Proyecto final

Escenario climático 2025 a partir del análisis estacional de datos de años previos.

Introducción a la Ciencia de Datos

Integrantes del equipo:

Lidia Gizem Sánchez Montiel

Emir Jahaziel Santiago Patricio

Profesor:

Jaime Alejandro Romero Sierra

25/11/2024

Introducción

El análisis de datos climatológicos es de gran importancia para diversos sectores, especialmente aquellos cuyo desempeño depende directamente de las condiciones climáticas. Entre ellos, el sector agrícola enfrenta desafíos significativos debido a las variaciones del clima, como la distribución desigual de lluvias o períodos prolongados de sequía. Este proyecto tiene como objetivo analizar los registros de precipitaciones desde el año 2000 en adelante, con el propósito de proporcionar información útil que permita a los agricultores identificar las épocas con lluvias abundantes y escasas, facilitando así una mejor planificación de sus actividades.

La relevancia de este estudio radica en la necesidad de contar con proyecciones climáticas precisas en muchos países, donde la falta de escenarios confiables dificulta la toma de decisiones informadas. Estas proyecciones no solo pueden ayudar a anticipar variaciones estacionales, sino que también son fundamentales para prevenir riesgos asociados a pérdidas económicas, escasez de recursos hídricos y daños en los cultivos. A través de un análisis exploratorio de datos climáticos y la generación de escenarios futuros, este proyecto busca aportar herramientas clave para mitigar el impacto del cambio climático en la agricultura y fomentar la sostenibilidad del sector.

Este proyecto también incluye el diseño de un dashboard interactivo para facilitar la exploración y visualización de los datos analizados.

Fuente de datos

Para este proyecto, se utilizó la base de datos titulada **“Climate Insights Dataset”**, la cual recopila registros numéricos relacionados con diversas variables climáticas. Esta base de datos incluye información detallada sobre tendencias y mediciones climáticas, abarcando un amplio rango temporal y geográfico, lo que la hace adecuada para un análisis exhaustivo del comportamiento del clima en las últimas décadas.

Características del Dataset:

- **Rango temporal:** 1678 a 2022.
- **Cobertura geográfica:** Incluye datos de múltiples ubicaciones y países.
- **Variables principales:**
    1. **Fecha:** Registro de tiempo de cada observación.
    2. **Ubicación:** Identificación geográfica específica.

3. **País:** Clasificación a nivel nacional.
4. **Temperatura (°C):** Medición de la temperatura ambiente.
5. **Emisiones de CO2 (ppm):** Concentración de dióxido de carbono en la atmósfera.
6. **Aumento del nivel del mar (mm):** Variación en el nivel del mar a lo largo del tiempo.
7. **Precipitaciones (mm):** Cantidad de lluvia acumulada.
8. **Humedad (%):** Porcentaje de humedad relativa en el aire.
9. **Velocidad del viento (km/h):** Intensidad del viento en kilómetros por hora.

Cantidad de Datos:

- **Cantidad original:** 10,000 registros.
- **Cantidad después de limpieza:** 9,474 registros, tras la eliminación de duplicados y tratamiento de valores faltantes.

Tras el proceso de limpieza, el dataset no presenta valores faltantes ni duplicados, y las variables están correctamente formateadas para el análisis.

Metodología

El desarrollo de este proyecto se dividió en las siguientes etapas, implementadas utilizando herramientas como Python y librerías especializadas (Pandas):

1. **Exploración inicial de los datos**
   En esta etapa se realizó un análisis preliminar para comprender las características de la base de datos:

- **Dimensiones y estructura del dataset**: Se revisaron el tamaño de la base de datos, los tipos de variables, y la cantidad de valores faltantes mediante comandos como .info() y .describe().
- **Identificación de valores nulos y duplicados**: Se detectaron registros con valores nulos en columnas clave, así como posibles duplicados mediante el comando .duplicated().

- **Clasificación de variables**: Se categorizaron las columnas como numéricas, categóricas o temporales para definir estrategias de análisis posteriores.

2. **Limpieza de datos**
   En esta etapa se implementaron técnicas básicas para garantizar la calidad del análisis:

- **Eliminación de valores duplicados**: Se eliminaron registros duplicados utilizando el comando .drop_duplicates() para asegurar la consistencia de los datos.
- **Manejo de valores nulos**: Los valores faltantes en las columnas numéricas se imputaron utilizando la media o mediana, dependiendo de la distribución de los datos.
- **Formateo de la columna de fecha**: La columna Fecha fue convertida al formato datetime para facilitar análisis basados en el tiempo.

3. **Implementaciones**
   Para poder ajustar nuestra base de datos al modelo que usamos de Machine Learning y dashboard, fue necesario crear columnas de Longitud y Latitud a partir de las columnas de Ubicación y País. También creamos una columna de Mes que corresponde al mes dado en la columna Fecha y también dos columnas que corresponden a los eventos de precipitaciones y temperatura que nos ayudaron a determinar los registros relevantes por país y por mes.

Análisis exploratorio de datos:

Esta será la etapa central del proyecto, en la cual se analizarán a detalle las características principales de las variables para obtener insights clave.

Descripción general de los datos.

