GASES DE EFECTO INVERNADERO Y SU RELACIÓN CON LA TEMPERATURA MUNDIAL

Mundos E - Hackathon

Grupo 2

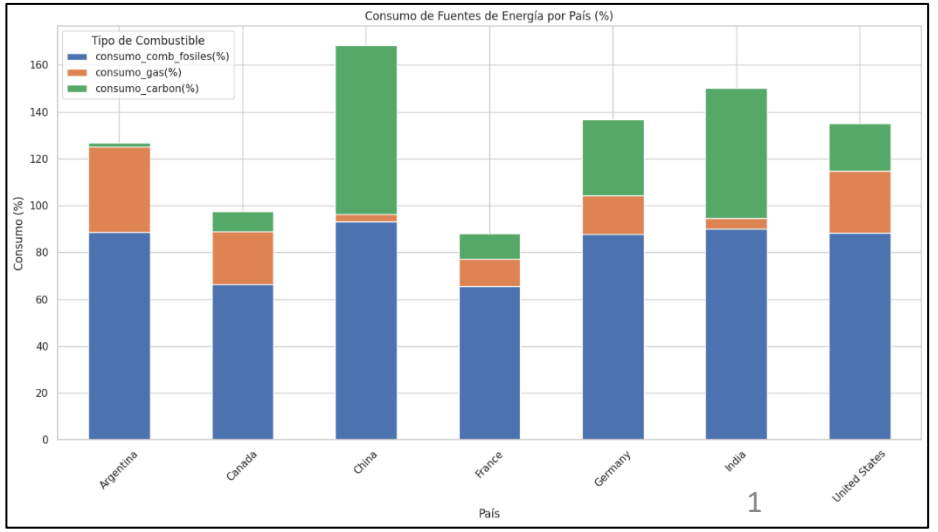
Lautaro Villafañe (ivilla1357@gmail.com) - Ing. Químico

Diego Murature (diegofmurature@gmail.com) - Ing. En Telecomunicaciones

Julio Mansilla (julio.tingo@gmail.com) - Geólogo

Marzo 2025

```python
# GRAFICOS DE LINEAS PARA VER LA TENDENCIA DE LAS EMISIONES A LO LARGO DE LOS AÑOS

# Selección de paises para trabajar con los gráficos:
paises_seleccionados = ['Argentina', 'China', 'India', 'United States', 'Germany', 'Canada']
df_seleccionado = union_df_filtrado[union_df_filtrado['Entity'].isin(paises_seleccionados)]

# Tamaño de gráfico:
plt.figure(figsize=(15, 10))

# Subplots (2x2)
fig, axs = plt.subplots(2, 2, figsize=(15, 10))

# Variables para graficar:
variables = ['emision_co2(Tn)', 'emision_n2o(Tn)', 'emision__ch4(Tn)', 'emision__so2(Tn)']
titulos = ['Emisiones de CO2', 'Emisiones de N2O', 'Emisiones de CH4', 'Emisiones de SO2']

# Graficar cada variable en su subplot correspondiente
for i, var in enumerate(variables):
    ax = axs[i//2, i%2]
    for pais in paises_seleccionados:
        subset = df_seleccionado[df_seleccionado['Entity'] == pais]
        ax.plot(subset['Year'], subset[var], label=pais)
    ax.set_title(titulos[i])
    ax.set_xlabel('Año')
    ax.set_xlim(1945, 2025)
    ax.set_ylabel(f'Emisiones Totales ({var.split("_")[1].upper()})')
    ax.legend()

```



AGENDA

- 1 - ¿QUÉ QUEREMOS HACER?
- 2 - OBJETIVOS DEL TRABAJO
- 3 - DESCRIPCIÓN DE LA PROBLEMÁTICA (¿POR QUÉ ES IMPORTANTE ESTE PROBLEMA?)
- 4 - METODOLOGÍA DE ANÁLISIS (¿COMO SE DERROLLO EL TRABAJO?)
- 5 - RESULTADOS Y CONCLUSIONES

1 - ¿QUÉ QUEREMOS HACER?

El propósito principal de este trabajo es utilizar Python, a través de Google Colab, para poder analizar las variables que influyen en el cambio climático y en las variaciones de temperatura de superficie medias mundiales.

Además, realizar un modelo que prediga las temperaturas a partir de las emisiones y concentraciones de gases de efecto invernadero emitidas por la población e industrias.

