Universidad del Rosario

Analisis Avanzado de Datos

W4. Evaluación Cruzada, boostrap

FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

Profesor

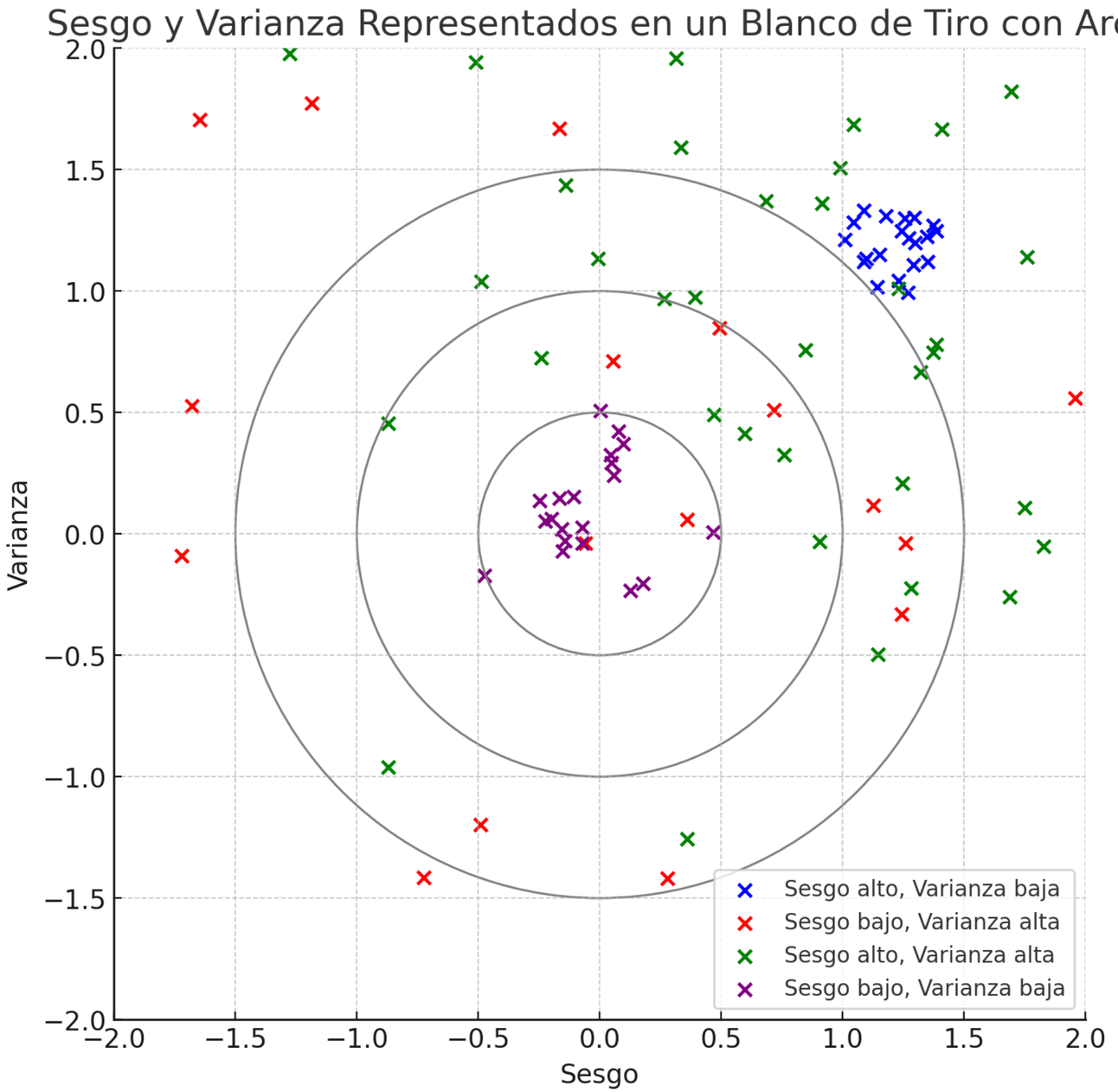
FERNEY ALBERTO BELTRAN MOLINA
ferney.beltran@urosario.edu.co