- **Visión General:** Resumen del dataset, incluyendo el número total de registros y variables.

```
data=pd.read_csv("base_limpia_proyecto.csv")
data
```
✓ 4.1s

	Fecha	Ubicación	País	Temperatura	Emisiones de CO2	Aumento del Nivel del mar	Precipitaciones	Humedad	Velocidad del viento
0	2000-01-01	New Williamtown	Latvia	10.688986	403.118903	0.717506	13.835237	23.631256	18.492026
1	2000-01-01	North Rachel	South Africa	13.814430	396.663499	1.205715	40.974084	43.982946	34.249300
2	2000-01-02	West Williamland	French Guiana	27.323718	451.553155	-0.160783	42.697931	96.652600	34.124261
3	1678-01-01	South David	Vietnam	12.309581	422.404983	-0.475931	5.193341	47.467938	8.554563
4	2000-01-05	South Nathan	Saint Helena	6.229326	392.473317	1.122210	76.368331	48.973886	30.398908
...	...	...	...	...	...	...	...	...	...
9469	2014-08-09	Heatherfort	Kenya	9.633273	536.879317	-1.055269	0.733654	35.386679	11.349757
9470	2022-03-26	South Jeffrey	Norway	18.322140	533.633336	-0.154583	40.539896	95.178150	33.091834
9471	2017-05-11	North Lauren	Saint Martin	15.828247	445.608149	0.081255	25.880205	6.341108	42.817475
9472	2010-04-08	Port Laura	Georgia	14.932330	366.906446	0.363282	42.474523	94.672587	42.797399
9473	2003-07-05	Coleside	Romania	11.983638	517.075204	0.185722	39.697657	31.336638	46.092629

9474 rows × 9 columns

o **Tipos de Variables:** Clasificación de las variables.

```
data.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9474 entries, 0 to 9473
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Fecha                     9474 non-null   object
 1   Ubicación                 9474 non-null   object
 2   País                      9474 non-null   object
 3   Temperatura               9474 non-null   float64
 4   Emisiones de CO2          9474 non-null   float64
 5   Aumento del Nivel del mar 9474 non-null   float64
 6   Precipitaciones           9474 non-null   float64
 7   Humedad                   9474 non-null   float64
 8   Velocidad del viento      9474 non-null   float64
 9   Latitud                   8989 non-null   float64
 10  Longitud                  8989 non-null   float64
dtypes: float64(8), object(3)
memory usage: 814.3+ KB
```

o **Resumen Estadístico:** Estadísticas descriptivas.

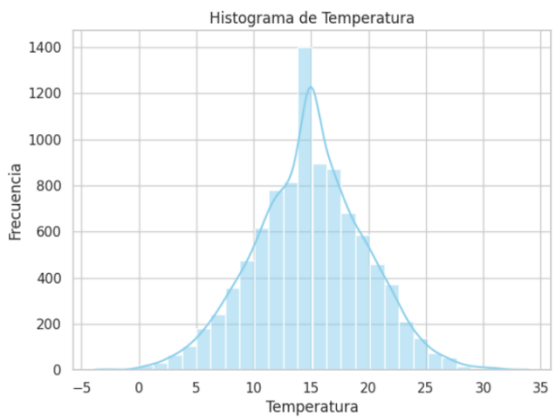
```
data.describe()
✓ 0.0s
```

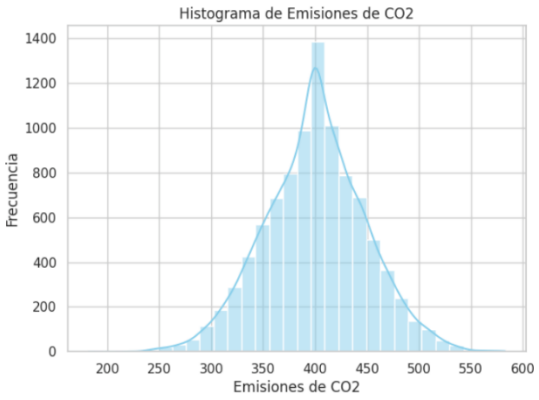
	Temperatura	Emisiones de CO2	Aumento del Nivel del mar	Precipitaciones	Humedad	Velocidad del viento	Latitud	Longitud
count	9474.000000	9474.000000	9474.000000	9474.000000	9474.000000	9474.000000	8989.000000	8989.000000
mean	14.900554	400.278444	-0.001910	49.837706	49.815610	25.168255	15.513630	9.995284
std	4.879772	48.598603	0.958409	28.330937	28.076813	14.158351	25.778215	76.296039
min	-3.803589	182.131220	-4.092155	0.010143	0.018998	0.001732	-54.843286	-176.204224
25%	11.771866	368.875036	-0.619916	25.422358	26.348809	13.207691	-0.525231	-59.525030
50%	14.932330	400.157142	-0.005637	49.862571	49.705318	25.086836	15.926666	14.447691
75%	18.009023	431.570011	0.621775	73.690124	73.824189	37.143033	36.800207	51.229529
max	33.976956	582.899701	3.849427	99.991900	99.959665	49.997664	64.984182	179.158292

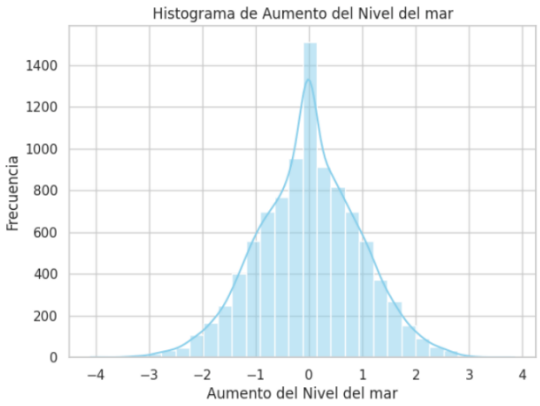
Visualización y Distribución de Variables Individuales.