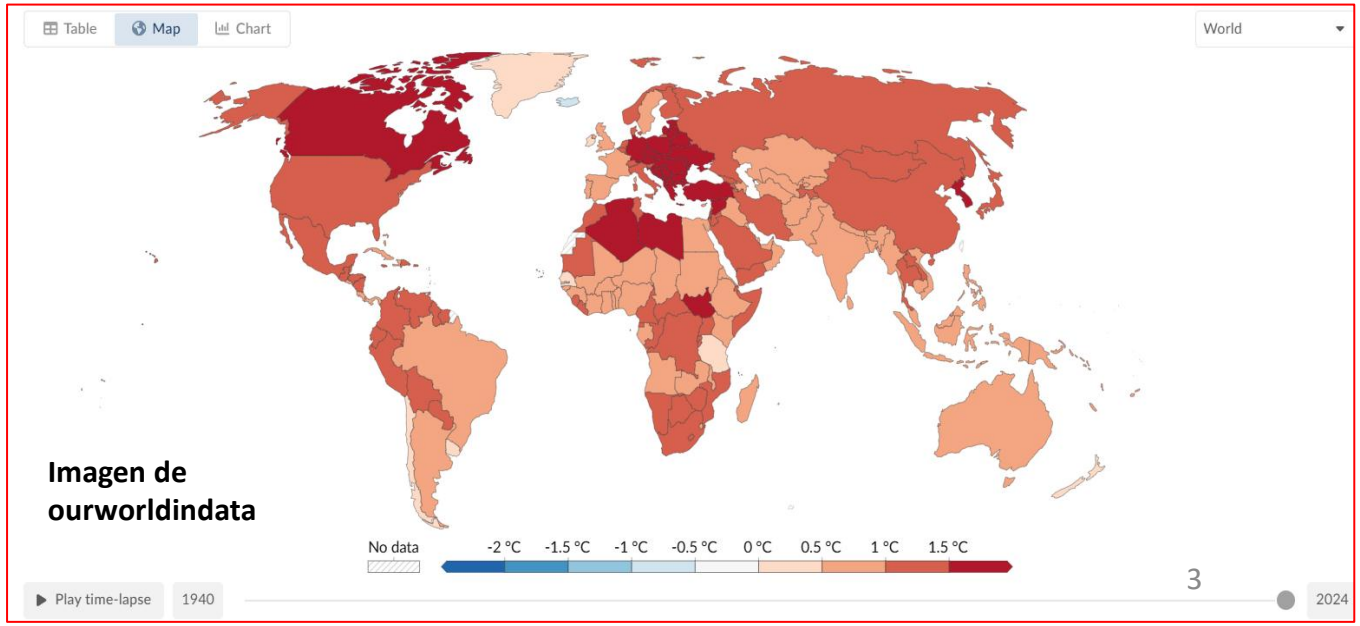


Imagen de ourworldindata

2 - OBJETIVOS DEL TRABAJO

- Analizar las tendencias de las emisiones de gases a lo largo de los años por países.
- Determinar cuáles son los países que mas colaboran con las emisiones de gases.
- Cuales son las concentraciones de gases que se encuentran en la atmosfera terrestre.
- De los países que más emisiones de gases liberan, cuales son sus fuentes de energía y de que tipo son (renovables o no renovables).
- Como es el consumo de combustibles fósiles en el mundo a lo largo de los años.
- Cuales variables son las que mas contribuyen a las anomalías de temperaturas mundiales.
- Realizar un Análisis de Regresión Lineal Múltiple, utilizando variables como las concentraciones de gases de efecto invernadero en la atmosfera para predecir las anomalías de T (°C) en superficie.
- Utilizar un método de aprendizaje automático de modelos supervisados como Random Forest para predecir las anomalías de temperatura global.

3 - DESCRIPCIÓN DE LA PROBLEMÁTICA (¿POR QUÉ ES IMPORTANTE ESTE PROBLEMA?)

En la última década, las emisiones de dióxido de carbono (CO_2), metano (CH4), y óxidos de nitrógeno (NO) han alcanzado altos niveles, resultado en un aumento significativo en la temperatura media global. Este incremento no solo afecta el clima, sino que también presenta riesgos para la salud pública, la biodiversidad y la seguridad alimentaria.

A medida que los países industrializados y en desarrollo contribuyen a estas emisiones, se observa una falta de consenso y acción global efectiva para mitigar el impacto del cambio climático. Por ello, es fundamental comprender las variables que inciden en este fenómeno para poder mitigar sus efectos y proponer soluciones sostenibles.

Impactos claves de esta problemática

1. Desigualdad en las Emisiones: Solo algunos países son responsables de una gran parte de las emisiones globales.
2. Fuentes de Energía No Renovables: Muchos de los países que más contribuyen a las emisiones dependen en gran medida de fuentes de energía no renovables.
3. Cambios en los Patrones Climáticos: Las anomalías de temperatura están relacionadas con cambios peligrosos en los patrones climáticos, lo que puede llevar a problemas en la agricultura, acceso al agua y sanidad pública.

4 - METODOLOGÍA DE ANÁLISIS (¿COMO SE DERROLLO EL TRABAJO?)

1 Recolección de datos

```
[ ]  1  emisiones_gases.head()
```

	Entity	Year	emision_co2(Tn)	emision_n2o(Tn)	emision__ch4(Tn)	emision__so2(Tn)
0	Afghanistan	1950	84272.0	2238381.0	8100568.5	748.6191
1	Afghanistan	1951	91600.0	2299787.2	8233321.0	767.4483
2	Afghanistan	1952	91600.0	2373210.5	8372900.5	778.1855
3	Afghanistan	1953	106256.0	2454824.2	8528310.0	816.0000
4	Afghanistan	1954	106256.0	2540803.5	8690891.0	838.3837

```
[ ]  1  emisiones_gases.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19218 entries, 0 to 19217
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Entity            19218 non-null  object
 1   Year              19218 non-null  int64
 2   emision_co2(Tn)   17300 non-null  float64
 3   emision_n2o(Tn)   16280 non-null  float64
 4   emision__ch4(Tn)  15910 non-null  float64
 5   emision__so2(Tn)  16790 non-null  float64
dtypes: float64(4), int64(1), object(1)
memory usage: 901.0+ KB
```

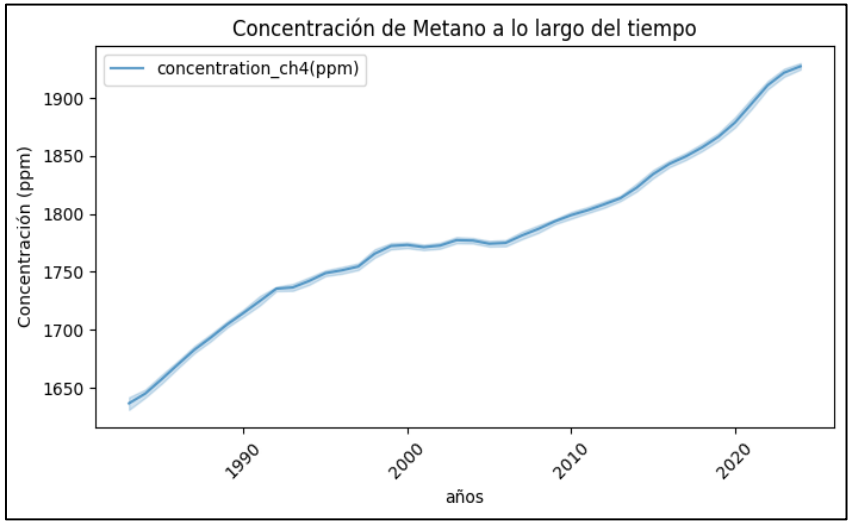
2 Limpieza y preprocesamiento

```
1  # Cambiar el nombres columnas para temperaturas_anuales:
2  temperaturas_anuales.rename(columns={'temperature_anomaly': 'T(°C)'}, inplace=True)
3
4  # Borrar columnas:
5  temperaturas_anuales = temperaturas_anuales.drop(columns=['Code'])
```

