Análisis de resultados

1. La matriz de correlación (`df_clima.corr()`) muestra relaciones entre las variables. Por ejemplo, la temperatura máxima y mínima tienen una correlación positiva moderada de 0.915255, lo cual es esperado ya que ambas variables están relacionadas con la temperatura. Además, la precipitación tiene una correlación positiva débil con el mes (0.141958) y el año (0.203246), lo que podría indicar patrones estacionales en las precipitaciones.

2. El gráfico de barras que muestra la cantidad de países repetidos (`conteo_paises.plot(kind='bar')`) destaca que algunos países tienen una mayor representación de observaciones del ave Passer domesticus. Esto puede deberse a varios factores, como condiciones ambientales favorables, esfuerzos de muestreo más intensos o incluso sesgos en la recolección de datos.

3. La selección de variables de entrada (`input = ['month', 'year', 'tempMax', 'precipitacion', 'tempMin']`) sugiere que el modelo intenta capturar la influencia de factores climáticos y temporales en la distribución geográfica del ave. Sin embargo, es posible que se requieran variables adicionales, como la altitud, el tipo de hábitat o la distancia a áreas urbanas, para mejorar el poder predictivo del modelo.

4. La precisión reportada del modelo (`Precisión: 0.602203182374541`) es moderada, lo que indica que el modelo tiene dificultades para predecir correctamente el código de país en una proporción sustancial de los casos. Esto puede deberse a varias razones, como la complejidad de los patrones migratorios, la presencia de ruido o valores atípicos en los datos, o la falta de variables predictoras relevantes.

Respuestas a lo que pide elias

a. Your objectives for estimation using the provided data.

El objetivo principal es desarrollar un modelo predictivo capaz de estimar la ubicación geográfica (país) de la especie Passer domesticus en función de variables climáticas y temporales. Esto permitiría comprender mejor cómo factores ambientales, como la temperatura, precipitación y estacionalidad, influyen en los patrones migratorios y la distribución espacial de esta ave. Además,

el modelo podría utilizarse para realizar predicciones sobre la presencia potencial de la especie en diferentes regiones, lo que sería valioso para estudios ecológicos y de conservación.

b. Details about your initial model selection.

La selección inicial del modelo de regresión logística multinomial se basa en la naturaleza de la variable objetivo, que es el código numérico del país. Al tener múltiples clases (países) posibles, la regresión logística multinomial es capaz de manejar este problema de clasificación multiclase de manera adecuada. Además, este tipo de modelo es ampliamente utilizado y tiene una interpretación relativamente sencilla, lo que lo convierte en una opción razonable para comenzar el análisis.

Sin embargo, es importante tener en cuenta que la elección del modelo también depende de los supuestos y las características de los datos. Si se violan los supuestos de la regresión logística (por ejemplo, la independencia de las observaciones o la ausencia de multicolinealidad), podría ser necesario explorar otros modelos alternativos, como los árboles de decisión, las redes neuronales o los modelos ensemble.

c. Information on validation methods and the metrics employed.

El método de validación utilizado es la división del conjunto de datos en conjuntos de entrenamiento y prueba (`train_test_split`). Al separar una parte de los datos como conjunto de prueba, se puede obtener una estimación más realista del desempeño del modelo en datos no vistos durante el entrenamiento, lo que ayuda a evitar el sobreajuste.

La métrica empleada es la precisión (`accuracy_score`), que es una métrica ampliamente utilizada para problemas de clasificación. La precisión mide la fracción de predicciones correctas sobre el total de predicciones realizadas por el modelo. Sin embargo, es importante tener en cuenta que la precisión puede ser engañosa en casos de conjuntos de datos desbalanceados, donde una clase tiene una representación mucho mayor que las otras. En estos casos, puede ser más informativo utilizar métricas adicionales, como la precisión, el recall o la puntuación F1 para cada clase.

Además, dependiendo de los objetivos del proyecto y los costos asociados a los diferentes tipos de errores de clasificación, podría ser más apropiado optimizar métricas alternativas, como el área bajo la curva ROC (AUC-ROC) o la entropía cruzada.

d. Preliminary conclusions drawn from your analysis to date.

- Existen correlaciones moderadas entre algunas variables climáticas y la ubicación geográfica (país), lo que sugiere que estos factores podrían influir en los patrones migratorios de la especie Passer domesticus. Sin embargo, es necesario explorar estas relaciones con más profundidad y considerar la inclusión de variables adicionales relevantes.

- El modelo actual de regresión logística multinomial tiene una precisión relativamente baja (alrededor del 60%), lo que indica que aún hay margen para mejorar su capacidad predictiva. Esto podría lograrse mediante ajustes en el preprocesamiento de datos, la selección de variables, la elección del modelo o la optimización de hiperparámetros.

- El conjunto de datos muestra que ciertos países tienen una mayor cantidad de observaciones del ave Passer domesticus, lo que podría estar relacionado con factores ambientales o geográficos específicos de esas regiones. Es importante considerar estos patrones y posibles sesgos en los datos al interpretar los resultados del modelo.

- Dependiendo de los objetivos específicos del proyecto y los costos asociados a los diferentes tipos de errores de clasificación, puede ser necesario explorar métricas de evaluación alternativas además de la precisión.