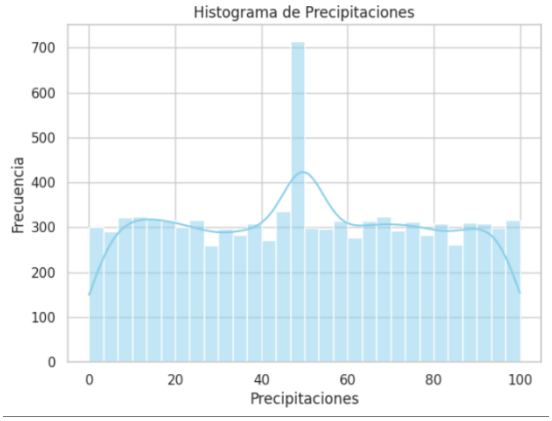
Variables Numéricas:

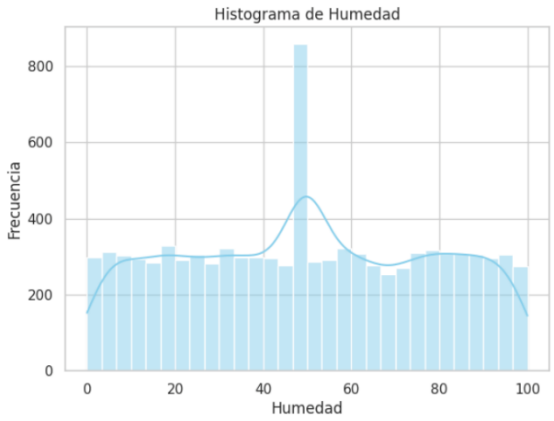
- **Histogramas:**

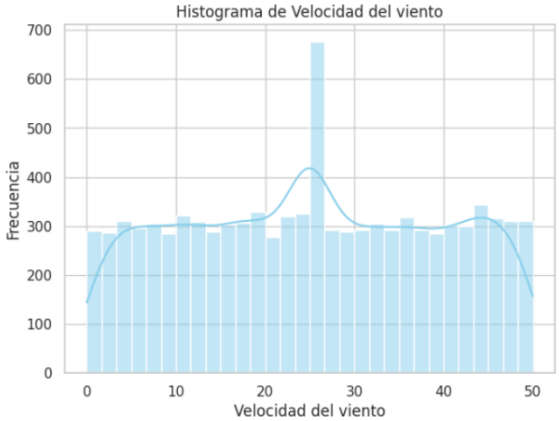






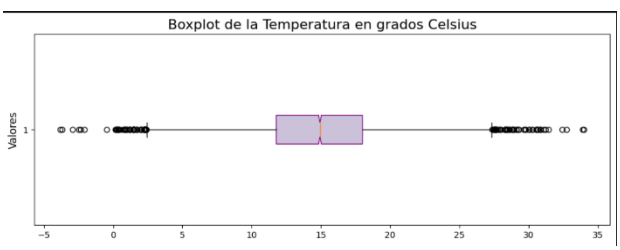


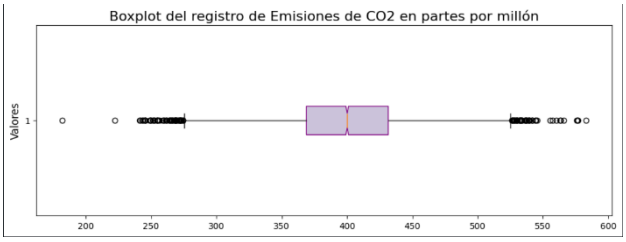


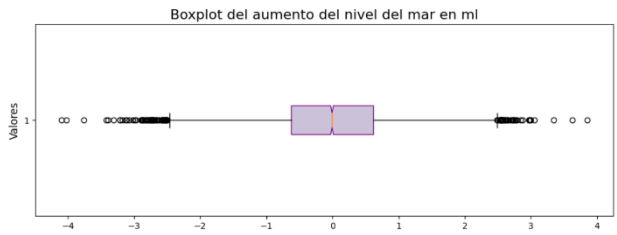


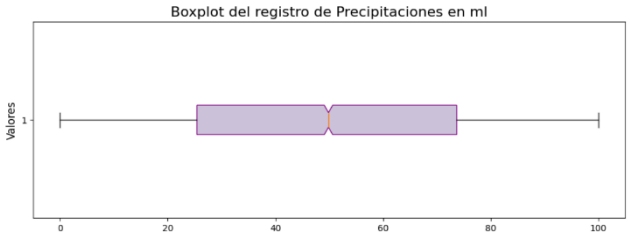
Descripción general: Las variables como temperatura, CO2 y aumento del nivel del mar son bastante estables, con valores que no varían mucho. Por otro lado, variables como las precipitaciones, humedad y velocidad del viento tienen "dos comportamientos", como si cambiaran dependiendo de la temporada o las condiciones del clima. Esto nos da una idea de qué esperar en términos generales y nos prepara para analizar más detalles, como qué pasa en cada estación o región.