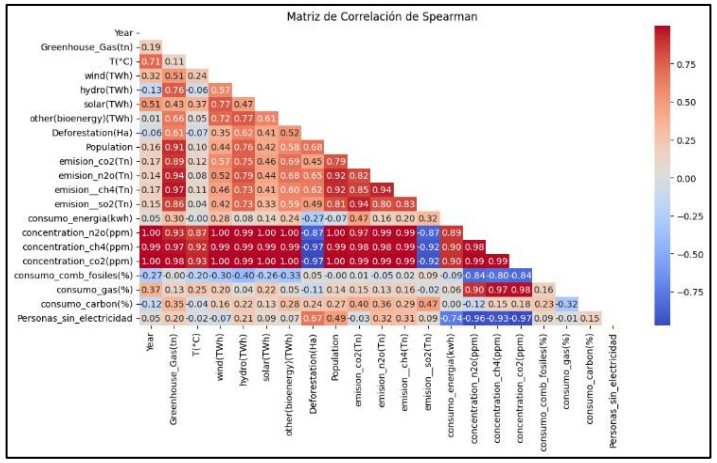
```
[ ]  1  temperaturas_anuales.tail()
```

	Entity	Year	T(°C)
16570	Zimbabwe	2020	0.229293
16571	Zimbabwe	2021	0.016838
16572	Zimbabwe	2022	0.035659
16573	Zimbabwe	2023	0.792851
16574	Zimbabwe	2024	1.364956

3 Análisis exploratorios



4 Selección de variables



5 Estandarización de los datos

```
[9]  1  # Estanderización de las variables #
2
3  from sklearn.preprocessing import StandardScaler
4
5  scaler = StandardScaler()
6
7  # Selección de variables que quieres estandarizar
8  columns_to_standardize = ['Greenhouse_Gas(tn)', 'T(°C)', 'wind(TWh)', 'hydro(TWh)', 'solar(TWh)',
9                            'other(bioenergy)(TWh)', 'Deforestation(Ha)', 'Population', 'emision_co2(Tn)',
10                           'emision_n2o(Tn)', 'emision__ch4(Tn)', 'emision__so2(Tn)',
11                           'consumo_energia(kwh)', 'concentration_n2o(ppm)', 'concentration_ch4(ppm)',
12                           'concentration_co2(ppm)', 'consumo_comb_fosiles(%)', 'consumo_gas(%)',
13                           'consumo_carbon(%)', 'Personas_sin_electricidad']
14
15 # Aplicar la estandarización
16 df[columns_to_standardize] = scaler.fit_transform(df[columns_to_standardize])
17
18 df.tail()
19
```

	Entity	Year	Greenhouse_Gas(tn)	T(°C)	wind(TWh)	hydro(TWh)	solar(TWh)	other(bioenergy)(TWh)	Deforestation
26438	Zimbabwe	2020	-0.247281	0.928929	-0.198068	-0.351949	-0.154867	-0.296902	
26439	Zimbabwe	2021	-0.246990	0.642477	-0.198068	-0.349074	-0.154867	-0.297938	
26440	Zimbabwe	2022	-0.246960	0.667852	-0.198068	-0.349141	-0.154779	-0.296902	

0s se ejecutó 7:21 p.m.

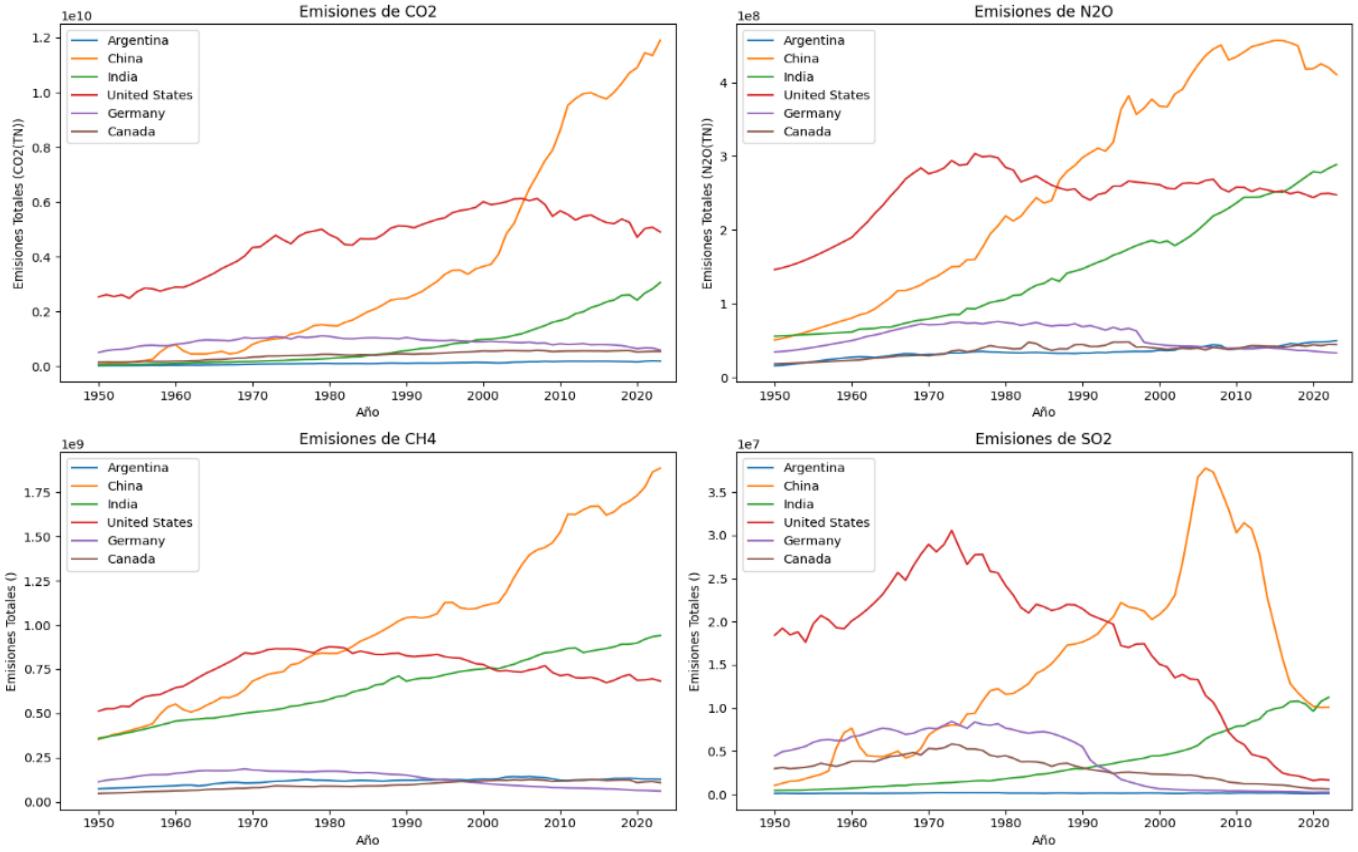
6 Creación de los modelos (RLM y RF)

