Nonlinear Econometrics (Maching Learning)

Winter School

Research Institute in Economics and Finance

Prof.: Cesar Ramos
TA: Luis Flores

October 2025

Contexto: ¿Por qué Machine Learning en economía?

Predicción de ciclos, inflación, crisis

- Los modelos de regresión de ML se utilizan para predecir variables objetivo en una escala continua (3).
- La regresión se utiliza en análisis de series temporales y muestras
- El ML es útil para manejar datos financieros y en tareas de previsión

Capacidad de manejar relaciones no lineales

- Los modelos lineales están sujetos a subajuste (*underfitting*) porque la realidad a menudo es más compleja que el modelo (1).
- Los Bosques Aleatorios (Random Forests) son un poderoso método no paramétrico capaz de modelar relaciones complejas y no lineales en los datos.

Diferencia entre modelos econométricos y ML

- El ML trabaja con modelos complejos con muchos grados de variabilidad (2).
- ML es una alternativa cuando las relaciones lineales fallan al modelar los datos (3).

Árboles de Decisión CAST: Concepto

¿Qué es un árbol de decisión?

- Modelo que divide los datos en ramas según preguntas sobre las variables.
- Cada nodo representa una condición sobre una variable.
- Cada hoja contiene la predicción final (clase o valor continuo).

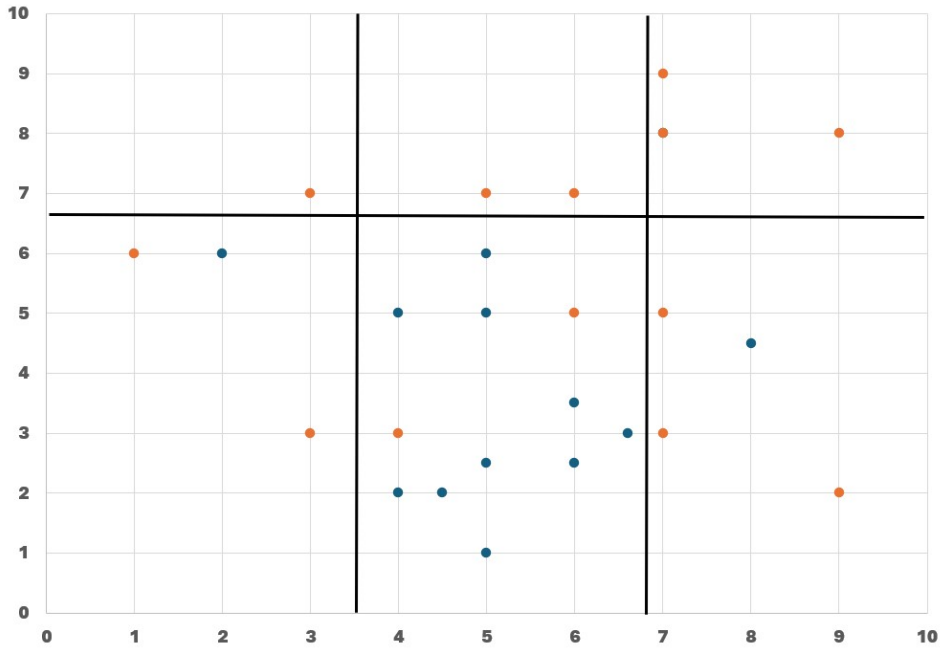
Clasificación vs. regresión

- Clasificación: resultado categórico (ej. recesión, expansión).
- Regresión: resultado numérico (ej. PIB futuro).