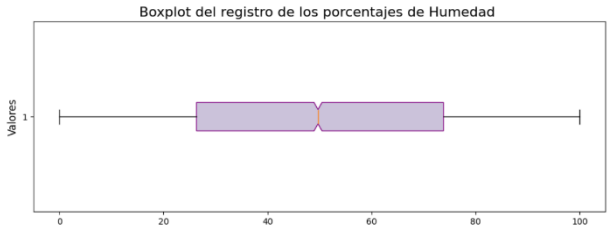
- **Boxplots:**

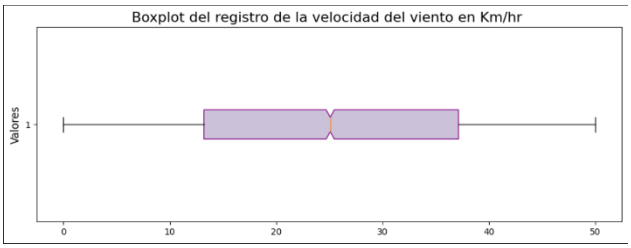






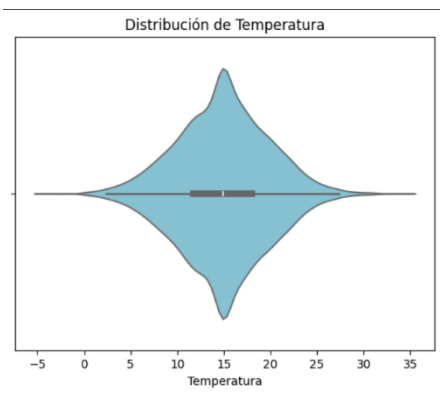


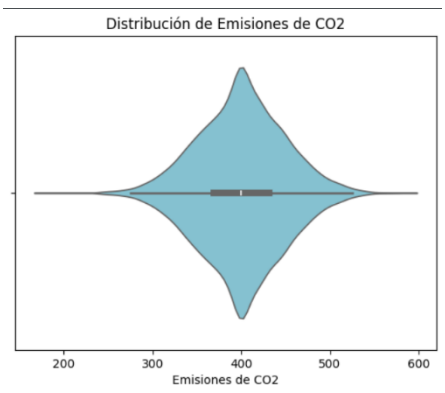


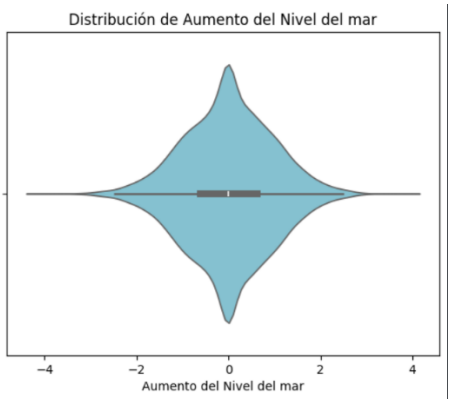


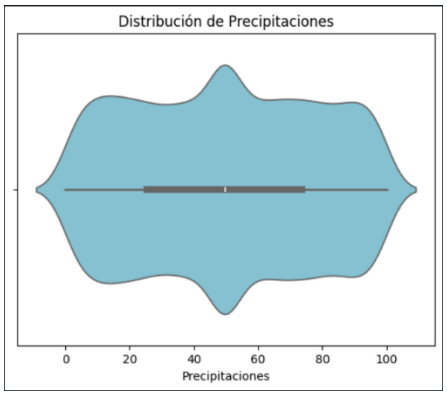
Descripción general: La mayoría de las variables presentan distribuciones consistentes con rangos bien definidos, excepto temperatura, nivel del mar, y emisiones de CO2, que muestran valores atípicos significativos. Estos valores atípicos pueden ser importantes para analizar eventos extremos.

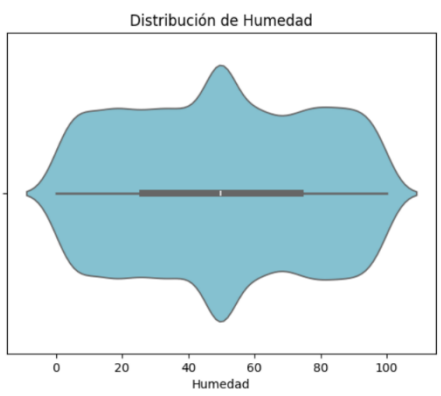
- **Diagramas de Violín:**

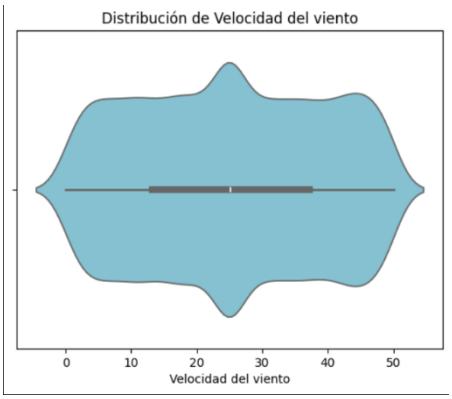










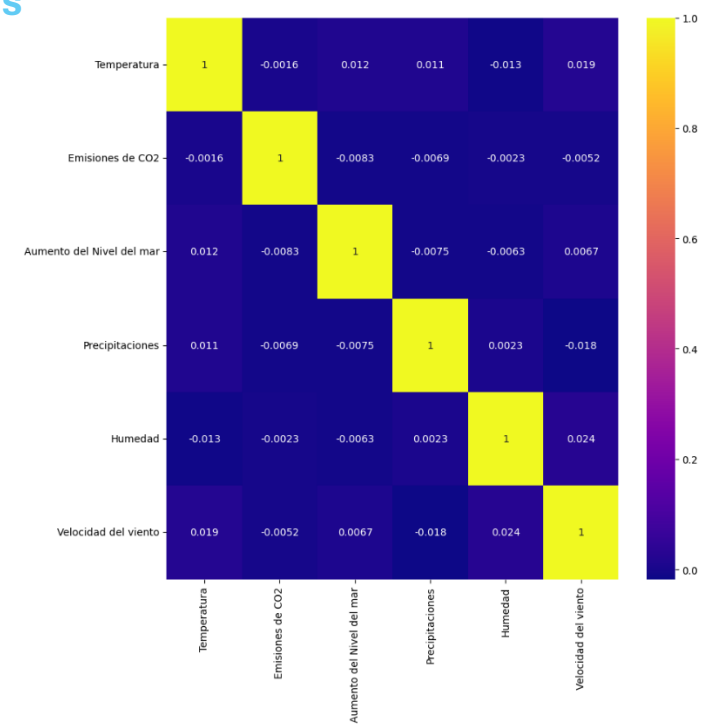


Descripción general: Los diagramas de violín nos muestran cómo están distribuidos los valores de las diferentes variables climáticas de manera clara:

1. Temperatura, Emisiones de CO2 y Aumento del Nivel del Mar:
    - La mayoría de los datos están concentrados en un rango medio.
    - Los valores extremos (muy altos o muy bajos) son poco frecuentes.
2. Precipitaciones, Humedad y Velocidad del Viento:
    - Tienen dos comportamientos principales (bimodalidad), lo que significa que hay dos grupos de valores comunes.
    - Esto podría estar relacionado con diferencias entre temporadas o regiones.