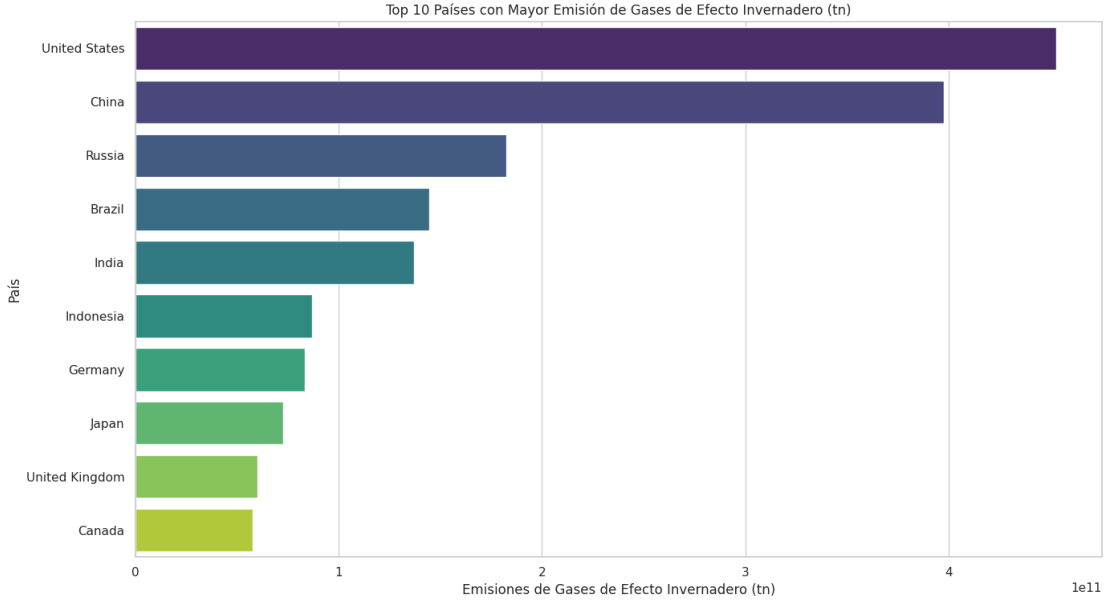
```
6  # Selección de variables
7  X = df[[ 'concentration_co2(ppm)','concentration_ch4(ppm)','concentration_n2o(ppm)']]
8  y = df['T(°C)']
9
10 # Eliminar filas con valores faltantes en X e y simultáneamente
11 df_cleaned = df[['concentration_co2(ppm)', 'concentration_ch4(ppm)', 'concentration_n2o(ppm)','T(°C)']].dropna()
12 X = df_cleaned[[ 'concentration_co2(ppm)', 'concentration_ch4(ppm)','concentration_n2o(ppm)']]
13 y = df_cleaned['T(°C)']  # Variable dependiente/a predecir o target.
14
15 # Dividir datos en entrenamiento y prueba
16 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42) # 70% de los datos para entren
17                                                                                          # #random_state=4 Si fijamos e
18
19 #Nos va a tirar datos aleatorios cada vez que lo llamemos, segun el porcentaje que coloquemos en "test_size=". Ejemplo:
20 #Cada vez que lo haga me tomara valores distintos aleatorios, esto se llama reproducibilidad de un experimiento.
21
22
23 # Crear y entrenar el modelo
24 model = LinearRegression()
25 model.fit(X_train, y_train)
26
27 # Predecir y evaluar el modelo
28 y_pred = model.predict(X_test)
29
30 # Calcular métricas de evaluación
31 r2 = r2_score(y_test, y_pred)
32 mse = mean_squared_error(y_test, y_pred)
33 mae = mean_absolute_error(y_test, y_pred)
34 rmse = np.sqrt(mse)
35
36 print("Train R2=", r2)
37 print(f'Mean Squared Error (MSE): {mse}')
38 print(f'Mean Absolute Error (MAE): {mae}')
39 print(f'Root Mean Squared Error (RMSE): {rmse}')
```

```
Train R2= 0.8305426757733156
Mean Squared Error (MSE): 0.014359585206407911
Mean Absolute Error (MAE): 0.09650904315586578
Root Mean Squared Error (RMSE): 0.11983148670699163
```

