Actividad Practica

Stephanie Ríos Rodríguez

Fundación Universitaria Compensar

Programación para Ciencia de Datos II

Docente

Sebastián Rodríguez Muñoz

Bogotá D.C, Colombia

17 de marzo de 2024

Actividad Practica

Stephanie Ríos Rodríguez

Docente

Sebastián Rodríguez Muñoz

Actividad # 6

Fundación Universitaria Compensar

Facultad De Ingeniería-Ciencia De Datos

Programación para Ciencia de Datos II

Bogotá D.C, Colombia

17 de marzo de 2024

Contenido

Introducción.

El presente proyecto surge a partir del análisis de datos referentes a los colombianos detenidos en el exterior, con el objetivo de explorar si existen diferencias en la cantidad de detenidos entre países, en particular entre Chile y Estados Unidos. A lo largo del proceso, se aplicaron métodos estadísticos básicos para contrastar hipótesis y evaluar las diferencias observadas en los registros.

Para facilitar la comprensión y la exploración de estos resultados, se desarrolló un dashboard interactivo utilizando Python y la librería Dash. Este dashboard permite filtrar la información por país y visualizar, de forma intuitiva, los hallazgos clave obtenidos del análisis. De esta manera, el proyecto no solo demuestra la capacidad de transformar datos complejos en visualizaciones claras, sino que también brinda una herramienta práctica para apoyar la toma de decisiones basada en evidencia.

Desarrollo

Primera parte.

1. Considerando su conjunto de datos y la cuestión de investigación definidas en el taller anterior. Plantee la hipótesis nula y alternativa. La hipótesis nula (H0) debe representar la afirmación inicial que presupone la ausencia de un efecto o relación significativa en la población. Por otro lado, la hipótesis alternativa (H1) debe desafiar esta suposición al afirmar la existencia de un efecto o relación. Es crucial que ambas hipótesis sean verificables a través de métodos estadísticos.

Pregunta de Investigación:
¿Existe una diferencia significativa en la cantidad de detenidos entre los registros correspondientes a Chile y Estados Unidos?

Hipótesis Nula (H0):
La media de la cantidad de detenidos en Chile es igual a la media de la cantidad de detenidos en Estados Unidos.
$(H0: \mu_{(Chile)} = \mu_{(EE.UU.)})$
Esta afirmación parte de la suposición inicial de que el país de detención no influye en la cantidad de detenidos.

Hipótesis Alternativa (H1):
La media de la cantidad de detenidos en Chile difiere significativamente de la media de la cantidad de detenidos en Estados Unidos.
$(H1: \mu_{(Chile)} \neq \mu_{(EE.UU.)})$
Esta hipótesis desafía la afirmación inicial al proponer que el país de detención sí tiene un efecto sobre la cantidad de detenidos.

2. Asegúrese de que las hipótesis nula y alternativa sean mutuamente excluyentes, lo que significa que no deben superponerse en términos de afirmaciones. Además, deben ser colectivamente exhaustivas, lo que implica que juntas deben cubrir todas las posibles situaciones en el contexto de su investigación. Esta condición es esencial para garantizar que el contraste de hipótesis sea completo y adecuado.

Para garantizar que el contraste de hipótesis sea completo y adecuado, es fundamental que las hipótesis formuladas cumplan con dos condiciones:

Mutuamente Exclusivas:
Esto significa que las hipótesis no pueden ser verdaderas al mismo tiempo. En nuestro caso, tenemos:

- **H0:** $\mu_{(Chile)} = \mu_{(EE.UU.)}$
- **H1:** $\mu_{(Chile)} \neq \mu_{(EE.UU.)}$
  Si la hipótesis nula (H0) se cumple (las medias son iguales), la hipótesis alternativa (H1) necesariamente es falsa, y viceversa. Por lo tanto, no existe solapamiento entre las dos afirmaciones.

Colectivamente Exhaustivas:
Esto implica que, entre ambas, se cubren todas las posibilidades respecto a la comparación de las medias. Es decir, para cualquier situación en la que se compare la cantidad de detenidos entre Chile y Estados Unidos, se cumple que:

- O bien las medias son iguales (H0),
- O bien son diferentes (H1).
  No existe una tercera posibilidad fuera de estas dos, lo que garantiza que las hipótesis abarquen todas las situaciones posibles en el contexto de la investigación.

En resumen, las hipótesis H0 y H1 son mutuamente excluyentes porque no pueden ser verdaderas simultáneamente, y son colectivamente exhaustivas ya que entre ellas se cubren todos los escenarios posibles para la diferencia (o ausencia de diferencia) en la cantidad de detenidos

entre Chile y Estados Unidos. Esta formulación asegura que el contraste de hipótesis sea completo y permite tomar decisiones estadísticas claras.

3. Seleccione una prueba estadística teniendo en cuenta los datos. La elección de la prueba estadística es crucial y debe basarse en la naturaleza de sus datos y las hipótesis formuladas. Por ejemplo, si está comparando medias de dos grupos, una prueba t podría ser apropiada. Si está analizando la asociación entre variables categóricas, una prueba chi-cuadrado podría ser la mejor opción. Justifique su elección explicando por qué la prueba seleccionada es la más adecuada para abordar su pregunta de investigación. Por ejemplo, si está utilizando una prueba t, explique que esta prueba es adecuada para comparar medias entre dos grupos independientes.

Prueba Seleccionada: Prueba t de Student para muestras independientes

Justificación de la Elección:

- **Comparación de Medias:**
  Dado que el objetivo es comparar la cantidad de detenidos entre dos grupos independientes (Chile y Estados Unidos), la prueba t es la más adecuada. Esta prueba permite determinar si existe una diferencia significativa entre las medias de dos poblaciones.