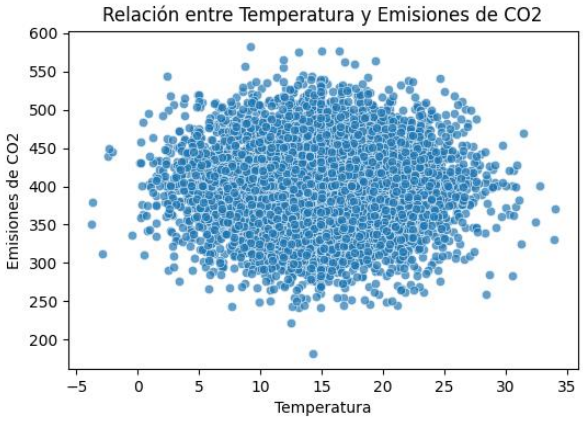
Correlación entre Variables

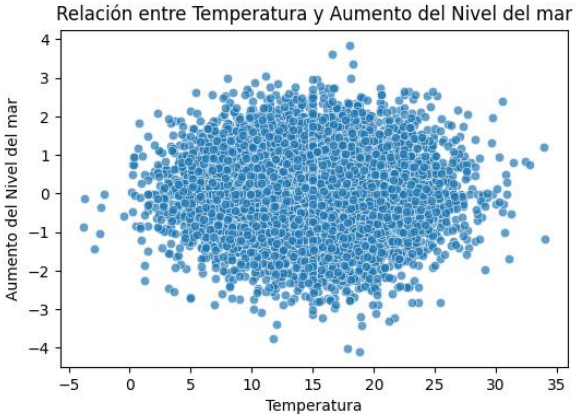
Matriz de Correlación (heatmap):

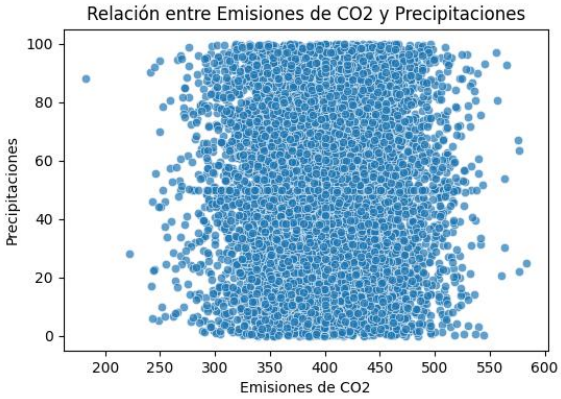


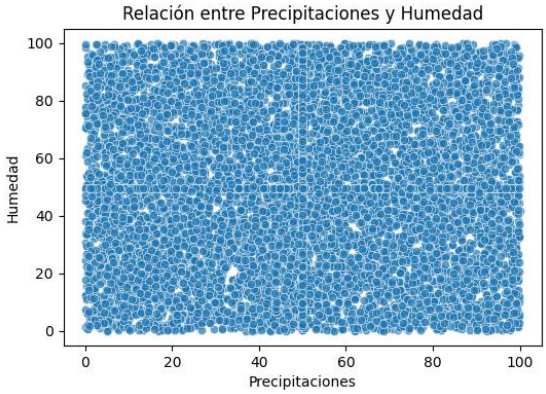
Descripción general: El heatmap indica que las variables no están correlacionadas de manera lineal. Esto podría significar que las relaciones entre estas variables, si existen, podrían ser no lineales o influenciadas por otros factores no representados en los datos. La baja correlación entre variables indica que no están altamente relacionadas entre sí, lo que es bueno para evitar colinealidad en el modelo.

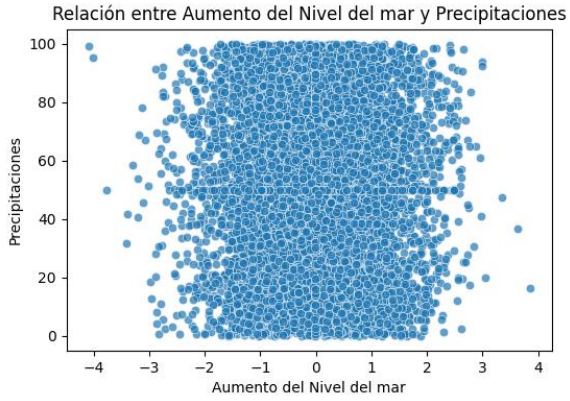
Parejas de Variables: Gráficos de dispersión.

