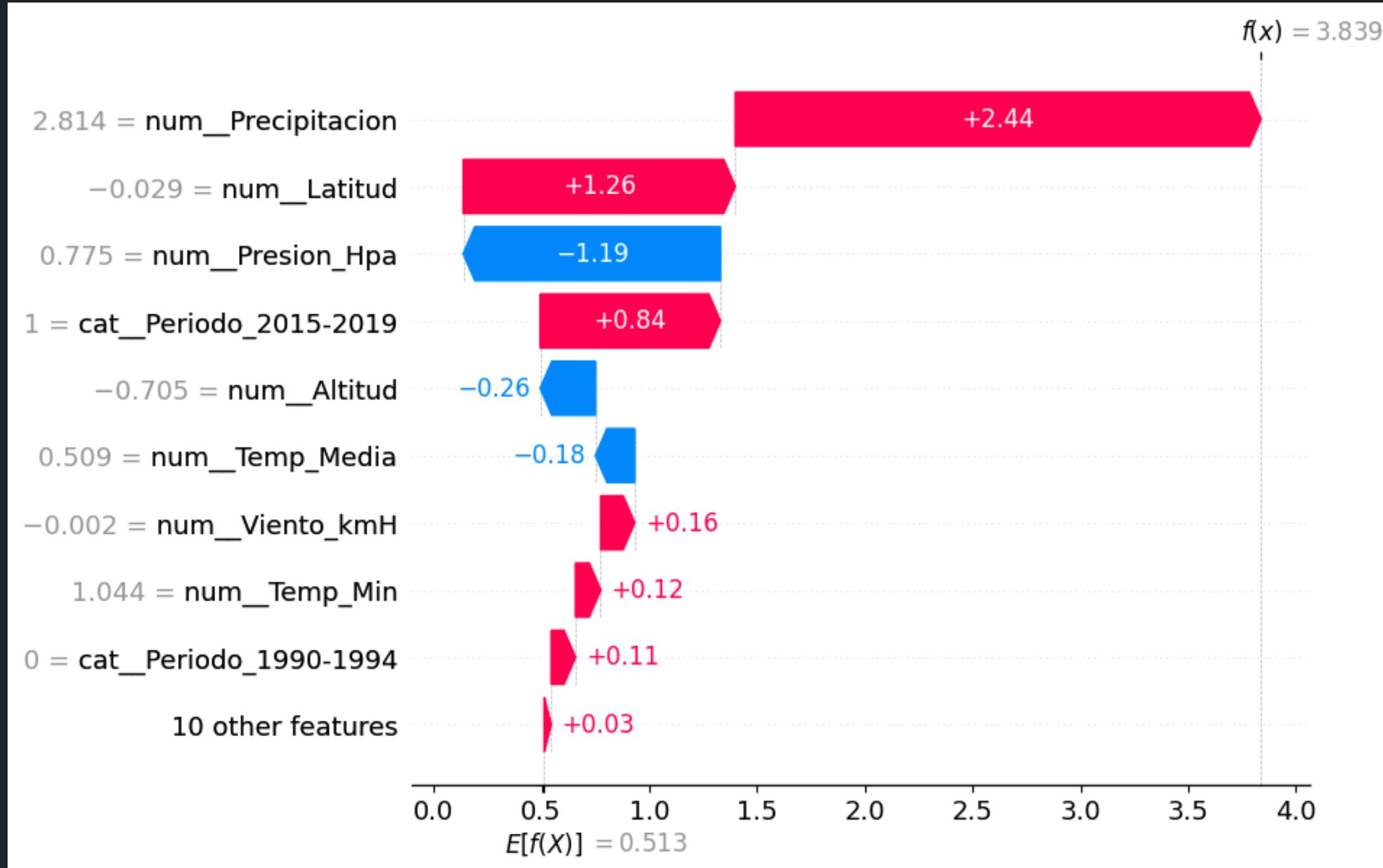
El modelo muestra que la latitud y longitud son variables clave, especialmente por la alta incidencia en Buenos Aires. Al excluir estas variables, la precipitación gana importancia, revelando patrones significativos en las zonas más afectadas.

INTERPRETABILIDAD



El modelo muestra que la latitud y longitud son variables clave, especialmente por la alta incidencia en Buenos Aires. Al excluir estas variables, la precipitación gana importancia, revelando patrones significativos en las zonas más afectadas.

INTERPRETABILIDAD LOCAL



Como se observa de manera local no siempre todo esta dado por latitud y longitud, en este caso, para el registro 1. Vemos una fuerte relacion del granizo o la posibilidad de una tormenta de granizo con la precipitación.

CONCLUSIONES

Este proyecto logró desarrollar un modelo predictivo eficaz para identificar la probabilidad de tormentas de granizo, destacando el rol clave de variables como latitud, longitud, altitud y presión. Los resultados obtenidos no solo demuestran la precisión del modelo, sino también su utilidad práctica en la prevención y gestión de riesgos climáticos. Este trabajo refleja el valor del análisis de datos para abordar desafíos ambientales y generar soluciones con impacto positivo en la sociedad.

RECOMENDACIONES

Hemos logrado grandes resultados pero si es cierto que esto recién es un comienzo, ya que esta beta solo contiene datos anuales y generar predicciones para un uso real no es tan eficaz, la idea es que esto crezca y se puedan reunir mas datos así como utilizar imágenes satelitales para generar mas confiabilidad en las predicciones, etc.

Gracias

<https://github.com/Nahuelito22>