- **Independencia de las Muestras:**
  Los datos de detenidos correspondientes a Chile y a Estados Unidos provienen de registros independientes entre sí, lo que cumple con uno de los supuestos fundamentales de la prueba t para muestras independientes.

- **Tamaño de Muestra Suficiente y Distribución:**
  Aunque se asume que las poblaciones tienen una distribución aproximadamente normal, el uso de la prueba t es robusto, especialmente cuando el tamaño de la muestra es suficientemente grande. En caso de sospechar que las varianzas son diferentes, se puede optar por la versión de Welch de la prueba t (especificando equal_var=False en

scipy.stats.ttest_ind).

- **Configuración Bilateral:**
  Como no tengo una hipótesis direccional (no se plantea cuál país debería tener una mayor cantidad de detenidos), la prueba se configura de forma bilateral, evaluando diferencias en ambas direcciones

La prueba t de Student para muestras independientes es la opción más adecuada para este contexto, ya que permite evaluar de manera estadística si las diferencias observadas en la cantidad de detenidos entre Chile y Estados Unidos son significativas o si pudieran atribuirse al azar.

4. Utilice la biblioteca scipy.stats para llevar a cabo el contraste de hipótesis. Proporcione código que ejecute la prueba y justifique los parámetros utilizados. Por ejemplo, si está realizando una prueba t de dos muestras, incluya el código que cargue los datos, calcule la estadística de prueba y el p-valor, y explique la interpretación de estos resultados. Asegúrese de detallar cualquier configuración específica de la prueba, como la elección de prueba bilateral o unilateral.

```python
# Bibliotecas necesarias
import pandas as pd
from scipy import stats

# Cargar la base de datos
ruta = r'C:\Users\steph\colombianos detenidos exterior.csv'
df = pd.read_csv(ruta, sep=';', encoding='utf-8', on_bad_lines='skip', dtype={'UBICACIÓN PAÍS': str})

# Filtrar los datos para los países de interés: CHILE y ESTADOS UNIDOS
paises_interes = ['CHILE', 'ESTADOS UNIDOS']
df_pais = df[df['PAIS PRISIÓN'].isin(paises_interes)]

# Extraer la variable "CANTIDAD" para cada grupo
detenidos_chile = df_pais[df_pais['PAIS PRISIÓN'] == 'CHILE']['CANTIDAD']
detenidos_estados = df_pais[df_pais['PAIS PRISIÓN'] == 'ESTADOS UNIDOS']['CANTIDAD']

# Realizar la prueba t de Student para muestras independientes
t_stat, p_valor = stats.ttest_ind(detenidos_chile, detenidos_estados, equal_var=False)

# Imprimir los resultados
print("Estadístico t:", t_stat)
print("p-valor:", p_valor)
```

```
Estadístico t: 6.848319505731843
p-valor: 7.753874078364868e-12
```

Realización de la prueba t:

La función stats.ttest_ind() compara las medias de dos muestras independientes.

- Se utiliza equal_var=False para no asumir que las dos poblaciones tienen la misma varianza, lo cual es la versión de Welch de la prueba t.
- Por defecto, la prueba es bilateral, lo que es adecuado ya que no se ha planteado una dirección específica en las hipótesis (sólo se evalúa si las medias son iguales o diferentes).

Interpretación de los resultados:

Al ejecutar el código, se obtuvo:

- **Estadístico t:** ~6.8483
  Esto indica que la diferencia observada entre las medias de los dos grupos es 6.85 veces la variabilidad esperada.
- **p-valor:** ~7.75e-12
  Un p-valor tan pequeño indica que, bajo la hipótesis nula de que las medias son iguales, la probabilidad de obtener una diferencia tan grande es prácticamente nula.

Dado que el p-valor es menor que el nivel de significancia comúnmente utilizado (alfa = 0.05), se rechaza la hipótesis nula. Esto significa que existe evidencia estadística muy fuerte para afirmar que la media de la cantidad de detenidos difiere significativamente entre CHILE y ESTADOS UNIDOS.

5. Después de realizar la prueba, interprete los resultados en el contexto de su investigación. Evalúe si puede rechazar la hipótesis nula o no, en función del nivel de significancia (alfa) que haya elegido (por ejemplo, alfa = 0.05). Reporte los valores obtenidos, como la estadística de prueba y el p-valor. Explique qué significan estos valores y cómo impactan en la evaluación de sus hipótesis. Por ejemplo, si el p-valor es menor que alfa, indique que tiene evidencia suficiente para rechazar la hipótesis nula.

Después de ejecutar la prueba t de Student para muestras independientes, se obtuvieron los siguientes valores:

- **Estadístico t:** 6.8483
- **p-valor:** 7.75e-12

Interpretación:

- **Estadístico t (6.8483):**
  Este valor indica que la diferencia entre las medias de la cantidad de detenidos para Chile y Estados Unidos es aproximadamente 6.85 veces mayor que la variabilidad esperada entre ambos grupos. Un valor t tan alto sugiere una diferencia considerable en las medias.
- **p-valor (7.75e-12):**
  El p-valor representa la probabilidad de obtener una diferencia igual o mayor que la observada, asumiendo que la hipótesis nula es cierta. En este caso, el p-valor es extremadamente bajo (muy menor que 0.05), lo que indica que es muy improbable obtener estos resultados si no existiera una diferencia real entre los grupos.

Evaluación según el Nivel de Significancia (alfa = 0.05):

Comparación con alfa:
Dado que el p-valor (7.75e-12) es mucho menor que el nivel de significancia elegido (alfa = 0.05), se cuenta con evidencia estadística suficiente para rechazar la hipótesis nula.
Se rechaza la hipótesis nula (H0: $\mu_{(Chile)} = \mu_{(EE.UU.)}$) en favor de la hipótesis alternativa (H1: $\mu_{(Chile)} \neq \mu_{(EE.UU.)}$). Esto significa que, en el contexto de esta investigación, existe una diferencia estadísticamente significativa en la cantidad de detenidos entre Chile y Estados Unidos. En otras palabras, el país de detención tiene un efecto relevante sobre la cantidad de detenidos.