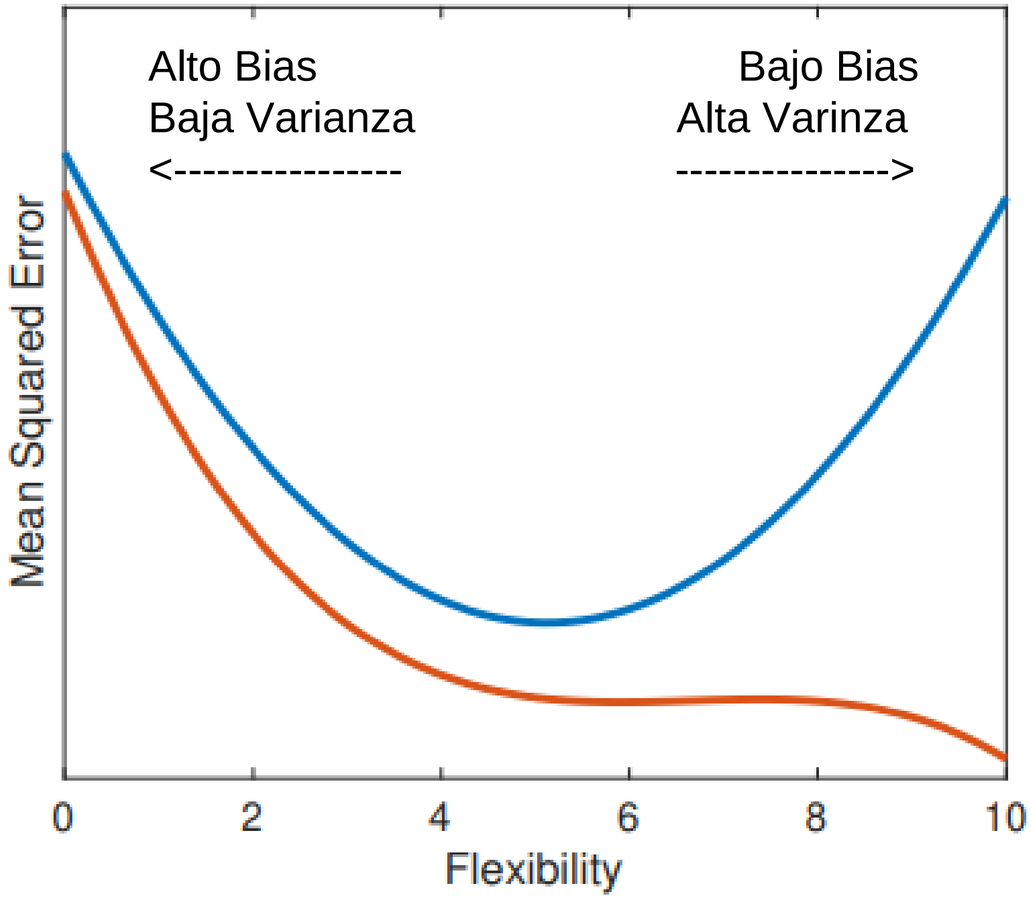
Ingeniero Electrónico.
Magister en TIC
Candidato Doctor en TIC

Director del Centro de investigación e innovación CEINTECCI.
Miembro de la junta directiva Avanciencia

Procesamiento y análisis de datos basadas en IA.
Simulación y modelado por computación,
Optimizan Sistemas de procesamiento en hardware y software
Diseño de sistemas electrónicos reconfigurables

Varianza /Sesgo (bias) tradeoff





Selección de Modelos

La elegir el modelo más apropiado (en términos de complejidad) que minimice el error de predicción en despliegue.

Prueba de diferentes modelos: A veces, la mejor manera de seleccionar un modelo es probar varios y ver cuál funciona mejor en un conjunto de validación

Técnicas de regularización: Como la regresión Lasso o Ridge, que pueden ayudar a evitar el sobreajuste añadiendo una penalización a los coeficientes del modelo.

Validación cruzada (Cross-validation): Se divide el conjunto de datos en varios subconjuntos y se entrena el modelo en algunos mientras se valida en otros. Esto proporciona una estimación más robusta del rendimiento del modelo en datos no vistos.

Mínimos cuadrados a menudo tienen un sesgo bajo pero una gran varianza

Puede mejorar la predicción al reducir o establecer algunos coeficientes en cero.

Sacrificamos un poco de sesgo para reducir la varianza de los valores predichos, y por lo tanto, podemos mejorar la precisión

Discusión

Tu empresa comenzará a producir vehículos autónomos y necesita un algoritmo de visión por computadora para detectar peatones.

En lugar de invertir en crear un equipo centrado en visión por computadora, tu empresa decide externalizar es el proceso y comprar una solución externa.

¿Cuál sería un buen enfoque para identificar la mejor solución?

Discusión

Tu empresa es una desarrolladora líder de algoritmos de visión por computadora.

Has descubierto que una empresa de coches autónomos está buscando un algoritmo de visión por computadora para detectar peatones y ha puesto a disposición un conjunto de datos para que cualquiera elabore una solución.

¿Qué harías con este conjunto de datos?

Validación de Modelos

Utilizamos conjuntos de datos para diferentes propósitos, por ejemplo, para evaluar el rendimiento del despliegue de un modelo final (**test dataset**) o para identificar el mejor modelo dentro de una familia de modelos (**training dataset**).