5 - RESULTADOS Y CONCLUSIONES

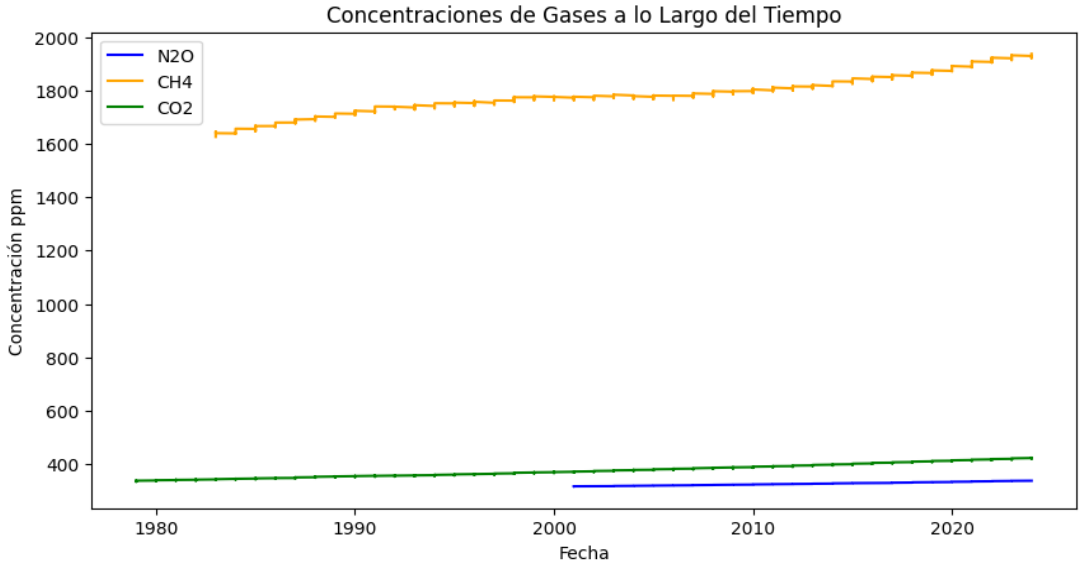
- Se identificaron los países que más contribuyen a las emisiones de gases de efecto invernadero, siendo China, Estados Unidos, India y Rusia los principales emisores.

