

# Aprendizaje estadístico para estudiar volatilidad y carteras

**Desde cero hasta ridge, KNN y gradient boosting**

Cómo entender las familias, justificar una selección y evaluar sus resultados.

Preparado para Matías Vukusic W.  
Edición integral revisada · 23 de octubre de 2026

Fundamentos · Ejemplos resueltos · Métodos ejecutados  
Resultados documentados · Límites · Defensa oral

Objetivo del proyecto: aproximar volatilidad histórica quinquenal a partir de la composición contemporánea agregada de cartera.

## GUÍA DE LECTURA

# Cómo estudiar este manual

Este manual parte desde cero y llega a las decisiones de tu investigación. Su objetivo es que puedas explicar qué aprende cada modelo, por qué lo comparas con otros y qué límites tiene la evidencia. No necesitas memorizar el código para comprenderlo.

## Tres niveles que nunca debes confundir

TEORÍA: propiedades generales de los métodos, sustentadas en ISL, ESL y Bishop. EJEMPLO: cifras ficticias para aprender una operación. EXPERIMENTO: decisiones y resultados reportados en los documentos del 18 y 19 de octubre de 2026. Estos últimos no se recalcularon desde la base original para escribir este manual.

## Orden recomendado

Primera lectura: capítulos 1 a 18, para construir lenguaje y entender las familias. Segunda lectura: capítulos 19 a 29, para reconocer cómo se aplicó ese lenguaje a tus datos. Tercera lectura: gráfico, ejercicios, respuestas y defensa oral. Usa el índice de la página siguiente para volver a cada tema.

## Qué cambia respecto de la edición anterior

Se integra el estado del experimento, se completa Y2, se explica el KNN ponderado, se distingue la configuración real de los ejemplos, se restituyen gráficos y fórmulas legibles y se incorporan las limitaciones del bootstrap y de la robustez. No se conserva como vigente el texto que trataba los conteos o las variables como decisiones aún no tomadas.

## Alcance temporal y probatorio

Edición revisada al 23 de octubre de 2026. La evidencia experimental utilizada llega a los informes del 18 y 19 de octubre; no se afirma una nueva corrida posterior. La auditoría documental motivó correcciones editoriales y metodológicas, pero no cerró los controles pendientes de las bases.

RECUERDA: puedes comprender y justificar una técnica sin afirmar que sus datos están certificados. La calidad del algoritmo, del experimento y de las mediciones son cuestiones diferentes.

# Índice de estudio

Cómo estudiar este manual · 2

1 · Qué son X e Y en tu problema · 5

2 · Qué significa aprender de datos · 6

3 · El mapa de las técnicas · 7

4 · Familias que estiman un número · 8

5 · Las matemáticas mínimas · 9

6 · Por qué acertar al entrenar no basta · 10

7 · Regresión lineal como punto de partida · 11

8 · Qué agrega la regresión ridge · 12

9 · KNN: vecinos y ponderaciones · 13

10 · Cuando la similitud de KNN deja de ayudar · 14

11 · Árboles y ensambles · 15

12 · Gradient boosting paso a paso · 16

13 · Por qué comparar estos tres · 17

14 · Qué dejo fuera y por qué · 18

15 · Alternativas viables que no debo descalificar · 19

16 · Qué implica trabajar con carteras agregadas · 20

17 · Evaluación y fuga de información · 21

18 · Cómo medir y elegir al final · 22

19 · Cómo se construye  $Y_1$  · 23

# Índice de estudio

|                                                |
|------------------------------------------------|
| 20 · Y2 y el indicador completo CMF · 24       |
| 21 · Qué mide realmente X · 25                 |
| 22 · De 28 candidatas a 26 entradas · 26       |
| 23 · Del universo a la muestra efectiva · 27   |
| 24 · Cómo se separaron RUN y fechas · 28       |
| 25 · Configuración realmente informada · 29    |
| 26 · Resultados: qué ganó y según qué · 30     |
| 27 · Incertidumbre y estabilidad temporal · 31 |
| 28 · Qué sigue pendiente de validación · 32    |
| 29 · Procedimiento reproducible y cierre · 33  |
| 30 · Ver las tres formas de ajustar · 34       |
| 31 · Ejercicios de fundamentos · 35            |
| 32 · Respuestas de fundamentos · 36            |
| 33 · Ejercicios sobre tu investigación · 37    |
| 34 · Respuestas sobre tu investigación · 38    |
| 35 · Lo que debo explicar de memoria · 39      |
| 36 · Defender la elección en voz alta · 40     |
| 37 · Dónde estudiar los métodos · 41           |
| 38 · Fuentes del proyecto y jerarquía · 42     |
| 39 · Qué quedó corregido y qué no · 43         |

## APRENDIZAJE ESTADÍSTICO

# 1 · Qué son X e Y en tu problema

Una observación es la unidad sobre la que el modelo aprende. En este estudio es un fondo identificado por RUN en un mes. Una posición individual de cartera no es un ejemplo independiente de aprendizaje. Tampoco debes duplicar una cartera por cada serie de cuotas y contar esas copias como fondos distintos.

X es el conjunto de características de la cartera: proporciones por clase, exposición extranjera observada, concentración y otros atributos agregados. Y es la respuesta que se aproxima. La principal es Y1, volatilidad retrospectiva anualizada calculada sobre 260 retornos semanales consecutivos. Una Y1 igual a 8 significa volatilidad anualizada de 8 %, no rentabilidad de 8 % ni probabilidad de pérdida de 8 %.

## De ejemplos conocidos a una regla

En entrenamiento se observan X e Y. El algoritmo aprende una regla que recibe X y devuelve una estimación de Y. En evaluación se oculta Y al procedimiento para comparar después la estimación con el valor observado. La respuesta puede referirse al pasado: estimarla no convierte el problema en pronóstico futuro.

## Estado actual del proyecto

El informe de cierre documenta 63.633 fondos-mes elegibles inicialmente; 31.973 se usaron en TRAIN, 659 en VALID y 1.761 en TEST. Otras 29.240 filas quedaron en reserva. Estas cifras no son millones de ejemplos ni 63.633 fondos independientes. [CIERRE, §§5-6]

El cálculo de Y nace al nivel RUN-serie-moneda-mes y se reduce a RUN-mes con una regla de continuidad observable hasta el corte. X y la Y elegida se empalman después. La moneda y los factores conservan limitaciones de validación económica. [PROY-Y, §§2-4]

RECUERDA: X describe la cartera observable; Y1 resume su historia de retornos; la fila analítica es fondo-mes. Y2 es secundaria y el indicador completo CMF no fue estimado.

## APRENDIZAJE ESTADÍSTICO

## 2 · Qué significa aprender de datos

Aprender significa ajustar una regla utilizando ejemplos. Supón que no conoces la volatilidad de una cartera con 40 % de acciones. Puedes asignarle a todos los fondos el mismo valor, usar una recta que relacione exposición accionaria y volatilidad o buscar fondos parecidos. Todas son reglas de estimación; difieren en cómo usan la información.

La notación habitual es  $Y = f(X) + \varepsilon$ .  $f$  representa la relación y  $\varepsilon$  lo no explicado. El modelo aprende una aproximación de  $f$  y devuelve  $\hat{Y} = f_{\text{estimada}}(X)$ . El sombrero de  $\hat{Y}$  significa «estimado». [ISL, §2.1]

En tu caso, la parte no explicada puede incluir exposiciones subyacentes omitidas, cambios históricos de estrategia y errores o limitaciones de medición. No todo residuo es azar puro. Ampliar o mejorar  $X$  puede reducir parte del error que parecía inevitable con una representación más pobre.

### Cuatro conceptos distintos

Algoritmo: procedimiento para aprender una regla. Modelo ajustado: regla concreta obtenida con ciertos datos. Parámetros: cantidades aprendidas, como los coeficientes de una recta. Hiperparámetros: elecciones que regulan el aprendizaje, como el número de vecinos, la penalización ridge o la profundidad de los árboles.

Elegir hiperparámetros también utiliza evidencia. Por eso no puedes probar muchas configuraciones sobre el examen final y después presentar el mejor resultado como si el examen hubiera permanecido intacto.

### Predicción no siempre significa futuro

En aprendizaje supervisado, predecir puede significar estimar una respuesta que se mantiene oculta, aunque corresponda al pasado. Tu modelo puede intentar reconstruir  $Y_1$  histórica desde  $X$  contemporánea. Pronosticar volatilidad posterior requiere definir otra  $Y$ , observada después de la fecha de  $X$ , y evaluar ese objetivo específicamente.

### Aprendizaje estadístico y machine learning

Son tradiciones muy superpuestas. La regresión lineal también es aprendizaje supervisado. No hay una frontera que convierta automáticamente los modelos antiguos en «solo estadística» y los recientes en «verdadera inteligencia artificial». El objeto común es aprender relaciones a partir de datos y evaluar su generalización. [ISL, §§2.1.1-2.1.4]

Para recordar. El algoritmo aprende; el modelo ajustado predice; los parámetros se estiman; los hiperparámetros regulan el ajuste.

## APRENDIZAJE ESTADÍSTICO

### 3 · El mapa de las técnicas

Las técnicas se ordenan según criterios distintos. Supervisado, regresión, árboles y bayesiano no son cuatro opciones del mismo nivel: responden preguntas diferentes.