6. Redacte una conclusión basada en los resultados del contraste de hipótesis. Explique qué indican los hallazgos sobre los datos y cómo responden a la pregunta de investigación inicial. Por ejemplo, si rechaza la hipótesis nula, indique que hay evidencia estadística de que existe una relación significativa entre las variables en estudio. Asegúrese de ser claro y específico en su interpretación, y vincule los resultados de la prueba con el contexto de su investigación.

Los resultados obtenidos en el contraste de hipótesis indican que existe una diferencia estadísticamente significativa en la cantidad de detenidos entre Chile y Estados Unidos. Con un estadístico t de 6.8483 y un p-valor de 7.75e-12, el valor del p-valor es mucho menor al nivel de significancia establecido (alfa = 0.05), lo que lleva a rechazar la hipótesis nula. Esto significa que la diferencia observada en las medias no se debe al azar, sino que existe evidencia suficiente para afirmar que el país de detención influye significativamente en la cantidad de detenidos. En el contexto de la investigación, estos hallazgos responden a la pregunta inicial, confirmando que factores relacionados con el país (en este caso, la comparación entre Chile y Estados Unidos) tienen un impacto relevante en los registros de detenidos.

Regresión Lineal.

En la actividad anterior se realizó una regresión simple, para esta actividad debe seleccionar al menos tres variables continuas independientes (características) y una variable dependiente (variable objetivo) que deseen predecir.

1. Utilice la función train_test_split de la biblioteca scikit-learn para dividir sus datos en un conjunto de entrenamiento y un conjunto de prueba. Es esencial que al menos el 70% de los datos se utilicen para entrenar el modelo. Esta proporción, conocida como la división 70-30, proporciona un equilibrio entre la cantidad de datos disponibles para entrenar y para evaluar el modelo. Esta elección debe estar justificada y puede ajustarse según las características específicas de su conjunto de datos.

```
# Importar las bibliotecas necesarias
import pandas as pd
from sklearn.model_selection import train_test_split

# Cargar la base de datos
ruta = r'C:\Users\steph\colombianos detenidos exterior.csv'
# Se especifica el separador, codificación y que "UBICACIÓN PAÍS" sea tratado como string para evitar advertencias.
df = pd.read_csv(ruta, sep=';', encoding='utf-8', on_bad_lines='skip', dtype={'UBICACIÓN PAÍS': str})

# Procesamiento de la variable de tiempo (FECHA PUBLICACIÓN)
# Convertir la columna "FECHA PUBLICACIÓN" a tipo datetime y crear la variable ordinal FECHA_NUM
df['FECHA_PUBLICACION'] = pd.to_datetime(df['FECHA PUBLICACIÓN'], dayfirst=True)
df['FECHA_NUM'] = df['FECHA_PUBLICACION'].apply(lambda x: x.toordinal())

# Paso 4: Seleccionar las variables para el modelo
# Variables independientes: LATITUD, LONGITUD y FECHA_NUM
# Variable dependiente: CANTIDAD
X = df[['LATITUD', 'LONGITUD', 'FECHA_NUM']]
y = df['CANTIDAD']

# Dividir los datos en conjuntos de entrenamiento y prueba (70% entrenamiento, 30% prueba)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Mostrar el tamaño de los conjuntos
print("Tamaño del conjunto de entrenamiento:", X_train.shape)
print("Tamaño del conjunto de prueba:", X_test.shape)
```

```
Tamaño del conjunto de entrenamiento: (246160, 3)
Tamaño del conjunto de prueba: (105498, 3)
```

2. Emplee la clase LinearRegression de scikit-learn para instanciar un modelo de regresión lineal multidimensional. Este modelo permitirá explorar y modelar las relaciones entre múltiples características de sus datos.

```
# Eliminar filas que contengan NaN en X o y
X_clean = X.dropna()
y_clean = y[X_clean.index]  # Aseguramos que y tenga el mismo índice que X_clean

# Dividir los datos limpios en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_clean, y_clean, test_size=0.3, random_state=42)

# Instanciar y ajustar el modelo
modelo_lr = LinearRegression()
modelo_lr.fit(X_train, y_train)

print("Coeficientes:", modelo_lr.coef_)
print("Intercepto:", modelo_lr.intercept_)
```

```
Coeficientes: [ 8.80772830e-09 -1.71063852e-08  9.02782196e-05]
Intercepto: -62.24127675661179
```

3. Utilice el método fit del modelo para llevar a cabo el ajuste a los datos de entrenamiento. Este paso es fundamental ya que el modelo aprenderá a representar la relación entre las características y la variable dependiente a partir de estos datos.

```python
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import pandas as pd

# Supongamos que X e y ya están definidos, donde:
# X contiene las características: LATITUD, LONGITUD, FECHA_NUM
# y es la variable dependiente: CANTIDAD

# Imputar valores faltantes usando la media
imputer = SimpleImputer(strategy='mean')
X_imputed = pd.DataFrame(imputer.fit_transform(X), columns=X.columns, index=X.index)

# Dividir en conjunto de entrenamiento y prueba (70-30)
X_train, X_test, y_train, y_test = train_test_split(X_imputed, y, test_size=0.3, random_state=42)

# Instanciar el modelo de regresión lineal
modelo_lr = LinearRegression()

# Utilizar el método fit para ajustar el modelo a los datos de entrenamiento
modelo_lr.fit(X_train, y_train)

# Una vez ajustado, se pueden visualizar los coeficientes y el intercepto
print("Coeficientes:", modelo_lr.coef_)
print("Intercepto:", modelo_lr.intercept_)
```

```
Coeficientes: [ 7.57119627e-09 -1.74503757e-08 -2.78205355e-04]
Intercepto: 209.55481361699495
```

4. Utilice el atributo coef_ del modelo para obtener los coeficientes de regresión asociados a cada característica. Estos coeficientes representan la contribución de cada característica en la predicción de la variable dependiente. Es crucial analizar si los resultados obtenidos concuerdan con las expectativas previas. Si hay discrepancias, es posible que algunas características tengan un impacto inesperado o que se necesite un ajuste adicional del modelo.