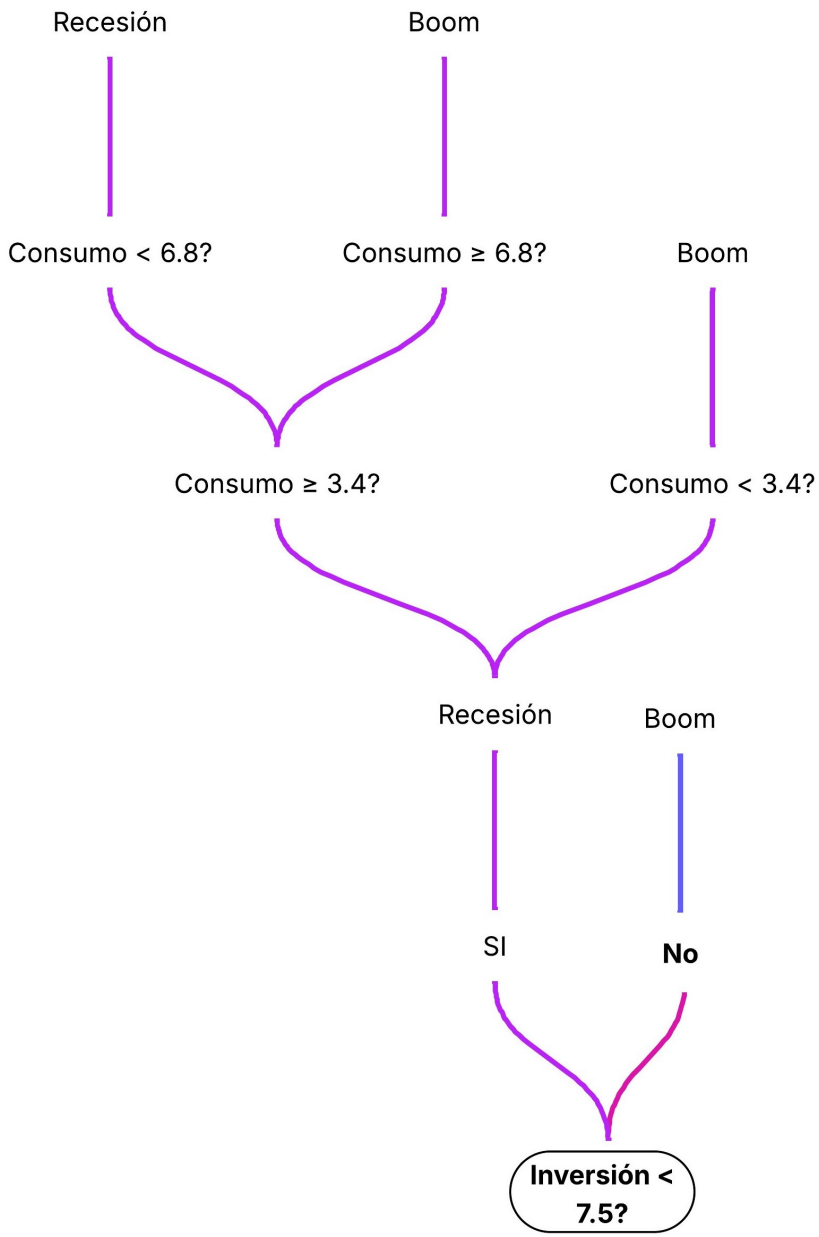
Consideraciones importantes

- El árbol selecciona la variable que mejor separa los datos (criterio de división: Gini o error cuadrático medio).
- Se detiene cuando no mejora la predicción o quedan pocos datos (criterio de parada) .
- Árboles muy profundos tienden a sobreajustar (*overfitting*).

Árbol de Decisión - Clasificación de regímenes económicos



*El árbol de decisión utiliza dos variables —**inversión(eje Y)** y **consumo (eje X)**— para clasificar la economía en **auge** o **recesión**. Los puntos naranjas indican **auge** y los azules muestran **recesión**.



Criterios de división y limitaciones del Árbol de Decisión

Objetivo: Evaluar qué tan buena es una partición (umbral) para separar los datos.

1. Criterio de impureza para clasificación: Índice de Gini

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2$$

- p_k: proporción de observaciones de la clase k en el nodo t.
- $G(t) = 0$ cuando el nodo es puro (todas las observaciones son de la misma clase).

2. Criterio para regresión: Error Cuadrático Medio (MSE)

$$\text{MSE}(t) = \frac{1}{n_t} \sum_{i \in t} (y_i - \bar{y}_t)^2$$

- $\bar{y}_t$: media de las predicciones en el nodo t.
- El umbral se elige para minimizar el MSE promedio ponderado de los hijos.

Problemas del Árbol de Decisión

- **Sobreajuste (Overfitting):** los árboles tienden a aprender ruido si crecen sin límite.
- **Alta varianza:** pequeños cambios en los datos pueden generar un árbol totalmente distinto.
- **Bajo sesgo (Bias) pero varianza alta:** modelo flexible pero poco estable.
- **Solución:** métodos de conjunto como **Random Forest** reducen la varianza promediando múltiples árboles.

¿Qué es Random Forest?

Ensamble de múltiples árboles de decisión

- Un **Bosque Aleatorio** es un conjunto (*ensemble*) de muchos árboles de decisión que trabajan de forma independiente.
- Cada árbol se entrena sobre una **muestra aleatoria** del conjunto de datos original, generada mediante el método de **bootstraping** (muestreo con reemplazo).
- En cada división (nodo), el árbol considera solo un **subconjunto aleatorio de características**. Esto introduce variabilidad y reduce la correlación entre los árboles.
- En tareas de **clasificación**, la predicción final del bosque se obtiene por **votación mayoritaria** entre los árboles. En **regresión**, se calcula el **promedio** de las predicciones individuales.
- Este proceso se conoce como **bagging** (*bootstrap + aggregating*), y su objetivo es **reducir la varianza** sin aumentar significativamente el sesgo.
- En series temporales, el muestreo se adapta mediante el uso de **subseries conjuntas** para preservar la estructura temporal.

Selección de parametros

Número de árboles (`n_estimators`)

- Aumentar árboles reduce la varianza hasta que el error se estabiliza.
- Más árboles → mayor costo computacional.

Número de variables por nodo (`max_features`)

- Controla cuántas características se prueban en cada división.
- Valores bajos → más diversidad; valores altos → menor sesgo.

Profundidad y criterios de parada

- `max_depth`, `min_samples_split`, `min_samples_leaf`.
- Ajustan el balance entre **sesgo y varianza**.

Aplicación: Clasificación de regímenes económicos

Ejemplo: Si queremos evaluar qué variables determinan un **régimen económico** (recesión o auge), podemos incluir:

- PIB total y PIB sectorial (industria, servicios, agricultura).
- Tasas de desempleo general y por sectores.
- Inflación, tasas de interés, producción industrial, consumo privado.

El modelo seleccionará automáticamente las variables más relevantes para clasificar los regímenes económicos.

Ponderación o importancia relativa:

- Cada variable recibe un peso según cuánto contribuye a reducir el **error cuadrático medio (MSE)** o el **índice Gini**.
- Las variables con menor contribución pueden eliminarse sin afectar el desempeño del modelo.

Interpretación: Esta medida de importancia permite identificar qué factores económicos tienen mayor influencia sobre los ciclos de auge o recesión.

References I

[1] A. Géron – *Hands-on machine learning with scikit-learn & tensorflow: Concepts, tools, and techniques to build intelligent systems*, first éd., O'Reilly Media, Inc., 2017.

[2] S. Marsland – *Machine learning: An algorithmic perspective, second edition*, CRC Press, Taylor & Francis Group, 2015.

[3] S. Raschka – *Python machine learning*, Packt Publishing, 2017, Referenciado también en Geron.