Supervisado significa aprender con parejas X-Y. Las etiquetas pueden calcularse: no necesitan ser anotadas manualmente. No supervisado significa explorar X sin una respuesta: clustering agrupa observaciones y PCA resume dimensiones. Semisupervisado combina ejemplos con y sin respuesta, bajo supuestos adicionales. Por refuerzo se aprenden acciones secuenciales con recompensas; aquí no se está aprendiendo una política de inversión. [ISL, §2.1.4]

Regresión estima una respuesta numérica. Clasificación estima categorías o probabilidades de pertenencia. Si se transformara Y1 en «bajo, medio, alto», harían falta cortes justificados y una evaluación distinta, posiblemente ordinal. [ISL, §2.1.5]

**RECUERDA:** la tarea actual es supervisada, de regresión y tabular. Esta clasificación precede a elegir ridge, KNN o boosting.

| Criterio       | Alternativas                                   | En el estudio                           |
|----------------|------------------------------------------------|-----------------------------------------|
| Información    | Supervisado / no supervisado / semisupervisado | Parejas X-Y                             |
| Respuesta      | Regresión / clasificación                      | Y1 continua                             |
| Representación | Lineal / vecinos / árboles / kernels / redes   | Tres familias                           |
| Combinación    | Modelo individual / ensamble                   | Boosting es ensamble                    |
| Incertidumbre  | Formulaciones probabilísticas y otras          | No la determina el nombre del algoritmo |

## APRENDIZAJE ESTADÍSTICO

## 4 · Familias que estiman un número

Regresión nombra la tarea; no exige una recta. KNN y boosting hacen regresión cuando su respuesta es cuantitativa. Muchas familias también tienen versiones clasificadoras. [ISL, §§2.1.5, 3.5 y 8.1]

Un modelo paramétrico fija una forma con una cantidad de parámetros dada la representación elegida. KNN se llama no paramétrico porque utiliza los ejemplos y su proximidad sin imponer esa fórmula global; esto no significa que carezca de supuestos.

Una regresión que incluye  $x^2$  o una interacción  $x_1x_2$  sigue siendo lineal en los coeficientes. Por eso ridge con atributos originales y ridge con términos polinómicos son procedimientos distintos. Las transformaciones también deben validarse.

Bayesiano describe una forma de expresar supuestos previos y actualizar incertidumbre; puede aplicarse a varias familias. No equivale a elegir naive Bayes ni define por sí solo la tarea.

| Familia       | Cómo estima                  | Ejemplos                |
|---------------|------------------------------|-------------------------|
| Lineal        | Suma ponderada de atributos  | OLS, ridge, lasso       |
| Local         | Respuestas de casos próximos | KNN, suavizadores       |
| Árboles       | Reglas y regiones            | Árbol de regresión      |
| Ensamblados   | Combina muchos modelos       | Random forest, boosting |
| Aditiva suave | Suma curvas por variable     | GAM                     |
| Kernels       | Similitud flexible           | SVR, kernel ridge       |
| Redes         | Transformaciones aprendidas  | Perceptrón multicapa    |

## APRENDIZAJE ESTADÍSTICO

## 5 · Las matemáticas mínimas

La media es el promedio. Si tres valores son 4, 7 y 10, su media es 7. La varianza mide dispersión mediante diferencias cuadradas respecto de la media. La desviación estándar es la raíz de la varianza y vuelve a la unidad original.

Para esos tres valores, las desviaciones son  $-3$ ,  $0$  y  $3$ . Sus cuadrados suman 18. La varianza muestral divide por  $n-1$ :  $18/2 = 9$ . La desviación estándar muestral es 3. No necesitas memorizar una demostración para entender que valores más separados suelen producir mayor dispersión.

### Volatilidad y anualización

La volatilidad de tu  $Y_1$  es una desviación estándar de retornos, multiplicada por  $\sqrt{52}$  y por 100. Si la desviación semanal es 0,01 en unidades decimales,  $Y_1 \approx 7,21\%$ . Es una convención de anualización; no vuelve independiente una serie con dependencia temporal ni garantiza una distribución normal. [PROY-Y, §2]

### Correlación y causalidad

La correlación describe asociación. Si los fondos con más acciones tienen mayor volatilidad, eso no demuestra cuánto cambiaría  $Y$  al obligar a un fondo particular a comprar acciones. Exposiciones omitidas, estilos y composición restante pueden explicar parte de la asociación. Un buen modelo predictivo no es automáticamente un modelo causal.

### Residuos y pérdida

El residuo es observado menos estimado:  $e = y - \hat{y}$ . Si  $Y$  es 12 y el modelo predice 9, el residuo es 3 puntos porcentuales. La pérdida asigna un costo al error. La pérdida cuadrática es  $e^2$ : un error de 3 cuesta 9 y uno de 6 cuesta 36. Al duplicar el error, el costo se cuadruplica.

Esto explica por qué unos pocos casos extremos pueden dominar el ajuste por mínimos cuadrados. También distingue dos decisiones: entrenar con una pérdida y evaluar con una métrica. Puedes entrenar minimizando error cuadrático y reportar tanto MAE como RMSE, siempre que el criterio de selección quede declarado. [ISL, §2.2.1]

Para recordar. La volatilidad mide dispersión de retornos; el error de un modelo mide distancia entre su estimación y la  $Y$  observada. Son dispersiones de objetos distintos.

Comprueba. Si predices 10 % y la volatilidad observada es 12 %, ¿tu error absoluto es 2 % relativo o 2 puntos porcentuales? Es lo segundo.

## APRENDIZAJE ESTADÍSTICO

## 6 · Por qué acertar al entrenar no basta

Un modelo muy flexible puede aprender particularidades de la muestra que no se repiten. Eso es sobreajuste. Uno demasiado rígido puede pasar por alto una relación importante: es subajuste. El objetivo es aproximar bien observaciones nuevas, no maximizar la fidelidad a las filas utilizadas para ajustar. [ISL, §2.2]

Sesgo. Error sistemático asociado a una representación limitada. Una recta puede fallar si la relación real tiene una curva marcada. Varianza del estimador. Cuánto cambiaría el modelo si lo entrenaras con otra muestra comparable. No es la volatilidad financiera que forma tu Y. Ruido o variabilidad no explicada. Diferencias que no puedes resolver con la información y representación disponibles.

La descomposición habitual del error cuadrático esperado reúne sesgo al cuadrado, varianza y error irreducible bajo sus supuestos. Es un marco para entender el equilibrio, no una identidad que puedas estimar automáticamente mirando un único resultado. [ISL, §2.2.2]

### Tres controles de flexibilidad

En KNN, muy pocos vecinos hacen depender la estimación de pocos casos; más vecinos suavizan, pero pueden mezclar carteras diferentes. En ridge, una penalización fuerte restringe el tamaño conjunto de los coeficientes. En boosting, profundidad, tasa de aprendizaje y número de árboles regulan cuánto puede adaptar el modelo sus reglas.

### Tres conjuntos con papeles distintos

Entrenamiento sirve para aprender parámetros. Validación sirve para escoger configuraciones y comparar candidatas. Prueba final sirve para estimar desempeño una vez tomadas esas decisiones. Si cambias repetidamente el modelo después de mirar la prueba, esta ya funciona como validación.

En tu panel, las divisiones deben respetar meses completos y orden temporal. No basta ordenar filas y cortarlas en tres partes: podrías dividir el mismo mes entre conjuntos o evaluar con información que no estaría disponible al aplicar el modelo. [PROY-Y, §4]

Para recordar. Regularizar es limitar flexibilidad para mejorar generalización. No garantiza mejora en todos los datos; esa mejora se comprueba fuera del entrenamiento.

Comprueba. Si KNN con un vecino acierta exactamente cada fila de entrenamiento, ¿eso acredita que aproxima bien la volatilidad de fondos distintos? No: puede estar consultando la propia observación.

## APRENDIZAJE ESTADÍSTICO

## 7 · Regresión lineal como punto de partida

Una regresión lineal combina las variables mediante pesos. Con dos características, escribe:

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

$\beta_0$  es el intercepto: el valor estimado cuando las características valen cero. Cada  $\beta$  indica cuánto cambia la estimación por una unidad adicional de la variable correspondiente, manteniendo las otras constantes. Los coeficientes son aprendidos; no son porcentajes de la cartera ni necesariamente probabilidades. [ISL, capítulo 3]

Supón una regla ficticia  $\hat{y} = 2 + 0,15 \times \text{porcentaje de acciones}$ . Con 40 puntos porcentuales de acciones, predice 8 % de volatilidad. Si expresaras acciones como 0,40, necesitarías un coeficiente de 15 para producir el mismo aporte. Las unidades importan.

### Cómo se aprende la regla

Mínimos cuadrados elige coeficientes para minimizar la suma de errores cuadrados:  $RSS = \sum (y_i - \hat{y}_i)^2$ .  $\sum$  significa sumar sobre observaciones. El procedimiento busca la combinación de pesos que hace menor esa cantidad en entrenamiento. No decide por sí solo si esa combinación será estable en otros meses.

### Qué aporta y qué supone

Ofrece un punto de comparación claro. Supone aditividad y efectos constantes en las características utilizadas. Si el efecto de exposición extranjera depende de la proporción accionaria, debes agregar una interacción o usar una familia que pueda representarla. Una recta puede ser útil aunque no sea literalmente verdadera; interesa si aproxima suficientemente bien para el propósito.