En el modelo, los coeficientes obtenidos fueron:

- **LATITUD:** 7.57e-09
- **LONGITUD:** -1.75e-08

- **FECHA_NUM:** -2.78e-04

```
# Obtener los coeficientes
coeficientes = modelo_lr.coef_
columnas = X_train.columns

# Imprimir los coeficientes asociados a cada característica
for col, coef in zip(columnas, coeficientes):
    print(f"Coeficiente para {col}: {coef}")

# Imprimir el intercepto
print("Intercepto:", modelo_lr.intercept_)
```

```
Coeficiente para LATITUD: 7.571196271338458e-09
Coeficiente para LONGITUD: -1.7450375730715175e-08
Coeficiente para FECHA_NUM: -0.0002782053551966476
Intercepto: 209.55481361699495
```

Los resultados actuales muestran que las variables geográficas (LATITUD y LONGITUD) tienen coeficientes extremadamente pequeños, mientras que FECHA_NUM presenta un coeficiente mayor, aunque aún pequeño en valor absoluto. Esto puede o no coincidir con las expectativas previas:

- **Expectativas Previas:**
  En principio, se esperaba que variables como LATITUD y LONGITUD tuvieran cierto impacto en la cantidad de detenidos si existe una relación geográfica clara. Sin embargo, en el modelo actual, su contribución es prácticamente insignificante.
  Por otro lado, el coeficiente de FECHA_NUM (aunque negativo) podría sugerir una leve tendencia decreciente a lo largo del tiempo, pero la magnitud pequeña indica que, en el contexto del modelo, el efecto del tiempo no es fuerte.

- **Posibles Discrepancias e Implicaciones:**

Escalado de Variables:
Dado que FECHA_NUM probablemente tenga valores muy grandes (por ejemplo, números ordinales que superan los 700,000) y las coordenadas geográficas se encuentran en una escala

distinta, la diferencia en las escalas puede estar afectando la magnitud de los coeficientes. Aplicar un escalado estándar a las variables independientes podría ayudar a interpretar mejor el impacto relativo de cada predictor.

Selección de Variables:

Si se esperaba que la ubicación (LATITUD y LONGITUD) tuviera un impacto mayor, los resultados indican que tal vez estas variables, en su forma actual, no capturan de manera efectiva la variabilidad en la cantidad de detenidos.

Podría considerarse incluir o transformar variables adicionales, por ejemplo, variables derivadas que agrupen regiones o que combinen información geográfica con otra dimensión, para explicar mejor el fenómeno.

Relación Lineal:

La relación entre las características seleccionadas y la cantidad de detenidos podría no ser estrictamente lineal. Esto sugiere la posibilidad de que se requiera un modelo más complejo o la inclusión de términos polinómicos o interacciones entre variables para capturar mejor las relaciones subyacentes.

Si bien el modelo ha sido ajustado y se obtuvieron los coeficientes, las magnitudes extremadamente pequeñas de los coeficientes para LATITUD y LONGITUD y el coeficiente moderado de FECHA_NUM sugieren que, en el modelo actual, estas variables tienen un impacto muy limitado sobre la predicción de la cantidad de detenidos. Esto puede no coincidir con las expectativas previas de un efecto geográfico o temporal significativo.

Por lo tanto, es posible que:

- Se requiera un escalado de las variables para mejorar la interpretabilidad.
- Se deba reconsiderar o transformar la selección de variables para captar mejor la variabilidad de los datos.

- Se explore un modelo alternativo que permita capturar relaciones no lineales o interacciones entre variables.

5. Utilice el atributo intercept_ del modelo para obtener el término de intercepción. Este término representa el valor esperado de la variable dependiente cuando todas las características son iguales a cero. Es una parte importante de la ecuación de regresión y puede influir significativamente en las predicciones del modelo.

```
# Intercepto:
print("Intercepto:", modelo_lr.intercept_)
```

```
Intercepto: 209.55481361699495
```

El intercepto obtenido es 209.5548, lo que indica el valor base de la cantidad de detenidos en el modelo. Es importante considerar que, si bien su valor puede no tener un significado práctico directo (debido a la escala de las variables independientes), juega un papel fundamental en la ecuación de regresión y en la precisión de las predicciones del modelo.

6. Utilice el conjunto de prueba que ha separado previamente para realizar predicciones utilizando el modelo ajustado. Utilice el método predict para obtener las predicciones del modelo sobre este conjunto de datos independiente.

```
# Utilizar el conjunto de prueba para hacer predicciones
y_pred = modelo_lr.predict(X_test)

# Mostrar algunas de las predicciones
print("Predicciones en el conjunto de prueba:", y_pred[:10])
```

```
Predicciones en el conjunto de prueba: [5.64787154 4.73947483 4.58646188 4.0376084  4.62068114 4.18946272
 5.02230947 4.60168293 4.66903659 5.64787154]
```

Predicción en el Conjunto de Prueba:

Con modelo_lr.predict(X_test) se obtienen las predicciones para la variable dependiente (CANTIDAD) utilizando el conjunto de prueba. Esto permite evaluar cómo el modelo generaliza a nuevos datos que no fueron utilizados durante el entrenamiento.

Visualización de Resultados:

Se imprimen las primeras 10 predicciones para tener una idea de los valores que el modelo estima para la cantidad de detenidos en el conjunto de datos independiente.