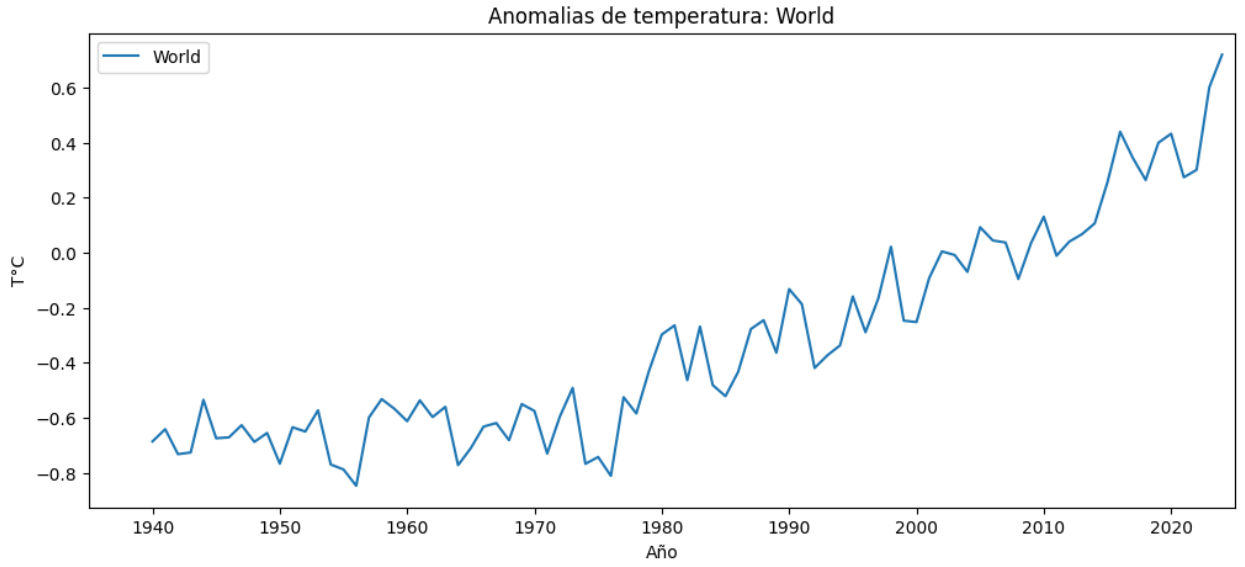


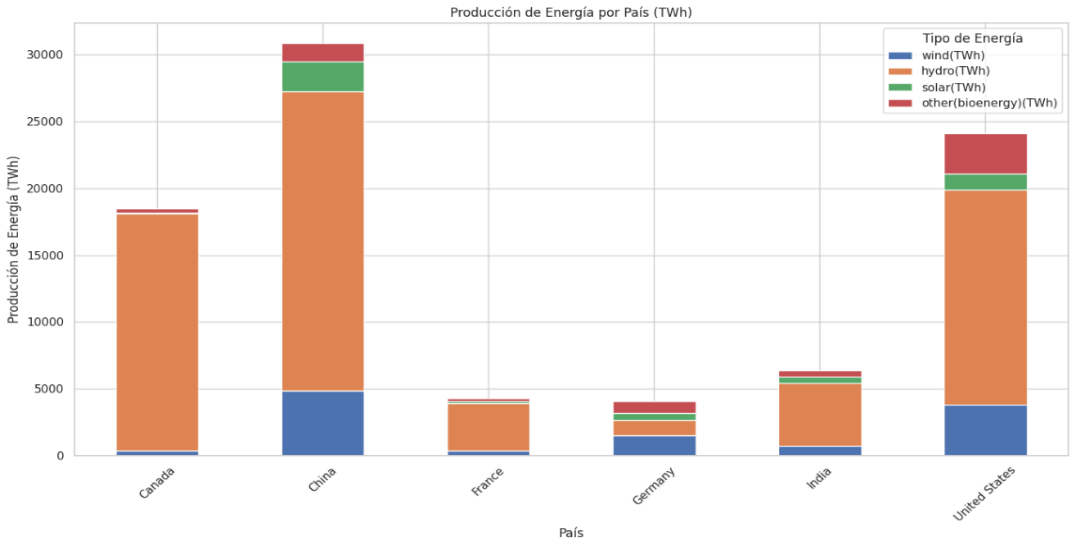


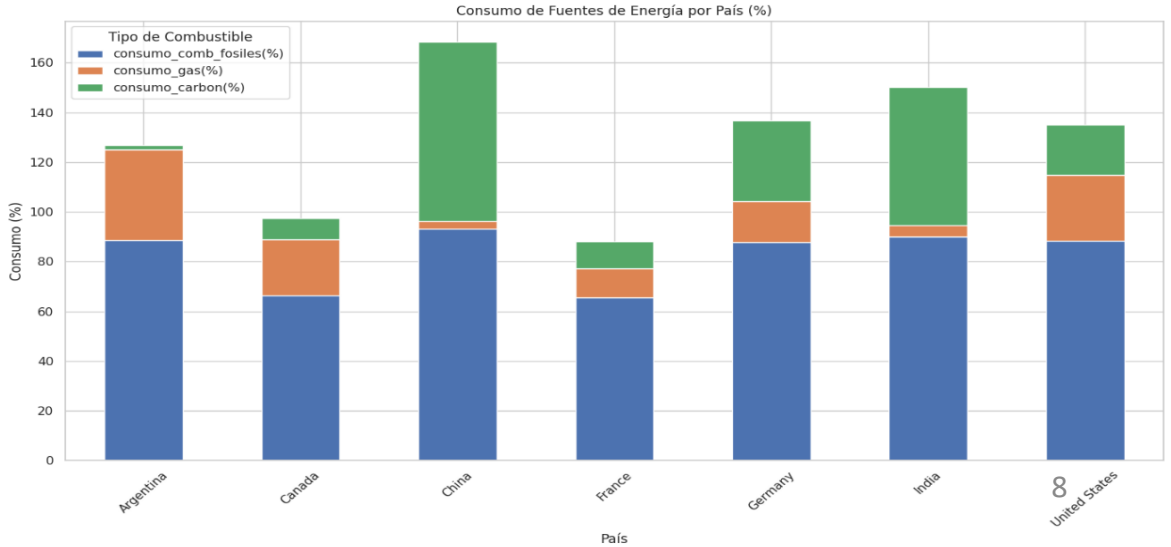
5 - RESULTADOS Y CONCLUSIONES

- ✓ Identificación las concentraciones de gases que se encuentran en la atmosfera terrestre. Cuales son sus fuentes de energía y de que tipo son (renovables o no renovables) para los países. Verificar las anomalías de temperatura mundiales.









5 - RESULTADOS Y CONCLUSIONES

✓ Análisis de Regresión Lineal Múltiple, utilizando variables como las concentraciones de gases de efecto invernadero en la atmosfera para predecir las anomalías de T (°C) en superficie.

```
# Dividir datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42) # 70% de los d
                                                                                          # #random_stat

#Nos va a tirar datos aleatorios cada vez que lo llamemos, segun el porcentaje que coloquemos en "test_s
#Cada vez que lo haga me tomara valores distintos aleatorios, esto se llama reproducibilidad de un exper


# Crear y entrenar el modelo
model = LinearRegression()
model.fit(X_train, y_train)

# Predecir y evaluar el modelo
y_pred = model.predict(X_test)

# Calcular métricas de evaluación
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mse)

print("Train R2=", r2)
print(f'Mean Squared Error (MSE): {mse}')
print(f'Mean Absolute Error (MAE): {mae}')
print(f'Root Mean Squared Error (RMSE): {rmse}')
```

```
Train R2= 0.8305426757733156
Mean Squared Error (MSE): 0.014359585206407911
Mean Absolute Error (MAE): 0.09650904315586578
Root Mean Squared Error (RMSE): 0.11983148670699163
```

El modelo de regresión lineal múltiple obtuvo un R^2 del 0.83054, mostrando una alta capacidad explicativa de las anomalías de temperatura, es decir, el 83.05% de la variabilidad de la temperatura puede ser explicada por las concentraciones de estos gases.

Además, el RMSE da 0.119831 lo sugiere que, en promedio, las predicciones del modelo tienen un error de aproximadamente 0.119 grados Celsius.

Interpretación P-VALUE:

-Concentration_co2(ppm): Un valor p de 0.007 indica que esta variable independiente es significativa al nivel de 0.05 (5%).

-Concentration_ch4(ppm): Un valor p de 0.000 también indica que esta variable es altamente significativa.

-Concentration_n2o(ppm): Un valor p de 0.092 indica que esta variable no es significativa al nivel de 0.05, pero podría considerarse marginalmente significativa si se usa un nivel de significación más alto (por ejemplo, 0.10).

OLS Regression Results

Dep. Variable:	T(°C)	R-squared:	0.799
Model:	OLS	Adj. R-squared:	0.796
Method:	Least Squares	F-statistic:	371.4
Date:	Tue, 11 Mar 2025	Prob (F-statistic):	2.02e-97
Time:	16:57:55	Log-Likelihood:	181.83
No. Observations:	285	AIC:	-355.7
Df Residuals:	281	BIC:	-341.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4008	0.110	3.631	0.000	0.184	0.618
concentration_co2(ppm)	0.3234	0.119	2.723	0.007	0.090	0.557
concentration_ch4(ppm)	0.2939	0.053	5.569	0.000	0.190	0.398
concentration_n2o(ppm)	-0.1512	0.089	-1.691	0.092	-0.327	0.025

Omnibus:	3.568	Durbin-Watson:	0.154
Prob(Omnibus):	0.168	Jarque-Bera (JB):	2.652
Skew:	-0.083	Prob(JB):	0.265
Kurtosis:	2.557	Cond. No.	41.2

5 - RESULTADOS Y CONCLUSIONES

✓ Análisis de Regresión Lineal Múltiple, utilizando el método de validación cruzada para evaluar el rendimiento de un modelo de aprendizaje automático de manera más robusta

```python
### VALIDACION CRUZADA ####
 # Lo realizo sobre el conjunto de datos de entrenamiento.

from sklearn.model_selection import cross_val_score

# Calcular la validación cruzada
scores = cross_val_score(model, X_train, y_train, cv=10, scoring='neg_mean_squared_error')
rmse_scores = np.sqrt(-scores)

print(f'Scores: {rmse_scores}')
print(f'Mean RMSE: {rmse_scores.mean()}')
print(f'Standard Deviation: {rmse_scores.std()}')

```

```
Scores: [0.15335553 0.11372592 0.11586044 0.12890906 0.12974855 0.10940188
 0.12873071 0.1333116  0.16196708 0.15869284]
Mean RMSE: 0.13337036103620367
Standard Deviation: 0.01783240521245515
```

Lo que se realiza es dividir los datos en varios subconjuntos que se van entrenando y evaluando el modelo en cada uno de ellos.