### Variables relacionadas y multicolinealidad

Si renta variable, deuda y otros suman exactamente 100 %, una variable puede reconstruirse a partir de las restantes y del intercepto. Los coeficientes de mínimos cuadrados no quedan identificados de manera única al incluir todo sin una convención. Incluso sin dependencia exacta, variables muy relacionadas pueden producir coeficientes inestables.

Ridge ayuda a estabilizar la estimación, pero no convierte «aumentar acciones dejando todo lo demás igual» en una intervención posible dentro de una composición cerrada. Debes definir una categoría de referencia o una interpretación composicional coherente. El problema conceptual permanece aunque el cálculo sea resoluble.

Para recordar. Lineal significa lineal en coeficientes; mínimos cuadrados aprende esos coeficientes minimizando RSS; un coeficiente predictivo no prueba causalidad.

## APRENDIZAJE ESTADÍSTICO

## 8 · Qué agrega la regresión ridge

Ridge modifica el objetivo de mínimos cuadrados. Además de castigar errores, penaliza coeficientes grandes:

$$\text{Objetivo ridge} = \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2$$

La primera suma mide falta de ajuste. La segunda mide el tamaño conjunto de los coeficientes, excluyendo normalmente el intercepto.  $\lambda$ , leído «lambda», controla cuánto pesa la penalización. Con  $\lambda = 0$  recuperas el problema de mínimos cuadrados; con penalización muy grande y variables centradas, las predicciones se aproximan al promedio de entrenamiento. [ISL, §6.2.1; ESL, §3.4.1]

### Un ejemplo del compromiso

Imagina dos soluciones sobre variables estandarizadas. La A tiene RSS de 100 y coeficientes 8 y  $-7$ ; sus cuadrados suman 113. La B tiene RSS de 110 y coeficientes 3 y  $-2$ ; sus cuadrados suman 13. Con  $\lambda = 1$ , sus objetivos son 213 y 123. Ridge prefiere B aunque ajuste algo peor los ejemplos observados. Este ejemplo compara soluciones hipotéticas; no calcula el óptimo completo.

¿Por qué aceptar más error de entrenamiento? Porque una combinación de coeficientes grandes que se compensan puede depender demasiado de la muestra. Restringirla puede reducir varianza y mejorar generalización. Es especialmente pertinente cuando los predictores están relacionados. [ISL, §6.2.1]

### Estandarización

Antes de penalizar, suele transformarse cada predictor mediante  $z = (x - \text{media}) / \text{desviación estándar}$ . Así, una unidad representa una desviación estándar y la penalización no depende arbitrariamente de medir una variable en pesos o millones de pesos. Media y desviación se calculan solo con entrenamiento y luego se aplican sin recalcular a validación y prueba.

Ridge no elimina habitualmente variables: reduce coeficientes, pero no induce selección exacta como lasso. Tampoco todos los coeficientes individuales tienen que disminuir monótonamente con  $\lambda$ ; lo que se restringe es su tamaño conjunto. No exige normalidad de  $X$  para producir predicciones.

### Por qué entra en tu comparación

Es una referencia lineal regularizada adecuada para preguntar si las exposiciones generales contienen una señal aproximadamente aditiva. Sus límites son las relaciones no lineales no representadas y la interpretación difícil de coeficientes entre porcentajes dependientes. Puede producir estimaciones negativas de volatilidad; cualquier transformación o restricción debe definirse y evaluarse, no aplicarse silenciosamente al final.

Para recordar. Ridge intercambia parte del ajuste por estabilidad mediante una penalización L2. Si se selecciona  $\lambda$  entre candidatos, debe hacerse con validación. En la corrida documentada se fijó  $\alpha=100$ ; no se informó una búsqueda de hiperparámetros.

## APRENDIZAJE ESTADÍSTICO

## 9 · KNN: vecinos y ponderaciones

KNN busca las  $k$  carteras de entrenamiento más próximas a una nueva cartera. La versión uniforme promedia sus  $Y$ . La versión ponderada asigna más influencia a los casos cercanos. Ambas suponen que proximidad en  $X$  es informativa de  $Y$ ; ninguna garantiza equivalencia económica. [ISL, §3.5; ESL, §2.3.2]

### Ejemplo uniforme, ficticio

Con acciones de 45 %, los dos vecinos más próximos de la tabla son B y C. Ambos están a cinco puntos; su promedio es  $(8+10)/2 = 9$  %. Con los cuatro vecinos, el promedio es 10 %.

### Ejemplo ponderado, ficticio

Para una nueva cartera con 42 % de acciones, B está a distancia 2 y C a distancia 8. Con pesos inversos a la distancia:  $\hat{y} = (8/2 + 10/8) / (1/2 + 1/8) = 8,4$  %. El promedio uniforme seguiría dando 9 %. Si la distancia es cero hay que aplicar una convención específica, no dividir por cero.

### Qué se ejecutó

El experimento usó 15 vecinos, distancia euclidiana, ponderación inversa por distancia, imputación por mediana y estandarización aprendidas en TRAIN. Por ello el ejemplo uniforme ilustra el mecanismo básico, pero no reproduce esa configuración. [CIERRE, §6.3]

**RECUERDA:** seleccionar vecinos y ponderarlos son dos decisiones. La distancia se calcula con  $X$ ; usar la  $Y$  desconocida para elegirlos sería fuga de información.

| Cartera | Acciones | Distancia a 45 | Y ficticia |
|---------|----------|----------------|------------|
| A       | 20 %     | 25             | 4 %        |
| B       | 40 %     | 5              | 8 %        |
| C       | 50 %     | 5              | 10 %       |
| D       | 80 %     | 35             | 18 %       |

## APRENDIZAJE ESTADÍSTICO

# 10 · Cuándo la similitud de KNN deja de ayudar

Con dos variables estandarizadas, una distancia euclidiana es la raíz de la suma de diferencias cuadradas. Si dos carteras difieren en 1 unidad en cada eje, su distancia es  $\sqrt{2}$ . La operación combina diferencias de características, no diferencias de volatilidad: usar Y para elegir vecinos de un caso nuevo sería hacer trampa.

## Escalas y redundancia

Si una variable se mide en millones y otra entre 0 y 1, la primera puede dominar la distancia. Estandarizar reduce esa dependencia de unidades, aunque no garantiza que todas las variables tengan igual relevancia económica. Una exposición casi constante puede adquirir demasiado peso al dividir por una desviación muy pequeña; requiere revisión.

Agregar varias copias o variantes muy correlacionadas de una exposición cuenta repetidamente la misma información dentro de la distancia. En composiciones que suman 100 %, incluir todos los porcentajes también fija una geometría particular. Por eso KNN exige justificar representación, escalas y distancia, no solo escoger k.

## Maldición de la dimensionalidad

A medida que agregas dimensiones, necesitas más observaciones para llenar el espacio de combinaciones. Los vecinos pueden quedar lejos y la idea de localidad perder utilidad. No significa que KNN falle siempre con muchas variables; significa que las dimensiones irrelevantes o redundantes pueden perjudicarlo y que importa la estructura efectiva de los datos. [ISL, §3.5; ESL, §2.5]

## Vecinos que son el mismo fondo

En tu panel, los vecinos de un fondo en abril podrían ser el propio fondo en marzo y febrero. Sus Y comparten casi toda la ventana histórica. Un resultado favorable en ese escenario no demuestra capacidad para fondos no vistos. La evaluación por grupos debe impedir que el fondo evaluado aparezca en entrenamiento si esa es la capacidad que se quiere afirmar. [PROY-Y, §4]

## Extrapolación y abstención

Un promedio con pesos no negativos no sale del intervalo de Y de sus vecinos. Si todos los comparables tienen Y entre 4 y 10, KNN no predice 20. Esto lo limita cuando una cartera está lejos de todo lo conocido. Conviene registrar distancias y considerar abstención por falta de soporte. El umbral se fija con entrenamiento y validación, no mirando retrospectivamente los errores de prueba.

Para recordar. La calidad de KNN depende de que «cercano» sea informativo. Una distancia calculable no equivale a una comparación económicamente adecuada.

## APRENDIZAJE ESTADÍSTICO

# 11 · Árboles y ensambles

Un árbol de regresión divide el espacio de características mediante preguntas. Por ejemplo: «¿acciones supera 50 %?». Dentro de una rama puede preguntar «¿exposición extranjera supera 30 %?». Cada región final, llamada hoja, entrega una predicción; con pérdida cuadrática, suele ser el promedio de  $Y$  de las observaciones de entrenamiento que llegaron a ella. [ISL, §8.1.1]

## Cómo se aprende una división

El algoritmo compara variables y umbrales candidatos y elige una división que reduzca la suma de errores dentro de las regiones. Repite el proceso sujeto a restricciones. En vez de aportar una recta, construye una función por tramos. Una división no representa necesariamente un umbral económico verdadero: es una aproximación elegida con la muestra.

## Interacciones

Supón que la exposición extranjera se relaciona de forma diferente con volatilidad en fondos accionarios y en fondos de deuda. Un árbol puede dividir primero por acciones y luego aplicar reglas distintas de exposición extranjera en cada rama. Ese efecto dependiente de otra variable es una interacción.

Una suma de árboles de un solo corte representa contribuciones aditivas por variable, pero no interacciones arbitrarias entre ellas. Para representar interacciones hacen falta árboles con estructura suficiente. «Usé boosting» no basta para afirmar que efectivamente modelaste interacciones complejas.