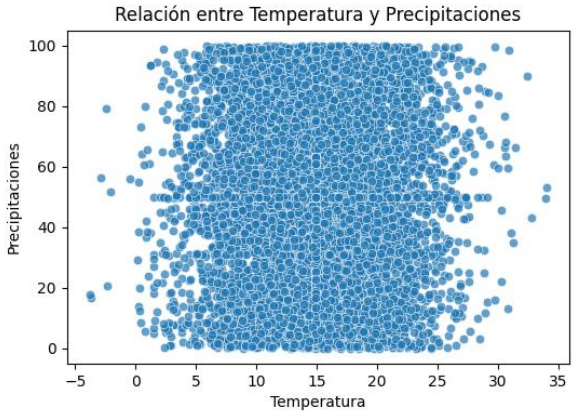


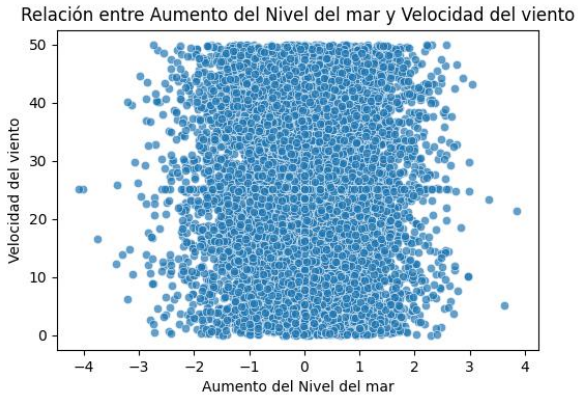


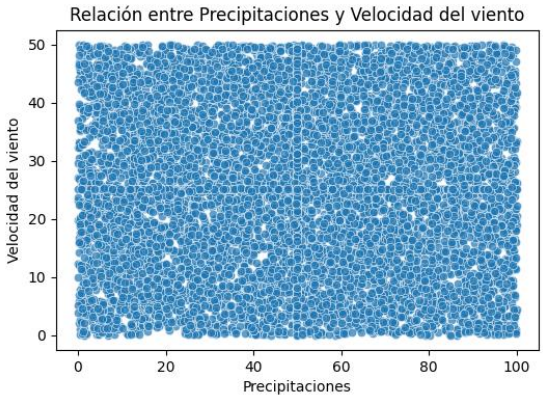


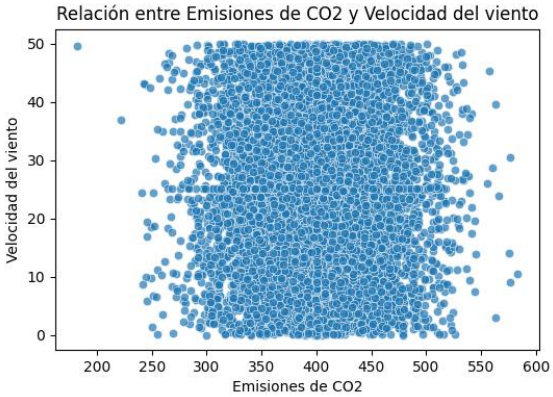


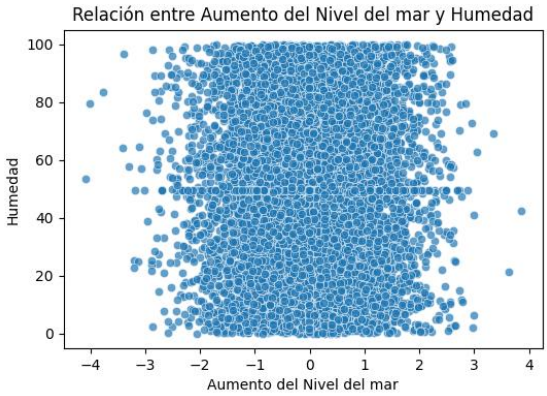


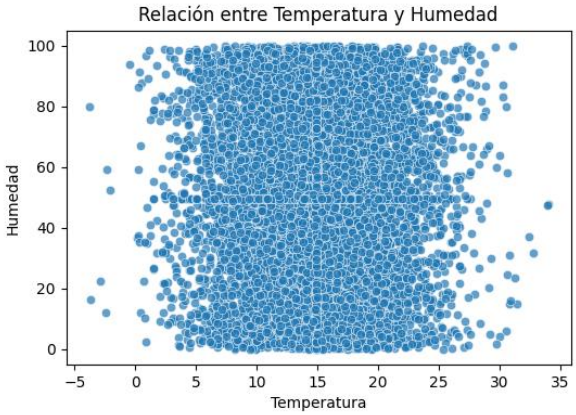


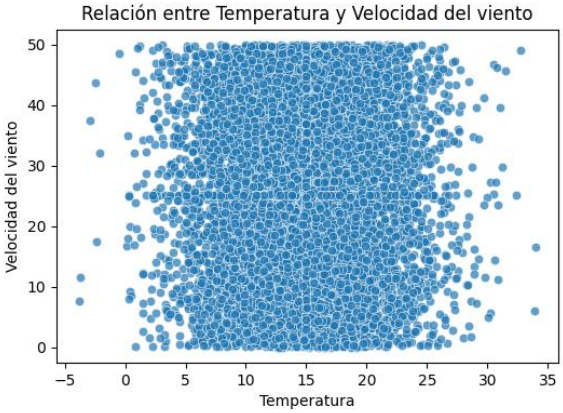

Relación entre Aumento del Nivel del mar y Velocidad del viento
Velocidad del viento
Aumento del Nivel del mar

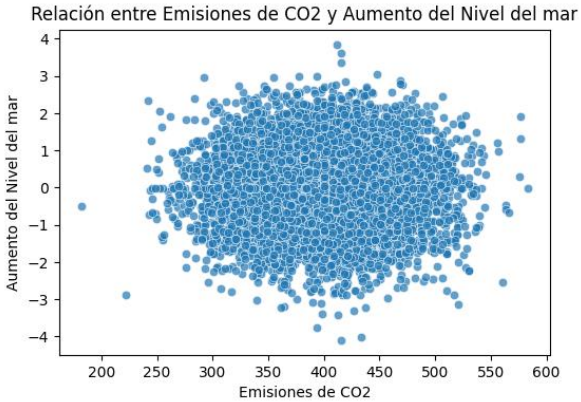

Relación entre Precipitaciones y Velocidad del viento
Velocidad del viento
Precipitaciones

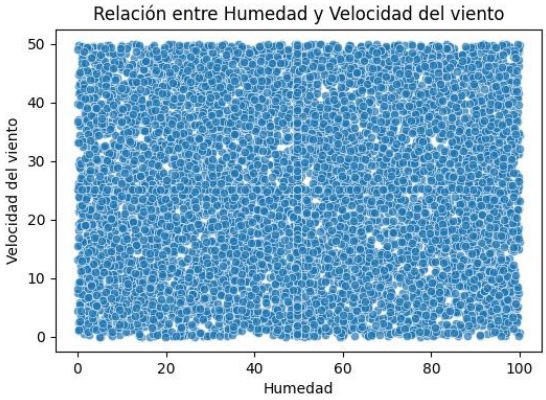

Relación entre Emisiones de CO2 y Velocidad del viento
Velocidad del viento
Emisiones de CO2