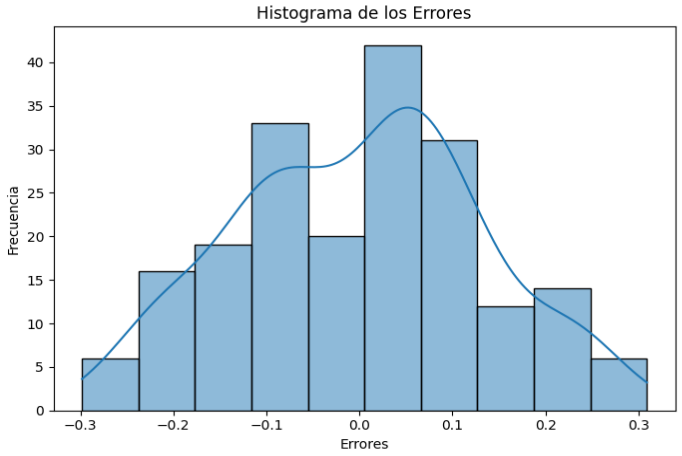
Los valores de Scores representan los errores cuadráticos medios (MSE) negativos, que luego se transforman a raíz cuadrada para obtener la RMSE (Root Mean Squared Error) en cada una de las 10 particiones de la validación cruzada.

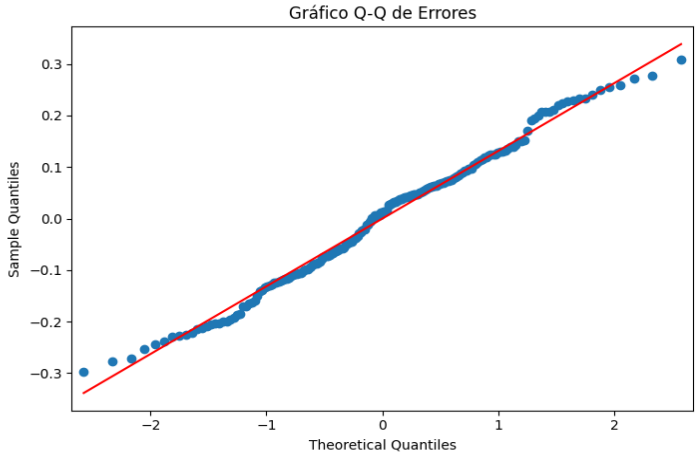
Desviación estándar de 0.01783 indica la variabilidad de los RMSE a través de las diferentes particiones. Un valor más bajo sugiere que el rendimiento del modelo es consistente en las diferentes particiones.

El RMSE medio de la validación cruzada (0.13337) es ligeramente superior al RMSE de la validación simple (0.11983).
Esto podría indicar que la validación cruzada está capturando mejor la variabilidad y es menos optimista sobre el rendimiento del modelo.

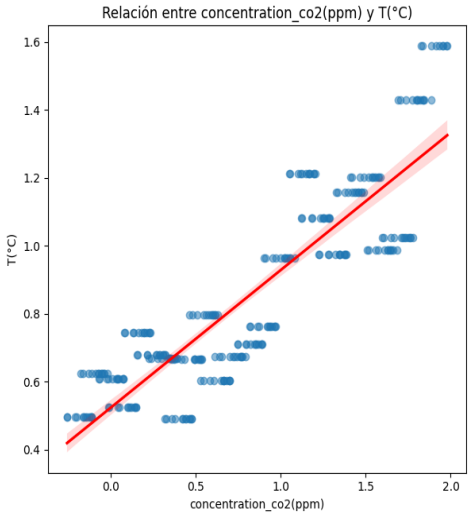
5 - RESULTADOS Y CONCLUSIONES

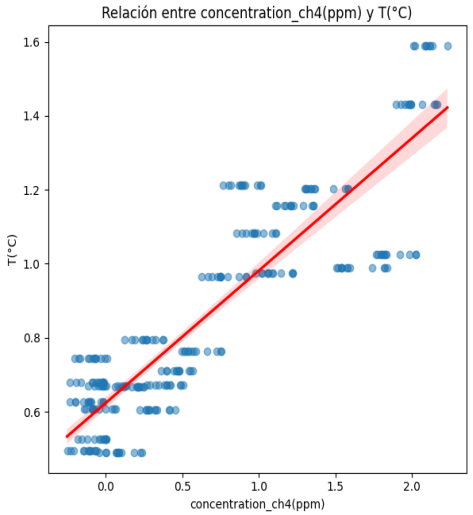
Análisis de supuestos para la regresión lineal múltiple