## Por qué no quedarse con un solo árbol

Los árboles profundos pueden ser inestables: cambios modestos en la muestra alteran sus divisiones. Los ensambles combinan varios modelos para obtener una regla más útil. Un árbol pequeño sigue siendo una buena herramienta didáctica y una referencia interpretable, pero puede resultar demasiado tosco. [ISL, §§8.1 -8.2]

## Dos maneras distintas de combinar

Bagging y random forest: construyen muchos árboles con variaciones de muestras y, en random forest, de variables candidatas; después promedian sus estimaciones. Buscan reducir la variabilidad de los árboles y su dependencia mutua.

Boosting: agrega modelos secuencialmente, haciendo que el siguiente atienda lo que el conjunto todavía está ajustando mal. Los árboles no se construyen como si los anteriores no existieran. [ISL, §§8.2.1-8.2.3]

Para recordar. Árbol significa partición por reglas; ensamble significa combinación de modelos. Random forest promedia árboles diversificados; boosting acumula correcciones.

## APRENDIZAJE ESTADÍSTICO

## 12 · Gradient boosting paso a paso

Gradient boosting construye una predicción sumando correcciones pequeñas. Con pérdida cuadrática, empieza normalmente por el promedio de  $Y$ . Calcula residuos, ajusta un árbol que los aproxime y agrega una fracción de esa corrección. Repite el proceso. [ISL, §8.2.3; ESL, §§10.10-10.12]

**predicción nueva = predicción anterior +  $\eta$  × corrección del árbol**

$\eta$ , leído «eta», es la tasa de aprendizaje. No es el  $\lambda$  de ridge. Los textos pueden usar letras distintas para hiperparámetros diferentes; memoriza su función, no solo su símbolo.

### Un ejemplo completo de una iteración

Tres carteras tienen acciones de 10, 50 y 90 %, y volatilidades ficticias de 4, 8 y 12 %. El promedio inicial es 8. Los residuos son  $-4$ , 0 y 4. Un árbol de un corte separa la primera cartera de las otras dos: predice una corrección de  $-4$  para la primera y de 2 para las restantes.

Con  $\eta = 0,5$ , las nuevas predicciones son 6, 9 y 9. Los residuos pasan a  $-2$ ,  $-1$  y 3. La suma de errores cuadrados baja de 32 a 14. El siguiente árbol se ajustará a los nuevos residuos. Reducir el error de entrenamiento en esta iteración no demuestra una mejora equivalente en datos nuevos.

### Qué significa gradiente

El gradiente indica cómo cambia la pérdida al modificar las predicciones. Su dirección negativa señala cómo reducirla localmente. Para pérdida cuadrática, el negativo del gradiente es proporcional al residuo: de ahí la explicación intuitiva de corregir errores. Para otras pérdidas se utilizan pseudo-residuos; no siempre son simplemente  $y - \hat{y}$ . [ESL, §10.10.2, tabla 10.2]

### Qué hay que regular

Número de árboles, profundidad o número de hojas, mínimo de observaciones por hoja y tasa de aprendizaje. Algunas variantes agregan muestreo parcial. Muchos árboles profundos pueden aprender ruido; una tasa pequeña suele requerir más iteraciones. Detener el entrenamiento según validación también es una decisión de selección y debe respetar la estructura temporal.

### Por qué entra en tu comparación

Permite investigar si umbrales, no linealidades e interacciones aportan respecto de una relación lineal o de vecinos. No reconstruye activos ocultos, no garantiza superioridad y no convierte importancia predictiva de variables en causalidad. Los árboles habituales tampoco extrapolan tendencias suavemente fuera de los rangos observados.

Para recordar. Boosting aprende una suma secuencial de correcciones; gradient boosting define esas correcciones a partir de la pérdida.

## APRENDIZAJE ESTADÍSTICO

## 13 · Por qué comparar estos tres

La selección es un experimento sobre la forma de la relación. Ridge prueba una estructura global lineal regularizada; KNN formaliza comparables; boosting permite relaciones no lineales e interacciones. Son hipótesis complementarias, no los únicos métodos posibles. [ISL, §§3.5, 6.2.1, 8.2.3]

La comparación tiene valor aunque gane ridge: significaría que, bajo esa representación y evaluación, mayor flexibilidad no compensó sus costos. Si gana boosting, no demuestra que siempre sea superior en datos tabulares ni que haya identificado relaciones causales.

### Qué debe mantenerse común

Definición de Y, claves elegibles, particiones y métricas. El preprocesamiento puede diferir según la familia, pero siempre debe aprenderse sin acceso privilegiado a VALID o TEST. Comparar procedimientos apropiados no significa imponer a todos exactamente las mismas transformaciones.

### Qué ocurrió aquí

Se documentaron configuraciones prefijadas, no una búsqueda exhaustiva de hiperparámetros. Las conclusiones comparan esas configuraciones concretas. No acreditan cuál sería el mejor representante posible de cada familia. [CIERRE, §6.3]

### Referencias sencillas

Comparar solo modelos complejos deja sin responder si aportan sobre una regla trivial. La mediana de TRAIN minimiza pérdida absoluta entre predicciones constantes; el promedio minimiza pérdida cuadrática. El experimento incluyó mediana global y por familia de cartera, ambas aprendidas en TRAIN.

RECUERDA: la justificación inicial es diversidad de supuestos con alcance manejable. El ganador es una conclusión empírica, limitada por el diseño y los datos.

| Modelo   | Pregunta que pone a prueba                   |
|----------|----------------------------------------------|
| Ridge    | ¿Basta una combinación lineal regularizada?  |
| KNN      | ¿Ayuda la similitud local entre carteras?    |
| Boosting | ¿Aportan reglas no lineales e interacciones? |

## APRENDIZAJE ESTADÍSTICO

## 14 · Qué dejo fuera y por qué

Excluir por tarea, postergar por alcance y descartar por evidencia son decisiones distintas. Ningún libro prescribe que este proyecto deba usar exactamente tres modelos. La selección combina propiedades generales de los métodos con recursos, datos y objetivos específicos.

Clasificación sería adecuada si se decidiera estudiar categorías, pero discretizar  $Y_1$  obliga a justificar umbrales y pierde diferencias internas. PCA puede combinarse con un regresor; no entrega por sí sola una volatilidad. Además, mucha variación de  $X$  no implica mucha información para  $Y$ . [ISL, §§2.1.5, 6.3.1]

Los modelos temporales y el aprendizaje supervisado no son incompatibles. Pueden combinarse retardos de retornos y atributos de cartera. Eso cambia el conjunto de información y la pregunta respecto del núcleo inicial.

**RECUERDA:** «no responde mi pregunta» es distinto de «podría servir, pero no lo incorpore» y de «lo probé y rindió peor».

| Técnica                        | Decisión                     | Motivo                                                             |
|--------------------------------|------------------------------|--------------------------------------------------------------------|
| Logística y clasificadores     | Fuera del objetivo principal | $Y_1$ es continua; reconsiderar con categorías                     |
| Ordinal                        | Otra respuesta               | Categorías ordenadas; no equivale a Spearman                       |
| Clustering                     | Exploración auxiliar         | Agrupar $X$ ; falta conectar con $Y$                               |
| PCA / PCR                      | Extensión viable             | Reducción de dimensión, seguida de regresión                       |
| Refuerzo                       | Fuera de la tarea            | No se aprende una secuencia de decisiones                          |
| Riesgo por activos subyacentes | Insumos insuficientes        | $X$ no identifica exposición completa ni covarianzas               |
| ARIMA / GARCH                  | Otra formulación             | Dinámica temporal; valorar para objetivos con historia de retornos |

## APRENDIZAJE ESTADÍSTICO

## 15 · Alternativas viables que no debo descalificar

Lasso. Penaliza la suma de valores absolutos de coeficientes y puede dejar algunos exactamente en cero. Es útil si la selección de variables es una prioridad o se sospecha que pocas concentran la señal. Ridge es una primera opción razonable cuando varias exposiciones relacionadas pueden aportar conjuntamente. Esto es una hipótesis de diseño, no prueba de que lasso perderá. [ISL, §§6.2.1-6.2.3]

Elastic net. Combina penalizaciones L1 y L2. Puede ser una ampliación valiosa cuando interesan selección y tratamiento de predictores correlacionados. Si la matriz final de  $X$  crece mucho, su incorporación se vuelve más pertinente. [ESL, §3.4.3]

Random forest. Es un rival directo y legítimo del boosting para datos tabulares. Ambos representan no linealidades e interacciones, pero combinan árboles de manera diferente. Postergarlo mantiene el experimento acotado; no sería defendible afirmar que boosting lo domina universalmente. Si solo añades un competidor flexible, random forest es una opción natural de robustez. [ISL, §8.2.2; ESL, capítulo 15]

GAM. Suma funciones suaves de las variables. Permite curvaturas con una estructura interpretable. Resulta atractivo si ridge deja patrones no lineales y se desea explicar efectos con curvas. Las interacciones deben especificarse o modelarse expresamente según la formulación. [ESL, §9.1]

SVR y kernels. La regresión de vectores de soporte admite respuesta numérica y relaciones no lineales. No debes descartarla diciendo «SVM solo clasifica». Añade decisiones de kernel, regularización, escalas y tolerancia al error. Puede reconsiderarse si los modelos iniciales dejan patrones sistemáticos; su costo debe medirse con la muestra real. [ESL, §12.3.6]

Redes neuronales. Pueden trabajar con datos tabulares y aprender relaciones complejas. No son exclusivas de imágenes ni requieren siempre millones de filas. Se postergan porque tu representación actual consiste en atributos agregados y ya hay tres hipótesis comparables con menor carga inicial de diseño. No hay evidencia previa que garantice que una red aporte valor adicional aquí. [ESL, §§11.3-11.4]

Mínimos cuadrados y árbol único. Conviene comprenderlos y pueden servir como referencias. Ridge y boosting agregan mecanismos para manejar limitaciones de esas bases, pero no las hacen inútiles en todas las muestras.