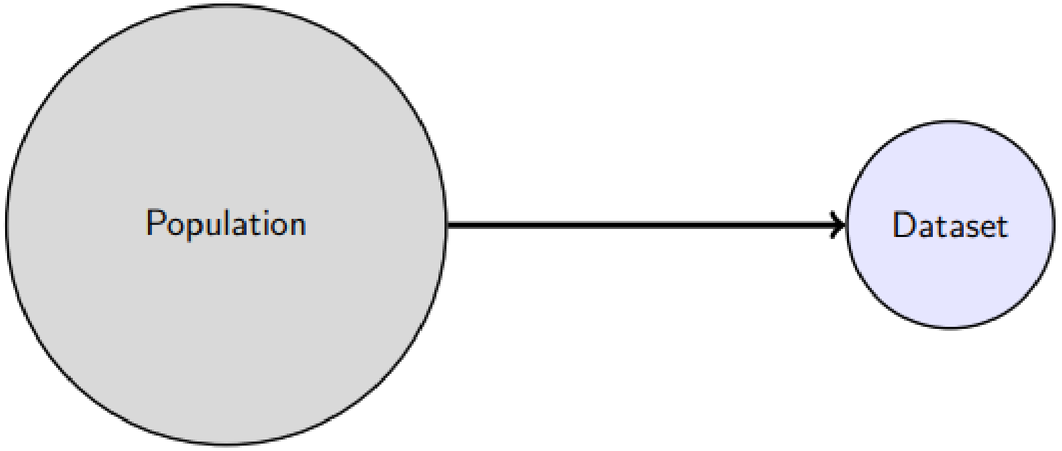
A menudo, necesitamos explorar diferentes opciones antes de entrenar un modelo final.
Por ejemplo, considera la regresión polinómica. El grado polinómico ***D*** es un **hiperparámetro**, ya que para cada valor de ***D*** se obtiene una familia diferente de modelos.

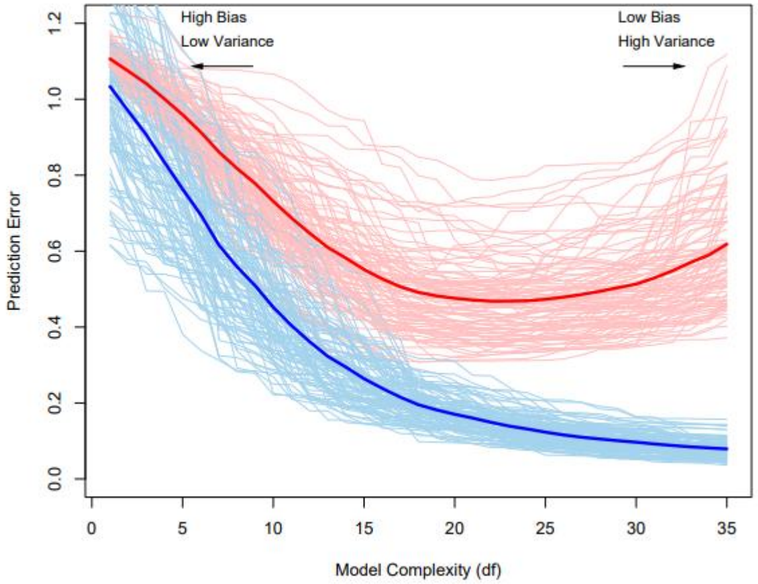
¿Cómo podemos seleccionar el valor correcto de D?

Otros hiperparámetros que podríamos necesitar evaluar antes del entrenamiento incluyen
λ en regularización ejemplo Ridge y lasso

Validación

Los métodos de validación nos permiten usar datos para evaluar y seleccionar modelos antes de entrenar el definitivo. Los mismos datos utilizados para la validación pueden ser usados luego para entrenar el modelo final