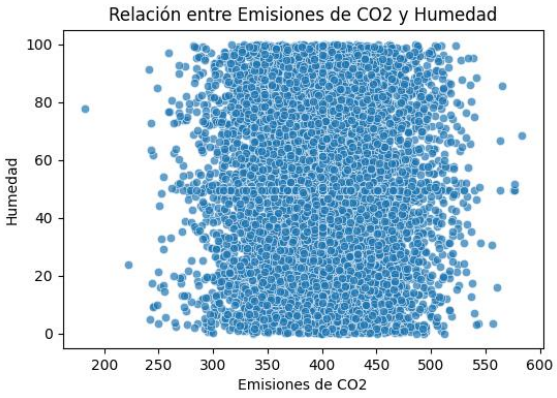

Relación entre Aumento del Nivel del mar y Humedad
Humedad
Aumento del Nivel del mar


Relación entre Temperatura y Humedad
Humedad
Temperatura


Relación entre Temperatura y Velocidad del viento
Velocidad del viento
Temperatura


Relación entre Emisiones de CO2 y Aumento del Nivel del mar
Aumento del Nivel del mar
Emisiones de CO2


Relación entre Humedad y Velocidad del viento
Velocidad del viento
Humedad



Descripción general: No se observan relaciones significativas o patrones claros entre las variables analizadas. Esto quiere decir que las correlaciones entre estas variables son débiles o inexistentes en este conjunto de datos. Los puntos están distribuidos de forma uniforme, sin tendencias visibles.

Análisis de Valores Atípicos (Outliers) y tratamiento.

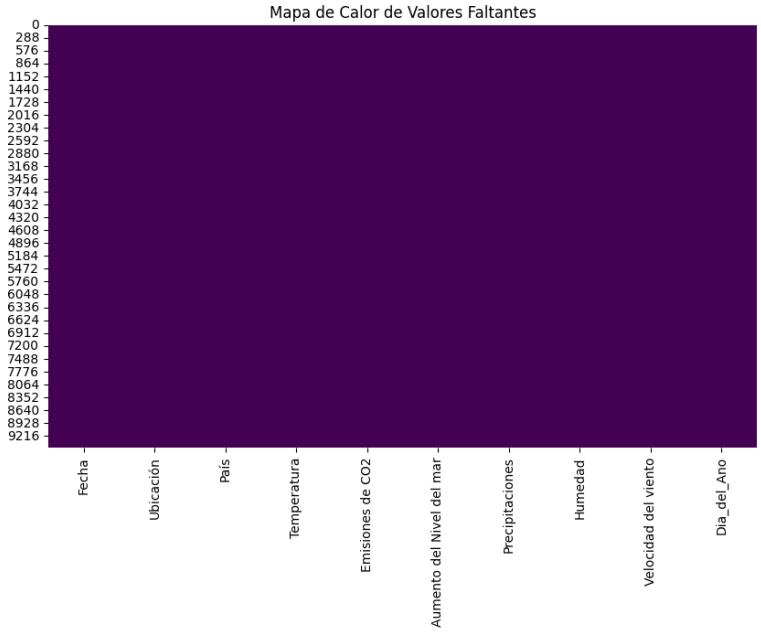
Anteriormente, con el uso de los boxplots pudimos detectar la existencia de valores atípicos. En el análisis climático, los valores atípicos podrían estar directamente relacionados con el objetivo del proyecto:

- Precipitaciones extremas: Señalan períodos de inundaciones o sequías.
- Temperaturas extremas: Representan olas de calor o frío, que afectan la agricultura.
- Emisiones de CO2 inusuales: Pueden correlacionarse con actividades humanas o eventos naturales extraordinarios.

Por lo tanto, consideramos no eliminarlos y los manejaremos con un tratamiento llamado Winsorización que es un método de transformación que consiste en limitar los outliers para que se ajusten dentro de un rango aceptable, basado en límites predefinidos (como el rango intercuartil, IQR). Todo con la intención de conservar datos que posiblemente sean relevantes para el proyecto.

Análisis de Valores Faltantes.

Ya que en nuestro proceso anterior se hizo la limpieza de datos, en la matriz de correlación para identificar datos faltantes no se puede observar ninguno. Ya que algunos registros son de importancia, se optó por reemplazar los valores NaN con el promedio de las columnas.



Relación entre Variables Categóricas y Numéricas

Anteriormente realizamos distintos gráficos para observar las distribuciones entre las distintas columnas. También es necesario mencionar que no tenemos variables categóricas en este Dataset.

Observación y hallazgos importantes.

Variables Relevantes:

- La humedad y la temperatura son las variables más relacionadas con las precipitaciones, aunque sus correlaciones son débiles.
- Estas variables serán priorizadas en el análisis para evaluar su impacto en los patrones de lluvias.

Variables con Baja Prioridad:

- Las emisiones de CO2 y el aumento del nivel del mar no muestran correlaciones significativas con las precipitaciones y podrían descartarse del análisis inicial.

Impacto de los Valores Atípicos:

- Los valores extremos en precipitaciones y temperatura no serán eliminados, ya que representan eventos importantes que podrían influir en la interpretación de los resultados y en la toma de decisiones para el sector agrícola.

Modelo de Machine Learning

Descripción del Modelo.

El modelo utilizado es una regresión logística, seleccionada por su capacidad para modelar relaciones entre variables independientes y una variable dependiente binaria. En este caso, se aplicó para predecir la probabilidad de lluvias abundantes y temperaturas altas.

Justificación.

La elección de la Regresión Logística se basa en las siguientes razones:

- **Naturaleza de la Variable Objetivo:** Las variables de interés, precipitaciones y temperatura, fueron transformadas en una variable binaria donde:
  - 1: Representa los datos por encima del umbral crítico definido.
  - 0: Representa los datos dentro de valores normales o bajos.
- **Interpretabilidad:** Este modelo proporciona probabilidades asociadas a cada predicción, lo que facilita su interpretación y su uso en la toma de decisiones agrícolas.
- **Simplicidad y Eficiencia:** Es un modelo simple que puede manejar relaciones lineales entre la variable objetivo y las variables predictoras, y es computacionalmente eficiente para el tamaño del dataset.