Para recordar. «No lo incluyo en la primera comparación» es más preciso que «no sirve». La ampliación debe responder a una necesidad o una hipótesis, no a acumular nombres.

## APRENDIZAJE ESTADÍSTICO

# 16 · Qué implica trabajar con carteras agregadas

Una cartera con 30 % invertido en otros fondos no revela necesariamente qué activos poseen esos vehículos. Dos fondos con los mismos porcentajes generales pueden diferir en sectores, emisores, duración, moneda, cobertura, liquidez o apalancamiento. Tu  $X$  observa una proyección parcial de la exposición económica. [PROY- $X$ , §§2.1-2.4]

## El límite de información

Si dos carteras tienen exactamente la misma  $X$ , un predictor determinista que usa solo  $X$  les asignará la misma estimación. Si sus  $Y$  difieren, no puede reproducir ambas exactamente. Aumentar complejidad no crea la información que falta. Puede aprovechar asociaciones indirectas, pero eso exige comprobar que se mantienen fuera de la muestra.

## La fotografía frente a la película

$Y_1$  resume cinco años;  $X$  representa un mes. Una cartera actual puede ser distinta de la que generó buena parte de los retornos de la ventana. El modelo aprende una asociación entre una fotografía y una historia. Un mal ajuste podría reflejar esa desalineación, además de problemas del algoritmo. Estudiar estabilidad de composición o usar resúmenes históricos sería una extensión que modifica el conjunto de información.

## Composiciones y variables cruzadas

Porcentajes de clases mutuamente excluyentes pueden sumar 100 %. Pero exposición extranjera y renta variable pueden superponerse: una acción extranjera pertenece a ambos criterios. No sumes todas las columnas esperando 100 %. Cada bloque necesita una definición y un denominador. Derivados pueden requerir tratamiento de exposición y no simple suma de valores contables.

Nulo tampoco equivale automáticamente a cero. Un archivo ausente puede significar falta de información. Imputar  $X$  exige una regla aprendida o fijada dentro del protocolo; no debe inventarse  $Y$  para observaciones sin la ventana válida. [PROY- $X$ , §2.4; PROY- $Y$ , §3]

## Selección de la muestra

Los fondos con cinco años completos forman una subpoblación. Aprender en ellos y aplicar a fondos nuevos introduce un problema de transporte de la relación. Evaluar fondos maduros no vistos simula desconocimiento de su historia para el modelo, pero no reproduce perfectamente un fondo recién creado, cuya  $Y$  histórica de cinco años no existe.

Para recordar. Un buen algoritmo no compensa cualquier representación. La definición económica de  $X$ , la calidad de  $Y$  y la población evaluada forman parte del método.

## APRENDIZAJE ESTADÍSTICO

# 17 · Evaluación y fuga de información

Evaluar exige simular qué se quiere afirmar. Una partición aleatoria de filas puede dejar meses del mismo fondo en entrenamiento y prueba. Como sus  $Y_1$  comparten gran parte de la ventana quinquenal, el resultado no representa necesariamente transferencia a fondos desconocidos. [PROY-Y, §4]

## Diseño documentado

La corrida separó RUN completos y fechas: TRAIN hasta diciembre de 2022, VALID en 2023 y TEST de enero de 2024 a septiembre de 2026. La regla exacta y la reserva se explican en el capítulo 24. Esta es una configuración ejecutada según el informe, no la propuesta genérica de validación cruzada del manual anterior. [CIERRE, §6.2]

## Preprocesamiento dentro de entrenamiento

Imputación, estandarización, selección de columnas y transformaciones ajustables deben aprenderse usando TRAIN. Si se realiza validación cruzada, deben reaprenderse en cada entrenamiento interno. Ajustarlas primero con todos los datos deja que el procedimiento vea la distribución reservada. [ESL, §7.10.2]

## Solapamiento no significa automáticamente fuga futura

Dos ventanas quinquenales terminadas en meses próximos comparten retornos. Fondos distintos comparten shocks. Separar RUN y fechas no elimina esas dependencias, pero tampoco prueba por sí solo que el algoritmo usó información futura: el objetivo es retrospectivo. La fuga debe examinarse en disponibilidad de  $X$ , corte de  $Y$ , selección de series y uso de TEST.

## Qué aporta un embargo

Una separación de períodos suficientemente amplia puede estudiar ventanas sin retornos compartidos. Un embargo breve no elimina cinco años de superposición. Esa sensibilidad consume observaciones y puede cambiar el régimen económico evaluado; no debe presentarse como remedio automático universal.

## Fecha de cartera y publicación

Para simular uso a una fecha concreta,  $X$  debe haber estado disponible entonces. Los escenarios 0/30/60/90/180 días están propuestos, no ejecutados en la evidencia examinada. Si falta publicación observada, los desfases serán supuestos. [DICTAMEN, II.5]

RECUERDA: el diseño debe corresponder a la afirmación. Fondos maduros no vistos no son fondos verdaderamente nuevos sin historia.

## APRENDIZAJE ESTADÍSTICO

## 18 · Cómo medir y elegir al final

El error absoluto medio o MAE es el promedio de  $|y - \hat{y}|$ . Conserva la unidad de Y. Un MAE de 2 significa un error absoluto medio de 2 puntos porcentuales de volatilidad anualizada, no un error relativo de 2 %.

RMSE es la raíz del promedio de errores cuadrados. También conserva la unidad de Y, pero da mayor peso a errores grandes. Si los errores son 1, 1 y 7,  $MAE = 3$  y  $RMSE = \sqrt{17} \approx 4,12$ . Reportar ambos permite distinguir error típico promedio y sensibilidad a fallos grandes.

$R^2$  compara la suma de errores cuadrados del modelo con la variación de Y respecto de su media en el conjunto evaluado:  $R^2 = 1 - SSE/SST$ . Puede ser negativo fuera de muestra. Si Y es constante, la fórmula usual queda indefinida. Además de  $R^2$ , compara explícitamente con una constante aprendida en entrenamiento, que sí representa una alternativa operativa.

Spearman evalúa concordancia de orden. Un modelo puede ordenar bien los fondos y equivocarse sistemáticamente en sus niveles. Por eso no sustituye MAE ni RMSE. Una regresión ordinal tampoco es «hacer Spearman»: una construye un modelo para categorías ordenadas; el otro es una medida de asociación de rangos.

### Cómo elegir sin sobreinterpretar

No basta una diferencia diminuta en un promedio. Revisa estabilidad entre períodos, distribución de errores por tipo de fondo, comportamiento en extremos y resultados en grupos no vistos. La incertidumbre debe respetar dependencia por fondo y tiempo; tratar cada fila como independiente produce una precisión aparente excesiva.

Si boosting mejora claramente, puedes afirmar que aportó bajo las condiciones evaluadas. No que descubrió la causalidad del riesgo. Si ridge queda muy cerca con menor complejidad, puede ser preferible. Si KNN ayuda solo donde hay vecinos próximos, puedes definir un ámbito de aplicación más restringido.

### Explicaciones y límites

Coeficientes ridge, vecinos KNN e importancias de árboles son explicaciones diferentes. Las variables correlacionadas complican interpretar todas ellas. La importancia de una variable no mide automáticamente el efecto de intervenir sobre ella.

Para recordar. Se elige por desempeño fuera de muestra, estabilidad y utilidad de la estimación. La complejidad es un costo que debe justificar su aporte, no un mérito por sí mismo. [ISL, §2.2; ESL, capítulo 7]

## APLICACIÓN AL PROYECTO · DEFINICIÓN OPERACIONAL

# 19 · Cómo se construye Y1

Y1 es la respuesta principal: una medida operacional retrospectiva de volatilidad anualizada. Se calcula primero por RUN, serie, moneda literal y mes. Después se elige una serie con continuidad histórica observable hasta el corte, sin usar su Y resultante ni información posterior para decidir. [PROY-Y, §§2-3]

## Ventana y anclas

El corte es el último martes calendario del mes. Cada ancla admite la última cuota de ese martes o hasta cuatro días calendario anteriores. Se requieren 261 anclas válidas para construir 260 retornos semanales consecutivos. Deben respetarse rupturas y la regla operacional de brechas de precios superiores a cinco días.

La transformación diaria documentada es  $r_d = (VC_d / VC_{d\_anterior}) \times FR_d \times FA_d - 1$ . Esto describe la operación implementada; no certifica la interpretación económica de los factores para todos los regímenes. La composición semanal debe reconstruirse desde el algoritmo y sus anclas; no basta sustituirla por un cociente de cuotas sin verificar el ajuste. [CIERRE, §3.1]

## Fórmula y unidades

Media semanal:  $r\_media = (r_1 + \dots + r_{260}) / 260$ .

Desviación muestral:  $s = \sqrt{[\sum_i (r_i - r\_media)^2 / 259]}$ .

$Y1 = 100 \times \sqrt{52} \times s$ .