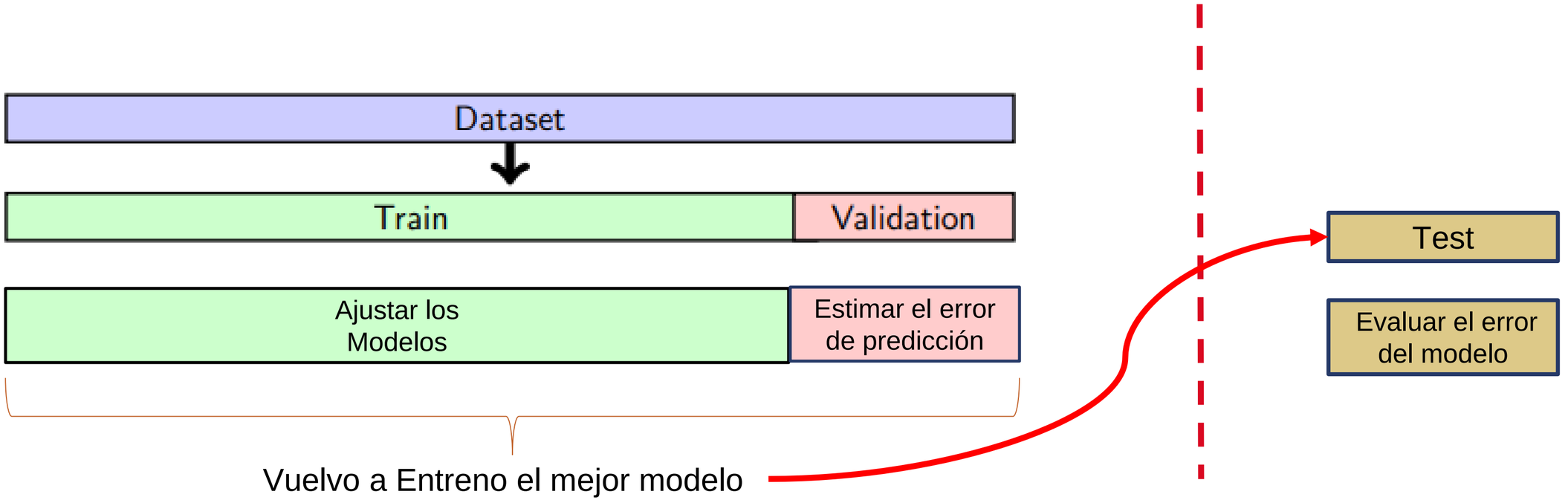




La validación implica una o más rondas de **entrenamiento** y **estimación de rendimiento** por familia de modelos, seguido de un promedio de rendimiento

Selección de Modelos

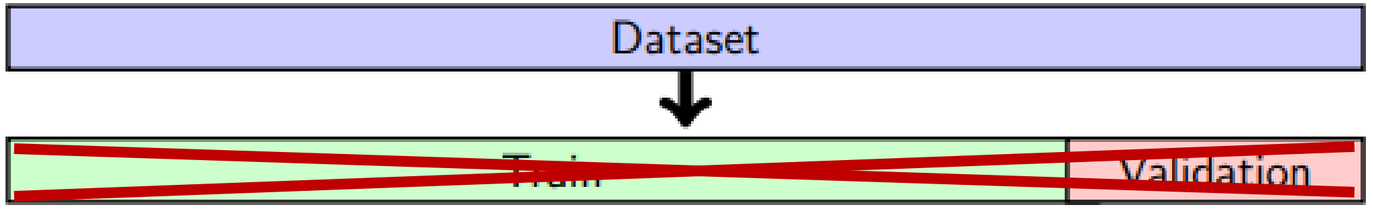
Si nos encontramos en una situación con abundancia de datos, el mejor enfoque para ambos problemas es dividir aleatoriamente el conjunto de datos en tres partes: un conjunto de entrenamiento, un conjunto de validación y un conjunto de prueba



Selección de Modelos

La cantidad de datos es limitada, lo que hace inviable la típica división

¿Qué hacemos ?



Maximizar el uso de los datos disponibles mediante un reusó eficiente de las muestras.
El objetivo es proporcionar una medida confiable del rendimiento del modelo en datos no vistos

Validación cruzada y **Bootstrap**

Leave-one-out cross-validation (LOOCV)

Este método también divide el conjunto de datos disponible en conjuntos de entrenamiento y validación. Sin embargo, **el conjunto de validación contiene solo una muestra**

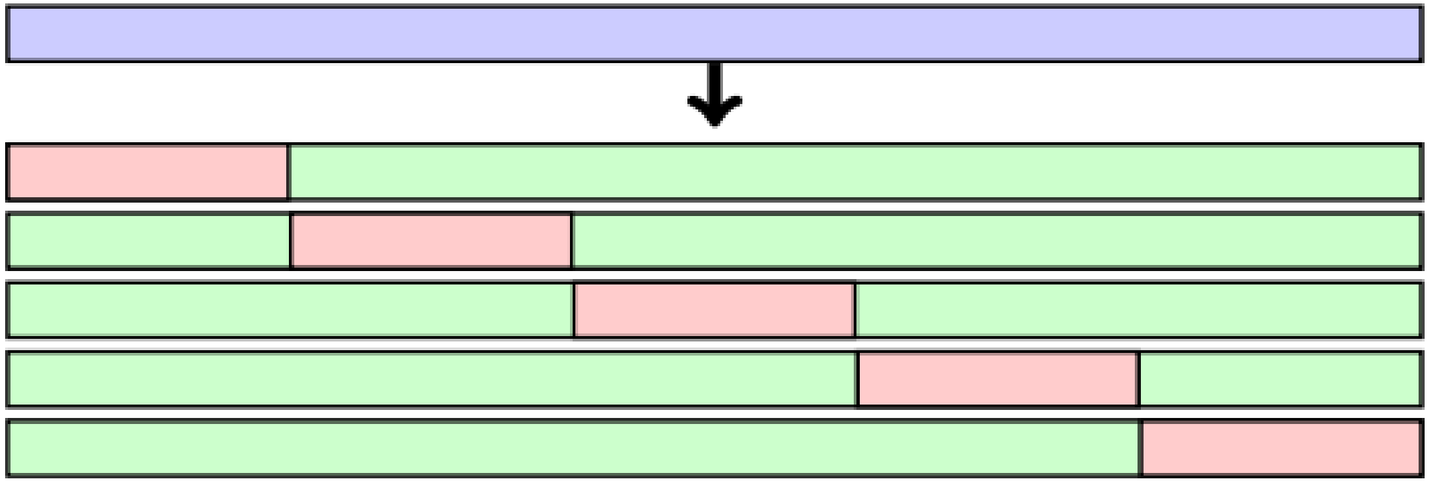
Se consideran **múltiples divisiones** y el rendimiento final se calcula como el promedio de los rendimientos individuales