7. Calcule el Error Cuadrático Medio (MSE), que es una medida de la calidad de las predicciones del modelo. Representa el promedio de los cuadrados de las diferencias entre las predicciones y los valores reales. Un MSE más bajo indica que las predicciones del modelo están más cerca de los valores reales, lo que sugiere un modelo más preciso.

```
from sklearn.metrics import mean_squared_error

# Calcular el Error Cuadrático Medio (MSE)
mse = mean_squared_error(y_test, y_pred)
print("Error Cuadrático Medio (MSE):", mse)
```

```
Error Cuadrático Medio (MSE): 299.63630772052596
```

El resultado obtenido, un MSE de aproximadamente 299.64, indica que, en promedio, el cuadrado de la diferencia entre las predicciones del modelo y los valores reales es de 299.64. Esto significa que, en promedio, la discrepancia al cuadrado entre la cantidad predicha de detenidos y la cantidad real es de ese valor.

8. Calcule el coeficiente de determinación (R^2), que proporciona una medida de la proporción de la varianza en la variable dependiente que es predecible a partir de las características. Un R^2 cercano a 1 indica un buen ajuste del modelo a los

datos.

```
from sklearn.metrics import r2_score

# Calcular el coeficiente de determinación (R²)
r2 = r2_score(y_test, y_pred)
print("Coeficiente de determinación (R²):", r2)
```

```
Coeficiente de determinación (R²): 0.002046274784165303
```

El bajo valor de R^2 indica que el modelo actual no es eficaz para predecir la cantidad de detenidos a partir de LATITUD, LONGITUD y FECHA_NUM. Esto sugiere la necesidad de:

- Revisar la selección y transformación de las variables.
- Considerar la inclusión de otras variables que puedan tener mayor relevancia en la explicación del fenómeno.
- Explorar modelos alternativos o técnicas de transformación para capturar relaciones no lineales.

9. Considerando la problemática inicial y los resultados obtenidos, analice si la regresión lineal ha proporcionado información adicional sobre los datos. Evalúe si el modelo ha logrado capturar las relaciones entre las características y la variable dependiente de manera significativa y si ha proporcionado conocimientos nuevos.

Evaluación General:

La regresión lineal aplicada utilizando LATITUD, LONGITUD y FECHA_NUM como variables predictoras para la cantidad de detenidos no ha logrado capturar de manera significativa la variabilidad de la variable dependiente. Con un R^2 de aproximadamente 0.002, el modelo explica menos del 0.2% de la variación en la cantidad de detenidos.

Puntos Clave del Análisis:

Bajo Poder Explicativo:

El valor de R^2 tan bajo indica que las variables seleccionadas no logran aportar información suficiente para predecir la cantidad de detenidos. En otras palabras, la relación lineal entre LATITUD, LONGITUD, FECHA_NUM y la cantidad de detenidos es muy débil, lo que sugiere que estos predictores, en su forma actual, no capturan los factores relevantes del fenómeno.

Posible Necesidad de Revisión en la Selección de Variables:

- Es posible que la problemática requiera considerar otras variables o condiciones (por ejemplo, tipo de delito, consulado, grupo de edad, etc.) que estén más directamente relacionadas con la cantidad de detenidos.
- La transformación o el escalado de las variables también podría mejorar la capacidad del modelo para detectar relaciones lineales si estas relaciones están ocultas por diferencias en las escalas de medición.

Conocimiento Nuevo y Limitaciones:

- **Conocimiento Adicional:**
  Aunque el modelo no ha proporcionado un buen ajuste, esto mismo es un hallazgo relevante: sugiere que, con los predictores seleccionados, no se puede explicar de forma significativa la variabilidad en la cantidad de detenidos.
- **Limitaciones del Modelo Lineal:**
  Es probable que la relación entre las variables de ubicación, tiempo y la cantidad de detenidos no sea estrictamente lineal. Esto puede motivar la exploración de otros modelos o técnicas de transformación (por ejemplo, modelos no lineales o la inclusión de interacciones y términos polinómicos) para capturar relaciones

complejas.

El modelo de regresión lineal aplicado no ha proporcionado información adicional valiosa sobre las relaciones entre LATITUD, LONGITUD, FECHA_NUM y la cantidad de detenidos. Esto sugiere que:

- Las variables seleccionadas, en el formato actual, no son adecuadas para explicar el fenómeno.
- Es necesario revisar la selección y transformación de las características o considerar otros predictores que tengan un impacto más directo.
- Se abre la posibilidad de explorar modelos alternativos que puedan capturar relaciones no lineales o interacciones entre variables.

10. Indique cuál sería la siguiente prueba a realizar y explique la razón detrás de esta elección. Mencione los resultados esperados y cómo estos podrían contribuir a una comprensión más profunda del problema o a la mejora del modelo.

Dado que el modelo de regresión lineal, basado en las variables LATITUD, LONGITUD y FECHA_NUM, explica solo un 0.2% de la variabilidad en la cantidad de detenidos ($R^2 \approx 0.002$), es evidente que estas variables en su forma actual no capturan adecuadamente los factores que influyen en el fenómeno.

Regresión Logística

Razón de la Elección:

- **Conversión de la Variable Dependiente:**
  Dado que la regresión lineal no está proporcionando un buen ajuste, una alternativa es reformular el problema como uno de clasificación. Por ejemplo, se puede transformar la

variable "CANTIDAD" en una variable binaria o categórica, mediante la definición de un umbral que distinga entre "alta" y "baja" cantidad de detenidos. Esto permite explorar si ciertos factores (por ejemplo, características del país, tipo de delito, etc.) influyen significativamente en la probabilidad de que se registre una alta cantidad de detenidos.