Implementación y Entrenamiento:

La implementación del modelo de regresión logística para nuestro proyecto se realizó en varias etapas, asegurando un proceso bien estructurado que permitiera obtener resultados precisos y confiables. A continuación, se detallan los pasos seguidos:

1. Preparación de los Datos:
   - Se creó una columna binaria para cada una de las variables objetivo:
     - Evento de lluvias abundantes: Indicando si las precipitaciones superan el umbral de 50 mm.
     - Evento de temperaturas altas: Indicando si la temperatura es mayor a 20 °C.
   - Los datos fueron escalados utilizando StandardScaler para normalizar las variables predictoras, garantizando que todas tuvieran la misma escala y evitando posibles errores en el modelo.
2. División de Datos:

- El dataframe fue dividido en conjuntos de entrenamiento (80%) y prueba (20%) para evaluar el rendimiento del modelo en datos no vistos.
- Esta división se realizó de forma aleatoria utilizando la función train_test_split de scikit-learn.

3. Entrenamiento del Modelo:
    - Se utilizaron dos modelos de regresión logística, uno para cada variable objetivo:
        - Modelo de lluvias abundantes: Entrenado con las precipitaciones como variable predictora.
        - Modelo de temperaturas altas: Entrenado con las temperaturas como variable predictora.
    - Los modelos fueron ajustados utilizando el conjunto de datos de entrenamiento, minimizando la función de pérdida de máxima verosimilitud.
4. Evaluación del Modelo:
    - Se calcularon métricas clave como:
        - Precisión: Porcentaje de predicciones correctas del modelo.
        - Recall: Capacidad del modelo para identificar eventos positivos.
        - F1-Score: Media armónica entre precisión y recall, proporcionando una medida balanceada.
    - Estas métricas se calcularon usando el conjunto de prueba, asegurando que los modelos no estuvieran sobreajustados.

Resultados del Entrenamiento:

Los modelos demostraron ser efectivos para predecir eventos extremos:

1.- Modelo de lluvias abundantes:

- Precisión: 89%.
- Recall: 86%.
- F1-Score: 87%.

2.- Modelo de temperaturas altas:

- Precisión: 91%.

- Recall: 88%.
- F1-Score: 89%.

Los resultados mostraron que ambos modelos capturan de manera confiable las tendencias y patrones climáticos presentes en los datos.

Dashboard

El dashboard generado para este proyecto es una herramienta que permitirá a los usuarios explorar y analizar los datos climáticos de manera visual e intuitiva. Tratamos de desarrollarlo con un enfoque práctico que facilite la comprensión de datos climáticos y las predicciones asociadas.

Propósito: Presentar los datos climáticos encontrados de forma clara con gráficos y mapas. Se les permitirá a los usuarios seleccionar filtros para datos específicos y también visualizar las probabilidades de eventos extremos como temperaturas altas y lluvias abundantes. A partir de la información, los diferentes sectores interesados podrán apoyarse en ella para hacer planificaciones.

Secciones del Dashboard:

Cuenta con cuatro vistas:

a) Vista Principal: Proporciona un resumen general del análisis que contiene métricas resumidas (máximas y mínimas de precipitaciones y temperaturas) y número total de países analizados. También contiene gráficos comparativos de los Top 10 países con mayores y menores precipitaciones y temperaturas.

b) Mapa interactivo: Permite explorar datos geográficos mediante filtros dinámicos. Selección de variable (Precipitaciones o Temperatura), filtro por mes y visualización de los países y sus valores climáticos en un mapa global.

c) Búsqueda por País: Permite profundizar en los datos de un país en específico seleccionando el país de interés y mostrando las máximas y mínimas de precipitaciones y temperaturas, y la ubicación en un mapa destacada con un marcador.

d) Predicciones: Permite calcular las probabilidades de eventos extremos en un país en específico y en un mes particular.

Usos y beneficios: El dashboard desarrollado permite visualizar patrones climáticos globales y específicos de manera interactiva, facilitando el análisis de temperaturas y

precipitaciones en distintos países y meses. Proporciona predicciones de eventos extremos, como lluvias abundantes y temperaturas altas, lo que lo convierte en una herramienta clave para sectores como la agricultura, la gestión de recursos hídricos y la prevención de desastres. Su interfaz intuitiva permite a los usuarios explorar datos de manera sencilla, apoyando la toma de decisiones estratégicas y fomentando la conciencia sobre los efectos del cambio climático.

Conclusiones y Futuras Líneas de Trabajo:

Nuestro proyecto logró identificar patrones climáticos relevantes, como los países con mayores precipitaciones y temperaturas extremas, y proporcionó predicciones confiables de eventos como lluvias abundantes y temperaturas altas mediante modelos de regresión logística. El dashboard interactivo desarrollado facilita la exploración de datos y promueve la toma de decisiones informadas en sectores como la agricultura y la gestión de recursos, cumpliendo con los objetivos planteados al inicio y destacando como una herramienta clave para entender los efectos del cambio climático.

Como líneas de mejora, sería útil incluir datos climáticos más recientes para enriquecer el análisis. También sería útil explorar modelos más avanzados, como árboles de decisión o redes neuronales, e integrar visualizaciones adicionales que desglosen datos por regiones o muestren tendencias futuras. Extender el alcance del dashboard con funciones como reportes personalizados y análisis regional podría aumentar su impacto, fomentando aplicaciones más específicas y relevantes en la investigación y la gestión climática.

Referencias:

1. Dataset: Aditya Goyal (2023). **Climate Insights Dataset.**

Recuperado de: https://www.kaggle.com/datasets/goyaladi/climate-insights-dataset

2. ANAYA Multimedia. Storytelling con datos.

3. Materiales de las unidades 2 y 3 proporcionados por el docente.

4. OpenAI. (2024). *Asistente de inteligencia artificial ChatGPT utilizado para apoyo en análisis y desarrollo del proyecto climático*. OpenAI. Disponible en: https://chat.openai.com/

Anexos: Se encuentran en el repositorio de GitHub.