Para N muestras, producimos N divisiones y obtenemos N rendimientos diferentes.

k-fold cross-validation

En este enfoque, el **conjunto de datos disponible se divide en k grupos** (también conocidos como pliegues o particiones) de tamaño aproximadamente igual:

- Realizamos **k rondas de entrenamiento seguidas de validación**, cada una usando un pliegue diferente para la validación y el resto para el entrenamiento.
- La estimación final del rendimiento es el promedio de los rendimientos de cada ronda



LOOCV es un caso especial de k-fold donde k=N.

k-fold cross-validation

- **El enfoque del conjunto de validación implica una ronda de entrenamiento**. Sin embargo, los modelos se entrenan con menos muestras y el rendimiento final es altamente variable debido a la división aleatoria.

- **La validación cruzada de dejar uno fuera** (LOOCV) requiere tantas rondas de entrenamiento como muestras hay en el conjunto de datos, sin embargo, en cada ronda casi todas las muestras se utilizan para entrenamiento. Siempre proporciona la misma estimación de rendimiento.

- **El enfoque de k particiones** (k-fold) es el más popular. Implica menos rondas de entrenamiento que LOOCV. En comparación con el enfoque del conjunto de validación, la estimación del rendimiento es menos variable y se utilizan más muestras para el entrenamiento.

Ejercicio académico y ético

Conjunto de Datos de Precios de Viviendas de Boston:
Contiene información sobre diferentes características de viviendas en Boston y su precio medio.

1. **CRIM**: Tasa de criminalidad per cápita por ciudad.
2. **ZN**: Proporción de tierra residencial dividida en zonas para lotes de más de 25,000 pies cuadrados.
3. **INDUS**: Proporción de acres de negocios no minoristas por ciudad.
4. **CHAS**: Variable ficticia Charles River. CHAS = 1 if traza limita con el río else 0
5. **NOX**: Concentración de óxidos nítricos (partes por 10 millones).
6. **RM**: Número promedio de habitaciones por vivienda.
7. **AGE**: Proporción de unidades ocupadas por sus propietarios construidas antes de 1940.
8. **DIS**: Distancias ponderadas a cinco centros de empleo en Boston.
9. **RAD**: Índice de accesibilidad a carreteras radiales.
10. **TAX**: Tasa de impuesto a la propiedad de valor total por $10,000.
11. **PTRATIO**: Proporción alumno-profesor por ciudad.
12. **B**: 1000*(Bk - 0.63)*2 donde Bk es la proporción de personas de origen afroamericano por ciudad.
13. **LSTAT**: Porcentaje de población de menor estatus.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.datasets import load_boston