- **Captura de Relaciones No Lineales:**
  La regresión logística es adecuada para modelar relaciones no lineales en términos de probabilidades y puede ser más sensible a detectar efectos de ciertas variables que, en el caso del modelo lineal, quedan diluidos por la escala o por relaciones complejas.

Resultados Esperados y Contribución al Problema:

- **Resultados Esperados:**
  Se espera que la regresión logística proporcione un modelo con mejores métricas de clasificación (por ejemplo, mayor precisión, sensibilidad y especificidad) al distinguir entre eventos con alta y baja cantidad de detenidos. Además, se podrán obtener coeficientes (o razones de verosimilitud) que indiquen la influencia relativa de cada predictor en la probabilidad de que se presente una alta cantidad de detenidos.

- **Contribución a una Comprensión Más Profunda:**
  - **Identificación de Factores Críticos:**
    Permite identificar de manera clara qué variables (o condiciones) incrementan significativamente el riesgo de tener un alto número de detenidos, lo que resulta crucial para orientar intervenciones o políticas de asistencia consular.
  - **Mejora del Modelo:**
    Al convertir el problema a uno de clasificación, se pueden explorar transformaciones y selección de variables adicionales (por ejemplo, incluir variables categóricas como el tipo de delito o el consulado), lo que podría mejorar la capacidad predictiva y ofrecer un modelo más ajustado a la realidad.

- **Perspectiva Complementaria:**
  La regresión logística ofrece una perspectiva complementaria a la regresión lineal. Mientras que la regresión lineal se centra en explicar la varianza en los valores numéricos, la regresión logística se orienta a clasificar eventos, lo que podría resultar en un mejor entendimiento de las condiciones que determinan la magnitud de las detenciones.

La siguiente prueba a realizar sería una regresión logística, que, al transformar la variable dependiente en una categórica, permitirá identificar y cuantificar los efectos significativos de las variables predictoras sobre la probabilidad de que ocurra un evento de alta incidencia. Este enfoque contribuirá a una comprensión más profunda del problema y ofrecerá una alternativa que podría mejorar el desempeño predictivo en comparación con el modelo de regresión lineal actual.

Regresión Logística.

Partiendo de la base datos con la que se ha estado trabajando, selecciona dos variables, una variable continua y otra variable categórica o binaria. Si no se tienen datos cate puede crear dos categorías utilizando algún umbral (solo para los propósitos del ejercicio, siempre que se tenga información continua se debe aprovechar).

1. Divida los datos en un conjunto de entrenamiento y un conjunto de prueba utilizando la función train_test_split de scikit-learn. Asegúrese de que al menos el 70% de los datos se utilicen para entrenar el modelo, como es común en problemas de clasificación.

Dado que la base de datos no tiene una variable categórica para clasificación, creare una nueva variable binaria a partir de "CANTIDAD". Para este ejercicio, se define que:

- Si CANTIDAD es mayor que 1 (por ejemplo, usando la mediana que en nuestro análisis previo fue 1), se clasifica como 1 (alta cantidad).
- Si CANTIDAD es igual o menor que 1, se clasifica como 0 (baja cantidad).

Selección de las Variables para la Regresión Logística

En este ejercicio:

- **Variable continua predictora:** FECHA_NUM
  (Esta variable se creó previamente a partir de "FECHA PUBLICACIÓN" y representa el valor ordinal de la fecha).
- **Variable objetivo-binaria:** CANTIDAD_BIN

Dividir los Datos en Conjunto de Entrenamiento y Conjunto de Prueba

Se usara la función train_test_split para separar al menos el 70% de los datos para entrenamiento y el 30% para evaluación. Esto es común en problemas de clasificación para asegurar que el modelo tenga suficientes datos para aprender y para validar su desempeño de forma independiente.

```
# Crear la variable binaria a partir de 'CANTIDAD'
df['CANTIDAD_BIN'] = (df['CANTIDAD'] > 1).astype(int)
```

```
# Seleccionar la variable continua predictora y la variable objetivo binaria
X_continuo = df[['FECHA_NUM']]
y_binaria = df['CANTIDAD_BIN']
```

```
from sklearn.model_selection import train_test_split

# Dividir los datos en conjunto de entrenamiento (70%) y de prueba (30%)
X_train, X_test, y_train, y_test = train_test_split(X_continuo, y_binaria, test_size=0.3, random_state=42)

# Imprimir tamaños de los conjuntos
print("Tamaño del conjunto de entrenamiento:", X_train.shape)
print("Tamaño del conjunto de prueba:", X_test.shape)
```

```
Tamaño del conjunto de entrenamiento: (246160, 1)
Tamaño del conjunto de prueba: (105498, 1)
```

2. Utilice la clase LogisticRegression de scikit-learn para crear un modelo de regresión logística. Recuerde que la regresión logística es ideal para problemas de clasificación binaria.

```
from sklearn.linear_model import LogisticRegression

# Instanciar el modelo de regresión logística
# Se usa 'liblinear' como solver, que es adecuado para datasets moderados y problemas binarios.
modelo_log = LogisticRegression(solver='liblinear', random_state=42)

# Ajustar el modelo utilizando el conjunto de entrenamiento
modelo_log.fit(X_train, y_train)

# Imprimir los coeficientes y el intercepto del modelo
print("Coeficientes:", modelo_log.coef_)
print("Intercepto:", modelo_log.intercept_)
```

```
Coeficientes: [[-3.02581167e-07]]
Intercepto: [-4.09799232e-13]
```