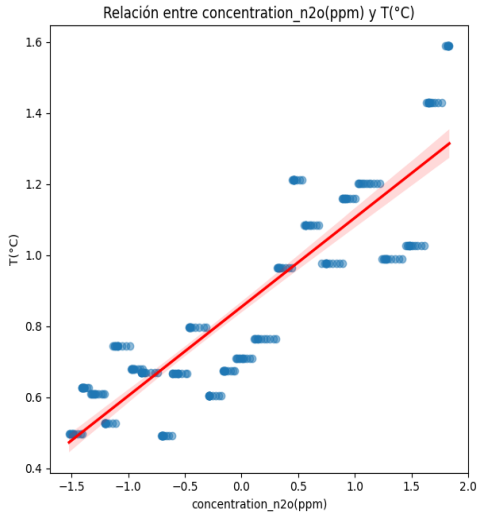




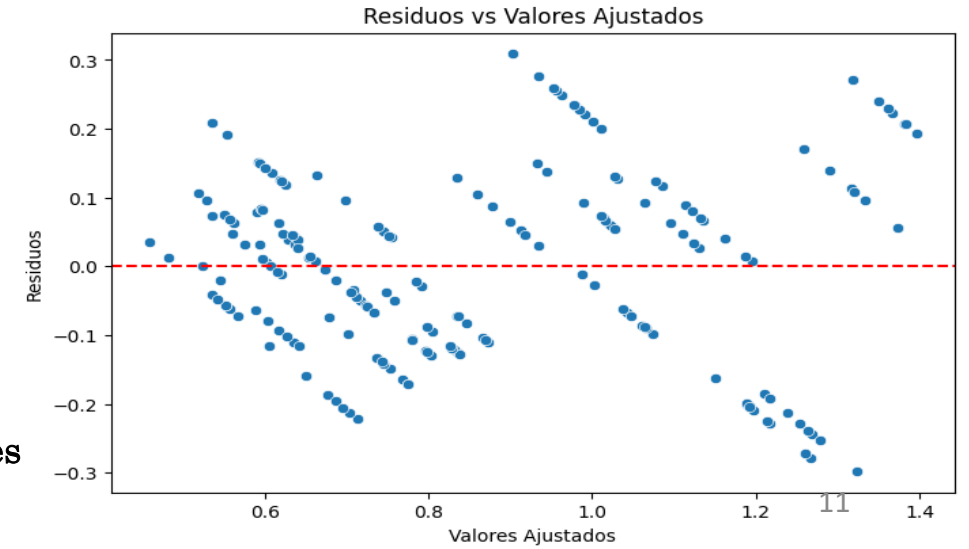
Supuesto de Normalidad







Supuesto de Linealidad

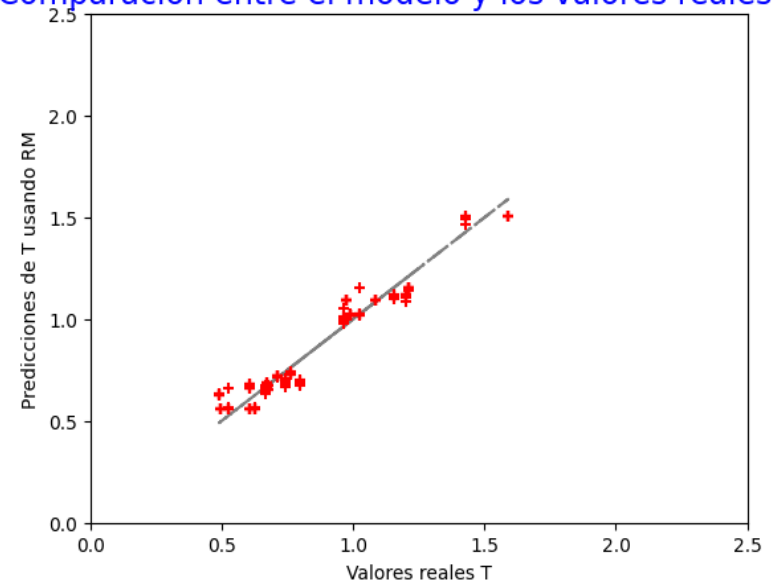


Independencia de los Errores (Homocedasticidad)

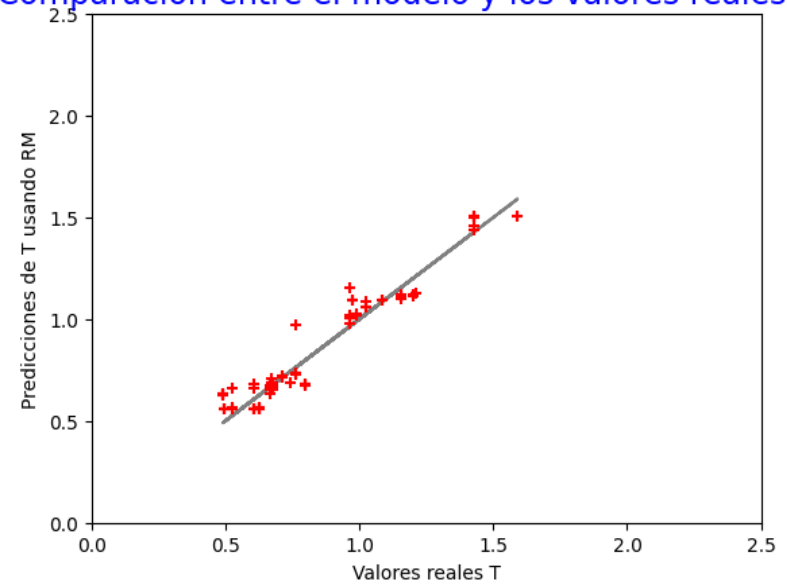
5 - RESULTADOS Y CONCLUSIONES

✓ Utilizar un método de aprendizaje automático de modelos supervisados como Random Forest para predecir las anomalías de temperatura global.

ENTRENAMIENTO - Modelo: RandomForestRegressor
Comparación entre el modelo y los valores reales de T



TESTEO - Modelo: RandomForestRegressor
Comparación entre el modelo y los valores reales de T




```
MESTRICAS EN ENTRENAMIENTO:

Error cuadratico medio:                0.004
Error absoluto medio:                  0.051
Raiz del error cuadratico medio:       0.225
R cuadrado:                            0.95073

 MESTRICAS EN TESTEO:

Error cuadratico medio:                0.005
Error absoluto medio:                  0.059
Raiz del error cuadratico medio:       0.242
R cuadrado:                            0.9354
```

Los resultados parecen indicar que el modelo es consistente entre los conjuntos de entrenamiento y testeo. No hay una diferencia significativa entre las métricas, lo que indica que el modelo tiene un buen equilibrio y no está sobreajustado.

Tanto el MSE como el MAE y la RMSE son bajos, lo que implica que el modelo hace predicciones precisas. El alto valor de R^2 dado indica que el modelo se ajusta bien a los datos.

El modelo captura con más profundidad la vinculación entre la temperatura y los gases de interés.
Se debe destacar el rendimiento del modelo a pesar del escueto conjunto de valores proporcionados a sus hiperparámetros. Esto nos permite concluir que la relación entre el target y los predictores es sumamente relevante.

Preguntas