# import el dataset
boston_data = load_boston()

print(boston_data.DESCR)
```

Ejercicio académico y ético

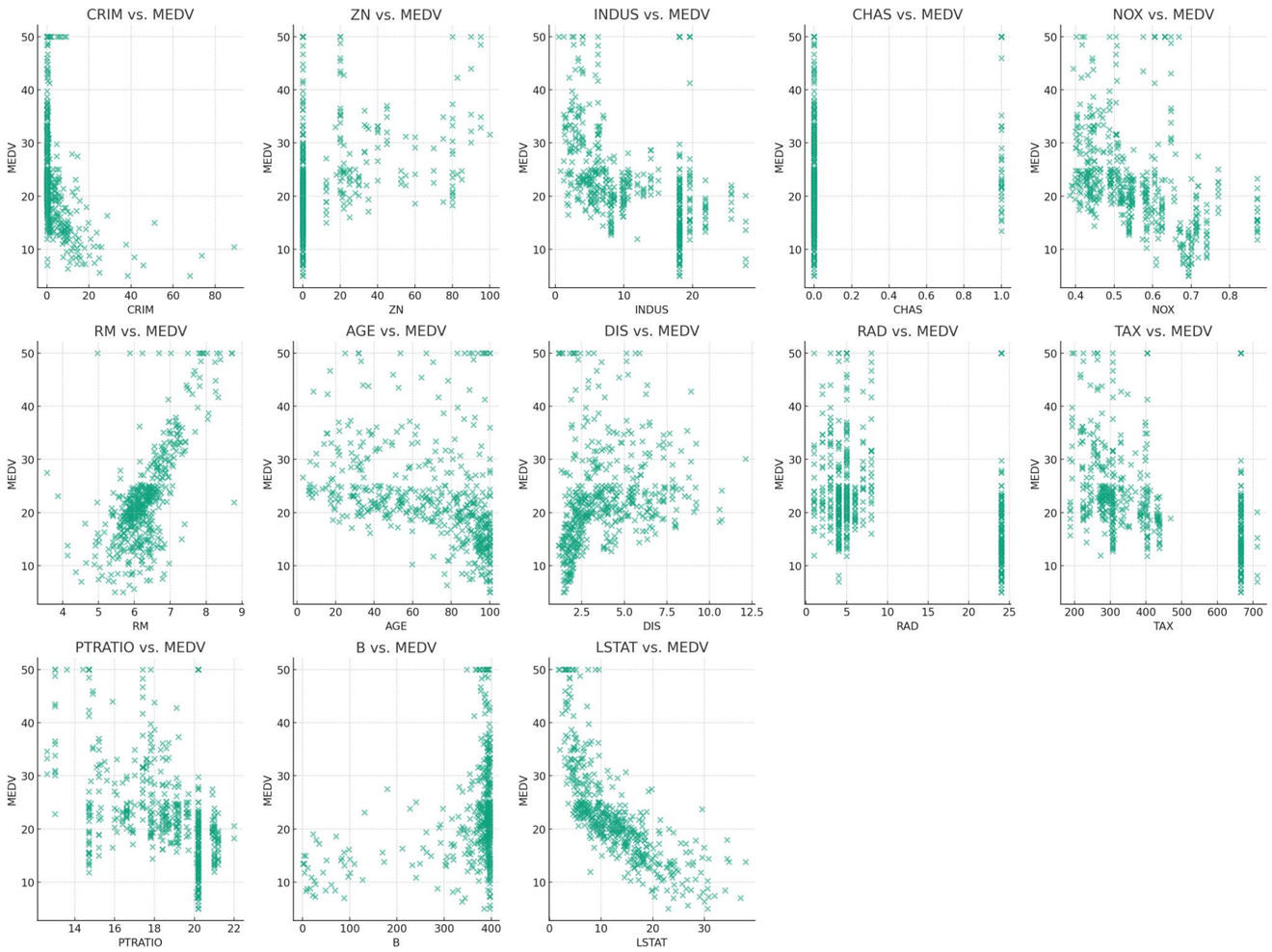
Realiza las técnicas de validación

Genera un conjunto de modelos de regresión Lineal y polinomio y entrega el mejor modelo usando validación k-fold y LOOCV

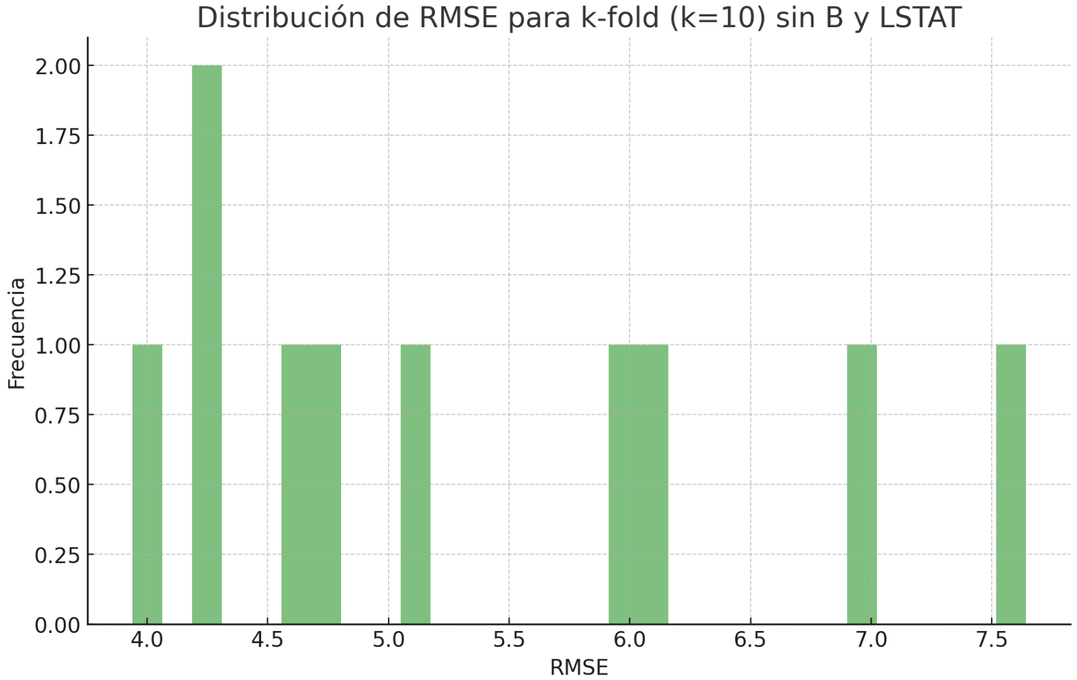
Que características seleccionas para realizar la predicción

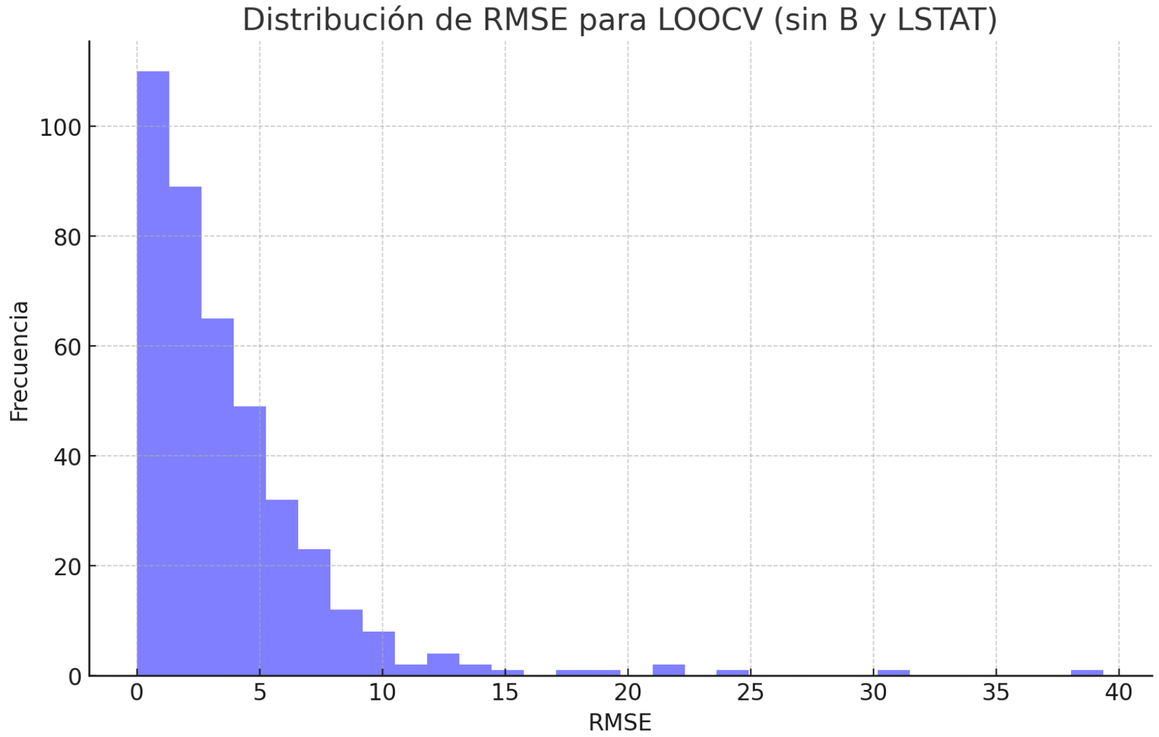
Para la medida de calidad usa RMSE

Puedes usar los datos dados en el colab del 1 laboratorio o usar las librerias sklearn



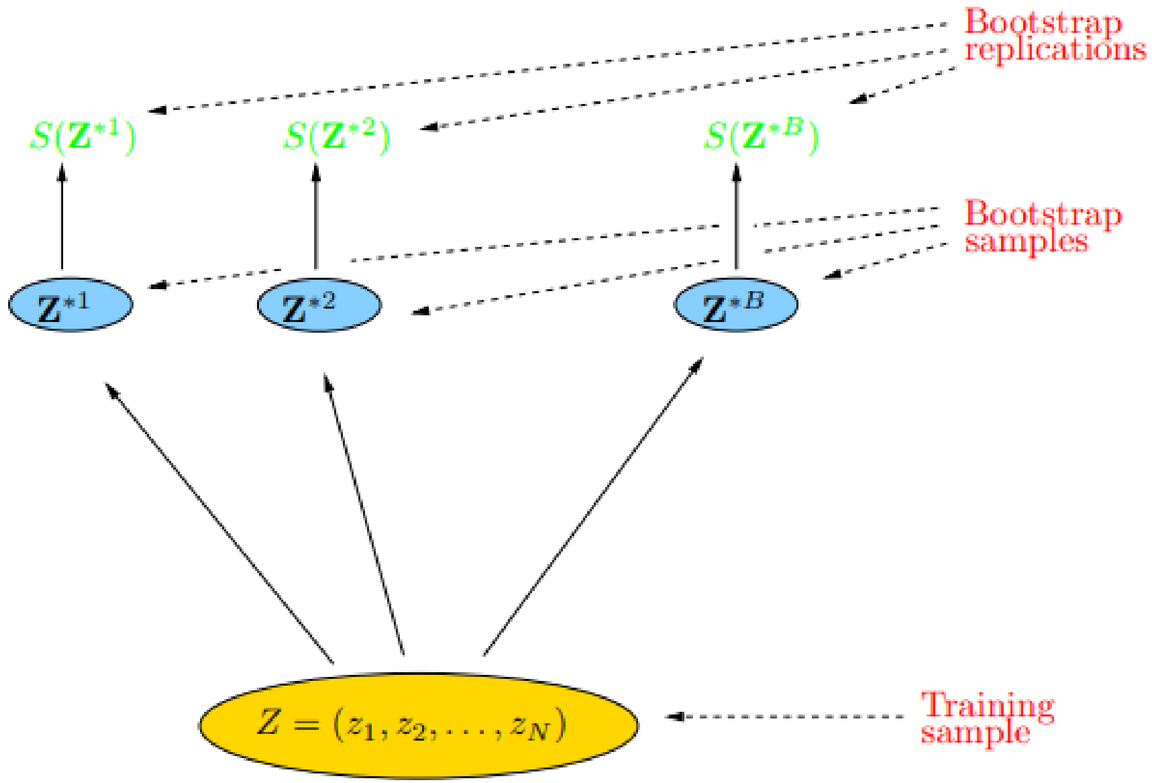
Que esperamos





Boostrap

El método de **bootstrap** implica tomar **muestra de reemplazo** repetidamente del conjunto de datos y recalculando el estadístico o modelo de interés en cada muestra. Estas re-muestras son denominadas **muestras bootstrap**



Validación Boostrap (Proceso)

- **Muestra Bootstrap:** Del conjunto de datos original de tamaño ***n***, selecciona aleatoriamente ***n*** observaciones con reemplazo para formar una **muestra bootstrap**.
- **Entrenar el modelo**: Entrena el modelo usando la muestra bootstrap.
- **Evalúa el modelo**: En todo el conjunto de datos original implementa el modelo y registra el error o métrica de interés (por ejemplo, RMSE en regresión, precisión en clasificación).