3. Ajuste el modelo a los datos de entrenamiento utilizando el método fit. Prueba con diferentes valores de C (parámetro de regularización) y diferentes algoritmos de optimización (solver). Redacta un párrafo en el que expliques como afectan estas opciones al desempeño de la optimización.

```python
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Definir una lista de valores para C y solvers a probar
valores_C = [0.01, 0.1, 1, 10, 100]
solvers = ['liblinear', 'lbfgs']

resultados = []

for solver in solvers:
    for C in valores_C:
        # Instanciar el modelo con el valor de C y solver actual
        modelo = LogisticRegression(C=C, solver=solver, random_state=42, max_iter=1000)
        modelo.fit(X_train, y_train)

        # Ajustar el modelo a los datos de entrenamiento
        y_pred = modelo.predict(X_test)
        acc = accuracy_score(y_test, y_pred)
        resultados.append((solver, C, acc))
        print(f"Solver: {solver}, C: {C}, Accuracy: {acc:.4f}")

# Opcional: visualizar los resultados en un DataFrame para mayor claridad
df_resultados = pd.DataFrame(resultados, columns=['Solver', 'C', 'Accuracy'])
print(df_resultados)
```

```
Solver: liblinear, C: 0.01, Accuracy: 0.5577
Solver: liblinear, C: 0.1, Accuracy: 0.5577
Solver: liblinear, C: 1, Accuracy: 0.5577
Solver: liblinear, C: 10, Accuracy: 0.5577
Solver: liblinear, C: 100, Accuracy: 0.5577
Solver: lbfgs, C: 0.01, Accuracy: 0.5577
Solver: lbfgs, C: 0.1, Accuracy: 0.5577
Solver: lbfgs, C: 1, Accuracy: 0.5577
Solver: lbfgs, C: 10, Accuracy: 0.5577
Solver: lbfgs, C: 100, Accuracy: 0.5577
      Solver       C  Accuracy
0  liblinear    0.01  0.557707
1  liblinear    0.10  0.557707
2  liblinear    1.00  0.557707
3  liblinear   10.00  0.557707
4  liblinear  100.00  0.557707
5      lbfgs    0.01  0.557707
6      lbfgs    0.10  0.557707
7      lbfgs    1.00  0.557707
8      lbfgs   10.00  0.557707
9      lbfgs  100.00  0.557707
```

El solver es el algoritmo que se encarga de encontrar los mejores parámetros para el modelo. Por ejemplo, el solver 'liblinear' funciona bien para problemas de clasificación binaria y con conjuntos de datos pequeños o medianos. En cambio, 'lbfgs' suele ser más rápido y manejar mejor problemas grandes o con muchas clases, aunque a veces necesita que los datos estén bien escalados para funcionar correctamente.

4. Obtenga los coeficientes de regresión para cada característica utilizando el atributo coef. ¿Cuál es la diferencia en cómo se utilizan estos coeficientes para realizar predicciones respecto o a los obtenidos mediante una regresión lineal?

En la regresión logística, al igual que en la regresión lineal, se obtiene un coeficiente para cada característica usando el atributo coef_. Sin embargo, la forma en la que se utilizan para hacer predicciones es diferente:

- **Regresión Lineal:**
  Los coeficientes se multiplican directamente por los valores de las características y se suman (junto con el intercepto) para obtener una predicción continua. Es decir, la ecuación es

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

  :

- Aquí, los coeficientes representan el cambio esperado en la variable dependiente por unidad de cambio en cada predictor.

- **Regresión Logística:**
  Los coeficientes representan el cambio en los logaritmos de las probabilidades (o "log-odds") de que el evento de interés ocurra, por cada unidad de cambio en la característica. La predicción se realiza primero calculando una combinación lineal similar a la regresión

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

lineal, pero luego se transforma mediante la función logística (sigmoide) para obtener una probabilidad:

En este caso, los coeficientes indican cómo varían los log-odds de la probabilidad del evento con respecto a los cambios en las características, y no la variable de salida directamente.

Diferencia:

Mientras que en la regresión lineal los coeficientes se usan para predecir valores numéricos directamente, en la regresión logística se usan para calcular los log-odds, que luego se convierten en probabilidades. Esto hace que la interpretación y el uso de los coeficientes en el contexto de la predicción sean distintos en cada caso.

5. Obtenga el término de intercepción utilizando el atributo intercept_. Esto representa la probabilidad de pertenecer a la clase positiva cuando todas las características son iguales a cero.

```
# intercepto
print("Intercepto:", modelo_log.intercept_)
```

```
Intercepto: [-4.09799232e-13]
```

En la regresión logística, el atributo intercept_ da el término de intercepción, que es básicamente el "valor base" del modelo. Este valor representa el logaritmo de las probabilidades (log-odds) de que la variable dependiente sea 1 cuando todas las características predictoras son cero.

En palabras simples, el intercepto dice cuál es la probabilidad base de que un registro pertenezca a la clase positiva si no se tienen en cuenta los efectos de las demás variables.

6. Utilizando el conjunto de prueba, realiza predicciones utilizando el modelo ajustado con el método predict. Asegúrate de que las predicciones estén en el rango [0, 1] y establece un umbral adecuado para clasificarlas en las clases, justifique su elección.

Utilizare el método predict_proba para obtener la probabilidad de pertenecer a la clase 1 (alta cantidad) para cada registro del conjunto de prueba. Esto garantiza que los valores estén en el rango [0, 1].

```
# Obtener las probabilidades de pertenecer a la clase 1
y_proba = modelo_log.predict_proba(X_test)[:, 1]
print("Algunas probabilidades:", y_proba[:10])
```

```
Algunas probabilidades: [0.44442022 0.44436127 0.44432018 0.44431794 0.44432937 0.44439123
 0.44435216 0.4444179  0.44432482 0.44442022]
```

Generalmente se utiliza un umbral de 0.5 para decidir la clase. Esto significa que:

- Si la probabilidad es mayor o igual a 0.5, clasificamos el registro como 1 (alta cantidad).
- Si es menor a 0.5, se clasifica como 0 (baja cantidad).

```
# Definir el umbral
umbral = 0.5
# Clasificar las predicciones
y_pred_bin = (y_proba >= umbral).astype(int)
print("Predicciones clasificadas:", y_pred_bin[:10])
```

```
Predicciones clasificadas: [0 0 0 0 0 0 0 0 0 0]
```

Elegí el umbral de 0.5 porque es lo más común en problemas de clasificación binaria. Esto significa que, si el modelo dice que hay un 50% o más de probabilidad de que un registro pertenezca a la clase positiva, lo clasificamos como positivo. Si la probabilidad es menor al 50%, lo clasificamos como negativo. Es una elección estándar cuando no tenemos razones para dar

más importancia a una clase que a la otra, ya que se asume que ambas tienen el mismo peso.