Los retornos  $r_i$  se expresan como proporciones. Si  $s = 0,01$ ,  $Y1 = 7,2111\%$ . El factor 100 convierte proporción a porcentaje;  $\sqrt{52}$  aplica la convención de anualización. La desviación muestral divide por 259, no por 260. No se presupone que anualizar haga independientes los retornos.

## Abstenerse también es un resultado

Sin 260 retornos válidos no se imputa Y1. Si  $s=0$ , se conserva  $Y1=0$ . Se deben registrar causas de abstención, serie, moneda y banderas. Ventanas con conflictos materiales acreditados quedan fuera de la muestra principal hasta resolver o acotar el efecto. No se eliminan automáticamente todos los retornos extremos finitos.

**RECUERDA:** un campo que dice “260 semanas” es un control de condición; una auditoría independiente necesita fechas, retornos y factores para recalcular la cifra.

## APLICACIÓN AL PROYECTO · DEFINICIÓN OPERACIONAL

## 20 · Y2 y el indicador completo CMF

Y2 es una sensibilidad a la cola izquierda. No fue la respuesta utilizada para entrenar los modelos informados. Su fórmula operacional se define sobre retornos estandarizados, no directamente sobre el percentil de retornos sin transformar. [PROY-Y, §2; CIERRE, §3.3]

### Cálculo completo

1. Calcular  $s$  y  $r_{\text{media}}$  sobre los mismos 260 retornos.
2. Estandarizar:  $z_i = (r_i - r_{\text{media}}) / s$ .
3. Ordenar los  $z$  y obtener  $q_{0,05}$  con interpolación lineal.
4. Calcular  $k = \text{máximo}(1, q_{0,05} / -1,65)$ .
5. Calcular  $Y2 = Y1 \times k$ .

La posición de interpolación en índices que comienzan en cero es  $(260-1) \times 0,05 = 12,95$ . Por tanto  $q = z_{\text{ordenado}}[12] + 0,95 \times (z_{\text{ordenado}}[13] - z_{\text{ordenado}}[12])$ . En posiciones que comienzan en uno, se interpola entre la 13.<sup>a</sup> y la 14.<sup>a</sup> observación. No confundir con nearest-rank.

### Ejemplo ficticio

Si  $Y1=10$  y  $q=-2,31$ ,  $k=1,40$  y  $Y2=14$ . Si  $q=-1,32$ , el cociente es 0,80, pero  $k$  se fija en 1:  $Y2=10$ . El máximo impide que este ajuste reduzca  $Y1$ .

Si  $s=0$  o  $0 < s \leq 10^{-12}$  en unidades semanales, se preserva  $Y1$  y  $Y2$  queda indefinida. No reemplazar esa ausencia por cero. La resolución mantiene percentil nearest-rank y cuantil normal exacto como sensibilidades alternativas, no como la definición principal.

### Tres objetos diferentes

La referencia técnica CMF se describe en los documentos del proyecto como  $B = \sigma \times k \times (1 + w)$ . El componente  $w$  requiere una definición de exposición adicional y no forma parte de  $Y1$  ni  $Y2$ . El experimento no estimó  $w$  ni  $B$  y no produjo categorías oficiales. Este manual reproduce esa referencia documental; no certifica su equivalencia regulatoria ni verifica su vigencia normativa.

**RECUERDA:**  $Y1$  es volatilidad;  $Y2$  agrega un ajuste operacional de cola;  $B$  requiere además exposición y convenciones regulatorias. No son nombres intercambiables.

## APLICACIÓN AL PROYECTO · CARACTERÍSTICAS

## 21 · Qué mide realmente X

La representación económica spot combina posiciones NACI y EXTR. Las declaraciones OPCI, FUTU y OPLA se cuentan, pero sus nocionales o subyacentes no se agregan al denominador spot. Las cinco clases son renta fija, renta variable, fondos, alternativos y otros. No hay look-through: las cuotas de otros fondos permanecen como fondos. [CIERRE, §4]

### Denominador y participaciones

Sea  $v_i$  la valorización válida de una posición spot y  $V = \sum v_i$ . El peso individual es  $a_i = v_i/V$ . Para una clase  $c$ :  $X_c = 100 \times \sum (v_i \text{ de clase } c)/V$ . Estas proporciones no son automáticamente porcentajes sobre patrimonio total ni exposición neta con derivados.

Los indicadores que dependen de  $V$  exigen cobertura válida de valorizaciones y  $V > 0$ . El código descrito descarta valores negativos de estos agregados; su interpretación frente a activos y pasivos debe revisarse económicamente. Los porcentajes declarados sobre activos son otro bloque y no se fuerza su suma a 100.

### Concentración y exposición

HHI por posición =  $\sum a_i^2$ . Top1 =  $100 \times \text{máximo}(a_i)$ . Top5 suma las cinco posiciones de mayor peso. Contar filas concentradas no equivale a agrupar emisores: un mismo emisor puede estar repetido. El HHI de emisor descrito agrega dentro de cada familia y no reconcilia automáticamente identidades entre NACI y EXTR.

La proporción EXTR es valor spot extranjero dividido por total spot. No equivale a riesgo cambiario neto ni identifica todo riesgo económico internacional. El vencimiento medio pondera años residuales de posiciones con fecha interpretable; no es duración modificada ni convexidad.

### Calidad de primer nivel

El diccionario contiene 174 pares familia-código. Los estados reportados son 125.780 MAPEO\_COMPLETO, 6.412 MAPEO\_PARCIAL, 566 SIN\_MAPEO y 24 SIN\_POSICIONES\_SPOT: suman 132.782. Mapeo completo significa que los códigos spot fueron clasificados bajo esas reglas. No significa moneda verificada, precios correctos, cobertura económica total o información disponible al momento de inferencia.

**RECUERDA:** antes de interpretar una columna, explica numerador, denominador, unidad, tratamiento de nulos y qué exposición queda fuera.

## APLICACIÓN AL PROYECTO · SELECCIÓN EN TRAIN

## 22 · De 28 candidatas a 26 entradas

La selección se documentó exclusivamente en 31.973 filas TRAIN. La regla retuvo columnas con más de diez valores no nulos y más de un valor distinto. Dos quedaron fuera por constancia: cobertura\_emisor y cobertura\_pct, ambas con 31.973 no nulos y un solo valor distinto. [DICTAMEN, II.2]

Vencimiento\_medio\_anos\_declarados se conservó con 22.063 observados: faltaban 9.910 valores, aproximadamente 31,0 % de TRAIN. Imputar no vuelve observados esos datos; solo permite aplicar una regla computacional. La mediana se aprende en TRAIN.

El listado siguiente reproduce la composición de las 26 entradas por grupos semánticos del informe; no inventa nombres de columnas no adjuntados. Para una ejecución debe usarse el esquema exacto de FEATURES y su CSV versionado. [CIERRE, §6.1]

**RECUERDA:** la prueba documentada de selección evita mirar TEST para descartar constantes. No certifica que todos los atributos midan óptimamente riesgo económico.

| Grupo de atributos retenidos                            | Cantidad |
|---------------------------------------------------------|----------|
| Pesos spot por las cinco clases                         | 5        |
| Proporción spot EXTR                                    | 1        |
| HHI por posición, Top1 y Top5                           | 3        |
| Recuentos de posiciones y familias                      | 6        |
| Números de códigos de instrumento, país y moneda        | 3        |
| HHI de emisor identificado                              | 1        |
| Vencimiento medio residual declarado                    | 1        |
| Suma de porcentajes declarados                          | 1        |
| Porcentajes declarados por cada una de las cinco clases | 5        |
| Total utilizado                                         | 26       |

La imputación y el escalado se ajustan después de separar los datos. No deducir de que una columna fue retenida que se conoce su importancia marginal o efecto causal.

## APLICACIÓN AL PROYECTO · DENOMINADORES

## 23 · Del universo a la muestra efectiva

El proyecto distingue posiciones, cuotas, series y fondos-mes. Los 10.973.842 registros instrumentales y 18.727.871 cuotas diarias describen insumos, no observaciones independientes del modelo. [PROY-X; CIERRE, §§2 y 5]

La tabla presenta el flujo vigente documentado. Las identidades aritméticas se comprobaron en la auditoría del manual; los conteos no se reejecutaron sobre SQLite.

No mezclar con la memoria del 12 de octubre: sus 132.698 X con cartola del mismo mes y 65.532 Y exploratorias corresponden a criterios/versiones anteriores. No son sinónimos del empalme operacional posterior. Una conciliación causal requiere comparar claves, no solo restar totales.

| Etapas                   | Fondo-mes | Diferencia o significado          |
|--------------------------|-----------|-----------------------------------|
| Panel X                  | 132.782   | Universo con características      |
| Empalme X-Y              | 132.616   | 166 X sin registro Y              |
| Y1 numérica en cruce     | 65.755    | 66.861 empalmadas sin Y1 numérica |
| Y1 numérica + X completa | 63.633    | 2.122 con X incompleta            |
| TRAIN                    | 31.973    | 421 RUN                           |
| VALID                    | 659       | 58 RUN                            |
| TEST                     | 1.761     | 63 RUN                            |
| Particiones activas      | 34.393    | TRAIN + VALID + TEST              |
| Reserva no utilizada     | 29.240    | 63.633 – 34.393                   |

En la cohorte de 63.633 hay 258 Y1 iguales a cero y 257 banderas de extremos. Moneda funcional y semántica histórica de factores permanecen pendientes para la cohorte. La elegibilidad operacional no es certificación económica. [CIERRE, §5]

**RECUERDA:** 63.633 no es el tamaño de entrenamiento, ni el número de fondos únicos, ni la suma de TRAIN+VALID+TEST.

## APLICACIÓN AL PROYECTO · VALIDACIÓN

## 24 · Cómo se separaron RUN y fechas

La partición documentada combina dos condiciones. Primero asigna cada RUN a un grupo determinista. Después admite solo las fechas autorizadas para ese grupo. Las filas que no cumplen ambas condiciones quedan en reserva. [CIERRE, §6.2]

### Regla reproducible declarada

Se calcula SHA256 del texto "Seminario2\_XY\_V1|" concatenado con el RUN. Se toman los primeros 12 caracteres hexadecimales, se convierten a entero y se aplica módulo 100. Resultado menor que 70: grupo TRAIN; entre 70 y 84: VALID; entre 85 y 99: TEST.