- **Repetir**: los pasos 1-3 muchas veces (por ejemplo, 1,000 o 10,000 veces) para obtener una distribución de la métrica de interés.
- **Interpreta los resultados:** Usa la distribución de errores para interpretar la robustez y generalización del modelo. Por ejemplo, puedes obtener un intervalo de confianza para tu métrica de interés.

Ventajas

- **No desperdicia datos:** A diferencia de la validación cruzada, donde una parte de los datos se deja fuera en cada iteración, la validación bootstrap utiliza todo el conjunto de datos para entrenar y evaluar en cada iteración.
- **Estimaciones más estables**: Dado que se basa en el remuestreo repetido, las estimaciones del error de validación tienden a ser más estables en comparación con un único conjunto de entrenamiento/prueba.
- **Flexibilidad:** Puede adaptarse fácilmente para evaluar modelos complejos o situaciones donde la validación cruzada es difícil de aplicar.

Ejercicio

Eres un analista especializado en el mercado inmobiliario. Con el conjunto de datos proporcionado anteriormente, tu tarea es predecir el valor de las viviendas utilizando solamente dos características. Para validar la precisión y robustez de tu modelo, emplearás la técnica de bootstrap. Para llevar a cabo este procedimiento en Python, debes hacer uso de la función resample que se encuentra en el módulo scikit-learn.utils.