7. Por ejemplo, en un examen para detectar cáncer se puede tener una menor tolerancia al error que seleccionar un producto en un sistema de recomendación.

En la práctica, el valor del umbral puede variar mucho dependiendo de las consecuencias de equivocarse en la predicción. Por ejemplo, si se trata de un examen para detectar cáncer, un falso negativo (decirle a alguien que no tiene cáncer cuando sí lo tiene) puede ser muy grave, así que preferimos poner un umbral más bajo para ser más "sospechosos" y reducir las posibilidades de pasar por alto un caso real. En cambio, si estamos recomendando un producto en un sistema de compras en línea, un error no es tan crítico: si recomendamos algo que no le interesa al usuario, la consecuencia no es tan grave. Por eso, en ese caso, podríamos mantener el umbral estándar (0.5) o incluso subirlo, dependiendo de cuánto grave sea equivocarse.

8. Calcula la precisión del modelo en el conjunto de prueba. La precisión es importante, pero también considera otras métricas como la sensibilidad y la especificidad para

```python
from sklearn.metrics import accuracy_score, recall_score, confusion_matrix

# Calcular la precisión
precision_modelo = accuracy_score(y_test, y_pred_bin)
print("Precisión (Accuracy):", precision_modelo)

# Calcular la matriz de confusión
matriz_confusion = confusion_matrix(y_test, y_pred_bin)
print("Matriz de Confusión:\n", matriz_confusion)

# Obtener la sensibilidad y la especificidad
TN, FP, FN, TP = matriz_confusion.ravel()

sensibilidad = TP / (TP + FN)  # Tasa de verdaderos positivos (recall de la clase 1)
especificidad = TN / (TN + FP) # Tasa de verdaderos negativos (recall de la clase 0)

print("Sensibilidad (Recall de la clase positiva):", sensibilidad)
print("Especificidad (Recall de la clase negativa):", especificidad)
```

```
Precisión (Accuracy): 0.557707255113841
Matriz de Confusión:
 [[58837     0]
 [46661     0]]
Sensibilidad (Recall de la clase positiva): 0.0
Especificidad (Recall de la clase negativa): 1.0
```

obtener una imagen completa del rendimiento del modelo.

9. Calcula la matriz de confusión del modelo en el conjunto de prueba. La matriz de confusión proporciona información detallada sobre los aciertos y los errores de clasificación.

```
from sklearn.metrics import confusion_matrix

# Matriz de confusion
matriz_conf = confusion_matrix(y_test, y_pred_bin)
print("Matriz de Confusión:\n", matriz_conf)

# Desglosar los valores en TN, FP, FN, TP
TN, FP, FN, TP = matriz_conf.ravel()
print("TN:", TN, "FP:", FP, "FN:", FN, "TP:", TP)
```

```
Matriz de Confusión:
 [[58837     0]
 [46661     0]]
TN: 58837 FP: 0 FN: 46661 TP: 0
```

La matriz de confusión se muestra en la forma:

- **TN (True Negatives): 58,837**
- **FP (False Positives): 0**
- **FN (False Negatives): 46,661**
- **TP (True Positives): 0**

la matriz de confusión permite una visión detallada de cómo el modelo clasifica cada ejemplo, ayudando a entender no solo cuántos aciertos y errores hay, sino también de qué tipo son esos errores, lo cual es fundamental para refinar y mejorar el modelo.

Conclusiones.

- ✓ Al comparar la cantidad de detenidos entre Chile y Estados Unidos, se encuentra una diferencia significativa (p-valor muy bajo). Esto indica que el país de detención sí influye en la cantidad de detenidos, lo que respalda la hipótesis de que la ubicación afecta los resultados.
- ✓ El modelo de regresión lineal, que utilizó LATITUD, LONGITUD y FECHA_NUM para predecir la cantidad de detenidos, mostró un R^2 muy bajo (aproximadamente 0.002). Esto significa que, con estas variables, el modelo no logra explicar casi nada de la variabilidad en la cantidad de detenidos, por lo que quizás se necesiten otros predictores o transformar las variables.
- ✓ Además, el error cuadrático medio (MSE) del modelo lineal fue alto, lo que sugiere que las predicciones no se acercan lo suficiente a los valores reales. Esto refuerza la idea de que las variables utilizadas no capturan bien el comportamiento de los detenidos.
- ✓ Cuando se transformó la variable "CANTIDAD" a una clasificación binaria (por ejemplo, alta vs. baja cantidad), el modelo de regresión logística terminó prediciendo todo como la clase negativa (baja cantidad). Esto se evidencia en la matriz de confusión, donde no se detectaron casos positivos. Esto podría deberse a que los datos están muy desbalanceados o a que la variable predictora elegida (FECHA_NUM) no es suficiente para distinguir entre ambas clases.
- ✓ En general, ambos enfoques (regresión lineal y logística) muestran que, con las variables actuales y el procesamiento aplicado, no se logra capturar completamente la complejidad del fenómeno de los detenidos. Esto sugiere que para mejorar la capacidad predictiva, habría que incluir más variables relevantes, aplicar técnicas de escalado o balanceo de clases, o explorar modelos alternativos que puedan capturar relaciones más complejas.