El formato exacto del RUN debe conservarse: cambiar su representación textual puede cambiar el hash. No se asigna por rendimiento, tamaño de Y ni resultado del modelo.

### Qué responde este diseño

Se informa ausencia de RUN compartidos entre las tres particiones activas. Se evalúa a la vez cambio de tiempo y diferencia de fondos. Por ello una caída en desempeño no puede atribuirse automáticamente al paso del tiempo: puede cambiar la composición de la cohorte.

Un fondo TEST puede tener larga historia de cuotas, aunque el algoritmo no lo haya visto. Esto simula desconocimiento de su historia para el modelo; no demuestra funcionamiento sobre un fondo recién creado cuya volatilidad quinquenal aún no existe.

### Qué no se comprobó de nuevo aquí

No se reejecutó el hash sobre la lista completa ni se verificaron intersecciones desde archivos de membresía. La descripción es coherente y está documentada, pero una réplica exige claves y scripts.

**RECUERDA:** grupo de RUN y filtro temporal son requisitos simultáneos. No se deben reasignar fondos de reserva para aumentar artificialmente TRAIN.

| Partición | Fecha admitida             | Filas / RUN  |
|-----------|----------------------------|--------------|
| TRAIN     | Hasta diciembre de 2022    | 31.973 / 421 |
| VALID     | Enero-diciembre de 2023    | 659 / 58     |
| TEST      | Enero 2024-septiembre 2026 | 1.761 / 63   |

## APLICACIÓN AL PROYECTO · IMPLEMENTACIÓN

## 25 · Configuración realmente informada

Esta ficha describe el primer experimento, no una recomendación universal. Los hiperparámetros estaban definidos en el script; no se documentó una búsqueda exhaustiva. Comparar sus resultados no equivale a comparar el óptimo de cada familia. [CIERRE, §6.3]

Los tres modelos supervisados limitaron sus predicciones a cero como mínimo:  $\hat{y}_{\text{final}} = \max(\hat{y}, 0)$ . El recorte importa para reproducir métricas y debe declararse. No forma parte de la definición matemática general de ridge o KNN.

Para histogram boosting deben conservarse también los parámetros efectivos, defaults, criterio de parada y número real de iteraciones. “Hasta 150” no demuestra que se hayan ejecutado exactamente 150. No se reconstruyen aquí defaults a partir de memoria ni se afirma una validación temporal interna que no esté acreditada.

Entorno informado: Python 3.12, NumPy 2.3.3, pandas 2.3.3 y scikit-learn 1.8.0. Son versiones reportadas por el informe, no comprobadas como entorno de la corrida en esta revisión.

| Método               | Configuración declarada                                                                                                                                   |
|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| Mediana global       | Mediana de Y1 de TRAIN                                                                                                                                    |
| Mediana por familia  | Clase dominante si alcanza 40 %; si no, mixta. Mediana TRAIN por familia; respaldo global cuando falta. Auditar desempates.                               |
| Ridge                | Imputación por mediana + escalado TRAIN; $\alpha=100$ ; intercepto normalmente no penalizado                                                              |
| KNN                  | Mediana + escalado TRAIN; $k=15$ ; distancia euclidiana; weights=distance; convención especial para distancia cero                                        |
| HistGradientBoosting | Mediana TRAIN; pérdida cuadrática; $\eta=0,05$ ; max_iter=150; max_leaf_nodes=15; max_depth=5; min_samples_leaf=30; l2_regularization=10; random_state=42 |

**RECUERDA:** el ejemplo de un vecino,  $\lambda=1$  o una sola corrección enseña una operación. Esta ficha identifica lo que produjo las métricas reportadas.

## APLICACIÓN AL PROYECTO · RESULTADOS DOCUMENTADOS

## 26 · Resultados: qué ganó y según qué

En VALID, boosting tuvo menor MAE, RMSE y mayor  $R^2$  y Spearman. KNN tuvo menos subestimaciones graves: 95 frente a 139. Por tanto “mejor” depende también de qué error interesa evitar. Las cifras siguientes son reportadas, no una nueva estimación. [CIERRE, §§7.1-7.2]

VALID: 659 filas de 58 RUN, año 2023.

| Modelo          | MAE   | RMSE  | $R^2$  | $\rho$ | Sub. >5 |
|-----------------|-------|-------|--------|--------|---------|
| Boosting        | 3,004 | 4,260 | 0,715  | 0,874  | 139     |
| KNN-15          | 3,158 | 4,735 | 0,647  | 0,821  | 95      |
| Ridge           | 3,741 | 4,895 | 0,623  | 0,782  | 134     |
| Mediana familia | 4,034 | 5,407 | 0,540  | 0,792  | 143     |
| Mediana global  | 6,668 | 8,557 | -0,151 | ND     | 259     |

TEST: 1.761 filas de 63 RUN, enero 2024-septiembre 2026.

| Modelo          | MAE   | RMSE  | $R^2$  | $\rho$ | Sub. >5 |
|-----------------|-------|-------|--------|--------|---------|
| Boosting        | 3,254 | 4,727 | 0,624  | 0,860  | 113     |
| KNN-15          | 3,616 | 5,623 | 0,468  | 0,802  | 118     |
| Ridge           | 3,649 | 4,996 | 0,580  | 0,777  | 171     |
| Mediana familia | 4,478 | 5,761 | 0,442  | 0,704  | 292     |
| Mediana global  | 6,248 | 8,229 | -0,139 | ND     | 599     |

MAE y RMSE están en puntos porcentuales;  $\rho$  es Spearman; Sub. >5 cuenta  $Y - \hat{y} > 5$  pp. ND indica no disponible/no definido en la tabla; una predicción constante no tiene variación de rangos.

La reducción relativa de MAE usa cifras no redondeadas:

$100 \times (4,477613 - 3,254105) / 4,477613 = 27,325\%$ . La diferencia absoluta es aproximadamente 1,224 pp. Son magnitudes diferentes.

**RECUERDA:** estas métricas acreditan una comparación exploratoria documentada. No prueban seguridad en cada fondo ni causalidad.

## APLICACIÓN AL PROYECTO · INCERTIDUMBRE

## 27 · Incertidumbre y estabilidad temporal

Un resultado puntual no explica cuánto podría cambiar con otra muestra ni dónde falla. El informe utilizó bootstrap pareado por RUN: 63 fondos completos remuestreados con reposición, 4.000 veces, conservando sus meses y calculando ambas pérdidas sobre las mismas remuestras. [CIERRE, §7.3]

### Qué significa el intervalo

El intervalo percentil del 95 % informado para la reducción relativa de MAE fue 17,6 %-36,8 %. Para diferencia absoluta: 0,767-1,713 pp. Es incertidumbre aproximada condicionada al TEST, no un intervalo que certifique todos los futuros mercados o fondos.

Preservar meses por RUN evita tratarlos como observaciones independientes, pero no resuelve completamente shocks simultáneos entre fondos, selección de modelo o cambio de régimen. Tampoco incorpora toda incertidumbre de reentrenamiento. Si la métrica agrega todas las filas, los fondos con más meses mantienen mayor peso; remuestrear por fondo no implica promediar sus MAE con pesos iguales.

### La mejora no fue constante

La comparación anual siguiente muestra deterioro de MAE del modelo y menor ventaja relativa. No prueba por sí sola deriva causal: los RUN presentes también cambian. [CIERRE, §7.4]

### Errores graves

Boosting subestimó más de 5 pp en 113/1.761 filas, aproximadamente 6,4 %. La referencia por familia lo hizo en 292/1.761, aproximadamente 16,6 %. Esto no elimina casos individuales adversos; la media puede mejorar mientras algunos fondos empeoran.

RECUERDA: un intervalo estrecho bajo un esquema de remuestreo incompleto no corrige sesgos de medición ni garantiza transporte a otra población.

| Año            | Filas | MAE ref. | MAE boost. | Mejora |
|----------------|-------|----------|------------|--------|
| 2024           | 652   | 4,580    | 2,993      | 34,7 % |
| 2025           | 641   | 4,393    | 3,187      | 27,4 % |
| 2026 ene.-sep. | 468   | 4,451    | 3,710      | 16,7 % |

## APLICACIÓN AL PROYECTO · LÍMITES ABIERTOS

## 28 · Qué sigue pendiente de validación

La resolución permite usar Y1 operacionalmente con limitaciones. Eso no certifica todos los datos. El dictamen del 19 de octubre distingue la selección de variables verificada en su paquete de cinco controles parcialmente resueltos. [PROY-Y; DICTAMEN, I-II]

### FA/FR: un conflicto concreto

RUN 8001, serie B, 19-12-2008: dos líneas del TXT presentan el mismo valor cuota 359,9104 y factores distintos. Una tiene FR=1 y FA=1,019020273; otra FR=1,000896629 y FA=1. Con el retorno base documentado, producen aproximadamente +0,37989 % y -1,40540 %: diferencia de 1,78529 pp diaria. No es una diferencia calculada de Y1.