```
pseudo codigo
1 Selecciona una muestra con reemplazo
    X_sample, y_sample = resample(X_selected, y, n_samples=bootstrap_size)

2 Entrena el modelo
3 Hace predicciones sobre todo el conjunto de datos
4 Calcula el error cuadrático medio y guardalo
    mse = mean_squared_error(y, y_pred)
    mse_values.append(mse)

4 Analizar y visualizar la distribución de mse_values para obtener insights sobre la
variabilidad y precisión del modelo.
```

Puntos importantes

- :
- No tenemos una descripción ideal de la población, todo lo que podemos hacer es **extraer datos**.
- Los datos se pueden utilizar para **probar** un modelo final, **entrenar** modelos, **evaluar**
- y **seleccionar** modelos.
- Los conjuntos de datos deben ser **representativos** de la población objetivo. Este reduce el riesgo de sobreajuste y conduce a mejores estimaciones del rendimiento en la implementación.
- Se debe **diseñar una estrategia de prueba** antes del entrenamiento. Además, evite
- observar el conjunto de datos de prueba durante el entrenamiento y probar un modelo final sólo una vez
- La calidad de un modelo final depende del **tipo** de modelo, la etapa de **optimización** y los **datos** de entrenamiento.
- Una estimación del rendimiento es una **cantidad aleatoria**.