



Módulo 6

Sesión N° 2



ACTIVIDAD:



¿Tú modelo generaliza? Aplicando validaciones cruzadas al dataset de diabetes

Objetivo: Aplicar diferentes técnicas de validación para medir el desempeño de un modelo de regresión lineal. Comparar los resultados y reflexionar sobre el impacto que tienen las decisiones de evaluación en la generalización del modelo.



Contexto

- Utilizar el dataset embebido `load_diabetes()` de `sklearn.datasets`.
- Entrenar un modelo de regresión lineal (`LinearRegression`).
- Evaluar el modelo utilizando:
 - Holdout
 - K-Fold ($k=5$)
 - Leave-One-Out (LOOCV)
- Comparar los resultados de cada técnica utilizando la métrica R^2 .
- Responder preguntas de reflexión sobre las diferencias observadas.



Tiempo estimado: 45 minutos

Formato de ejecución: grupal





Requerimientos:

Estás desarrollando un modelo para predecir progresión de una enfermedad crónica utilizando datos médicos (como presión arterial, edad, índice de masa corporal, etc.). Antes de confiar en los resultados, necesitas evaluar correctamente si el modelo generaliza bien.

A continuación, se presentan las celdas que debes ejecutar. Recuerda observar los resultados y tomar nota para las preguntas de reflexión.

CELDA 1 – IMPORTACIÓN DE LIBRERÍAS Y DATOS

```
from sklearn.model_selection import train_test_split, KFold, cross_val_score, LeaveOneOut
from sklearn.linear_model import LinearRegression
from sklearn.datasets import load_diabetes

X, y = load_diabetes(return_X_y=True)
modelo = LinearRegression()
```

CELDA 2 – VALIDACIÓN HOLDOUT

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
modelo.fit(X_train, y_train)
score_holdout = modelo.score(X_test, y_test)
print(f"R² Holdout: {score_holdout:.3f}")
```

CELDA 3 – VALIDACIÓN K-FOLD

```
kf = KFold(n_splits=5, shuffle=True, random_state=1)
scores_kfold = cross_val_score(modelo, X, y, cv=kf, scoring='r2')
print(f"Resultados K-Fold: {scores_kfold}")
print(f"Promedio R² (K-Fold): {scores_kfold.mean():.3f}")
```

CELDA 4 – VALIDACIÓN LEAVE-ONE-OUT (LOOCV)

```
from sklearn.datasets import make_regression
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import LeaveOneOut, cross_val_score

# Generar un dataset pequeño con una sola característica
X, y = make_regression(n_samples=20, n_features=1, noise=10, random_state=0)

# Crear el modelo
modelo = LinearRegression()

# Crear esquema LOOCV
loocv = LeaveOneOut()

# Evaluar con MSE (negativo)
scores = cross_val_score(modelo, X, y, cv=loocv, scoring='neg_mean_squared_error')

# Mostrar promedio del error (recordando que sklearn devuelve negativo)
print(f"Promedio del error (LOOCV - MSE): {-scores.mean():.2f}")
```

1. ¿Cuál técnica entregó el valor de R^2 más alto? ¿Y cuál el más bajo?
2. ¿Cuál técnica crees que refleja mejor la capacidad real de generalización del modelo?
3. ¿Por qué crees que LOOCV es más costosa en tiempo de cómputo?
4. ¿En qué caso elegirías Holdout sobre K-Fold, o viceversa?
5. ¿Qué aprendiste sobre la importancia de evaluar modelos con distintos enfoques?