El efecto exige enlazar evento, serie seleccionada, ventana y partición. Un retorno de 2008 no entra directamente en ventanas quinquenales terminadas en 2024-2026; podría afectar datos antiguos de entrenamiento o reserva y, por esa vía, el ajuste. No seleccionar el factor que mejora los resultados.

### Moneda y continuidad

Los literales \$\$, PROM, EUR y MONEDA NO DEF. no acreditan moneda funcional por RUN-fecha. No convertirlos por intuición. Una etiqueta de continuidad y una guarda que exige 260 semanas son útiles, pero las 63.633 claves de la tabla de trazabilidad no contienen todas las anclas y retornos para recálculo independiente.

### Disponibilidad y robustez

No hay ejecución acreditada de escenarios de publicación 0/30/60/90/180 días. Las cuatro etiquetas de sensibilidad del TEST conservaron las mismas filas y métricas, porque las banderas utilizadas no retiraron observaciones de TEST. Esto no prueba que no existan anomalías no identificadas ni una depuración efectiva.

### Dos sensibilidades diferentes

Primero se pueden evaluar predicciones congeladas en una cohorte realmente elegible. Después se puede reentrenar con datos depurados y reglas fijas. Son preguntas distintas; ambas, si se deciden después de ver TEST, son diagnósticas y no una nueva confirmación independiente.

RECUERDA: una limitación declarada con precisión es mejor evidencia que una etiqueta “validado” sin prueba que la respalde.

## APLICACIÓN AL PROYECTO · PLAN DE VERIFICACIÓN

## 29 · Procedimiento reproducible y cierre

Reproducir es poder reconstruir una decisión y un resultado desde insumos identificados. No basta conservar una tabla final de métricas. La secuencia siguiente es un plan de verificación pendiente, no trabajo ejecutado para este manual.

### Antes de volver a entrenar

1. Congelar versiones, hashes, scripts, diccionario, CSV de Y, lista de variables y entorno. Abrir SQLite con mode=ro y query\_only=ON; escribir derivados en otra carpeta.
2. Conciliar claves y cohortes. Verificar 132.782 → 132.616 → 65.755 → 63.633 y las particiones, o explicar cada divergencia.
3. Reconstruir serie-moneda, 261 anclas y 260 retornos de las 34.393 filas activas. Comparar Y1 con tolerancias numéricas prefijadas; no usar Y1 almacenada como entrada del recálculo.
4. Resolver o acotar eventos de factores, cambios de denominación y disponibilidad de X. Crear matriz de elegibilidad por RUN-mes con motivos y fuentes.

### Al verificar el experimento

5. Comprobar selección de columnas, mediana, escalado y ajuste exclusivamente en TRAIN. Conservar parámetros efectivos y criterios de parada de todos los modelos.
6. Recalcular métricas de predicciones congeladas sobre las mismas claves. Verificar clipping, denominadores, pérdidas y bootstrap pareado.
7. Ejecutar sensibilidades efectivas sin optimizar para recuperar un resultado favorable. Reportar filas y RUN excluidos por cada condición. Si una cohorte queda vacía, declararlo.

### Cómo informar el estado

PASS significa que la prueba especificada pasó. FAIL significa que contradijo su condición. NO VERIFICABLE significa que faltan insumos para resolverla. Un PASS de sumas no es un PASS de validez económica; un hash verifica identidad de bytes, no verdad de los datos.

### Qué congelar en el cierre

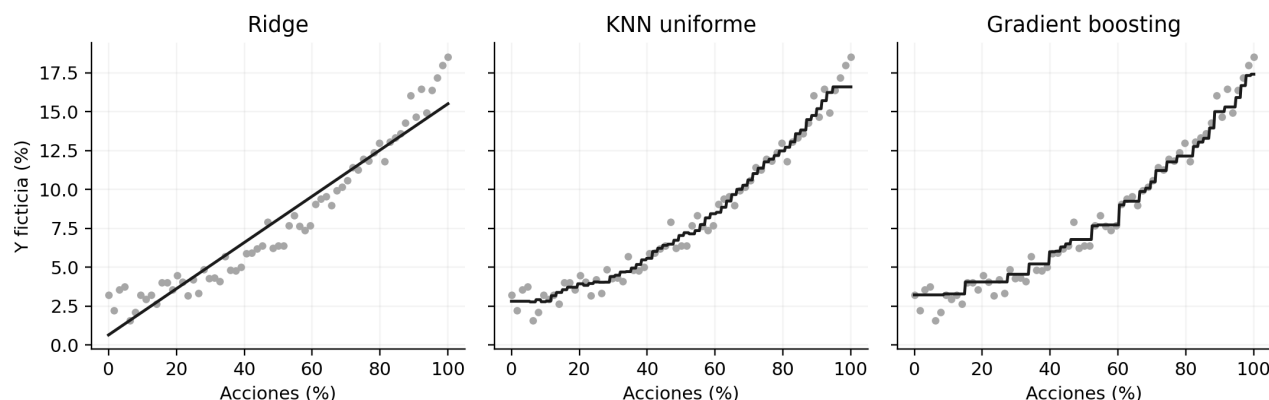
Objetivo, fórmula de Y, regla de serie, códigos de moneda, definición de X, nullos, filtros, particiones, hiperparámetros, predicciones y métricas. Distinguir integridad, consistencia computacional, validez económica, desempeño exploratorio y equivalencia regulatoria.

RECUERDA: la siguiente decisión metodológica debe responder a evidencia nueva o una hipótesis explícita, no al deseo de que gane un algoritmo.

## EJEMPLO FICTICIO · GRÁFICO GENERADO PARA ESTE MANUAL

## 30 · Ver las tres formas de ajustar

Los tres paneles usan los mismos datos sintéticos y las mismas escalas. Los puntos son observaciones ficticias y las líneas son modelos ajustados para esta ilustración. No contienen fondos chilenos ni reproducen la corrida del proyecto.



Ridge recibe solo una característica original y por eso entrega una recta. KNN produce promedios locales; cambia cuando cambian los vecinos. Boosting aproxima la curva mediante reglas por tramos.

La curva ficticia se generó con  $Y=3+0,0015 \times X^2 + \text{ruido}$ , con semilla 42. La no linealidad fue introducida deliberadamente; la figura no es una competencia imparcial sobre qué método será superior en cualquier problema.

Configuraciones didácticas: ridge  $\alpha=1$  sobre X estandarizada; KNN uniforme con siete vecinos sobre X estandarizada; GradientBoostingRegressor con 60 árboles, profundidad máxima 2 y tasa 0,05. Es boosting clásico para ilustrar el mecanismo, no HistGradientBoosting del experimento.

### Ejercicio visual

Imagina datos casi alineados y una muestra pequeña. Una curva flexible podría seguir ruido sin aportar una reducción útil de sesgo. Por eso capacidad de representación y desempeño fuera de muestra son propiedades diferentes. [ISL, §§2.2.2 y 3.5]

RECUERDA: el gráfico ayuda a comprender cómo se construye una estimación. Las tablas del capítulo 26 informan lo que ocurrió en el experimento.

## PRÁCTICA · DATOS FICTICIOS

# 31 · Ejercicios de fundamentos

Resuelve sin consultar el capítulo siguiente. Todos los datos son ficticios.

## A Identificar la tarea

1. Tienes proporciones de cartera y una volatilidad numérica por fondo-mes. ¿Cuál es el paradigma, la tarea y la unidad de observación?
2. Si cambias  $Y$  a cinco categorías ordenadas, ¿qué cambia en el problema? ¿Basta con medir Spearman?
3. Agrupas carteras con clustering y encuentras cuatro grupos. ¿Ya construiste un predictor de volatilidad?

## B Calcular y explicar

4. Tres vecinos tienen  $Y$  de 5, 8 y 11. ¿Qué predice KNN uniforme con  $k = 3$ ? ¿Qué información falta para decidir si esos vecinos son buenos comparables?
5. Una predicción es 9 y la  $Y$  observada es 12. Calcula residuo, pérdida absoluta y pérdida cuadrática.
6. Dos soluciones ridge tienen RSS de 50 y 54, y sumas de coeficientes cuadrados de 30 y 8. Con  $\lambda = 1$ , ¿cuál tiene menor objetivo? ¿Demuestra eso menor error de prueba?
7. En boosting, la predicción actual es 8, el árbol entrega una corrección de 4 y la tasa es 0,1. ¿Cuál es la nueva predicción? ¿Por qué no se agrega necesariamente toda la corrección?

## C Detectar fallos metodológicos

8. Estandarizas usando todos los años y después separas entrenamiento y prueba. ¿Qué está mal?
9. Tu KNN encuentra como vecinos los meses anteriores del mismo fondo. ¿Qué capacidad podría estar demostrando y cuál no?
10. El modelo más complejo ajusta mejor entrenamiento, pero peor validación. ¿Qué explicación considerarías?
11. Dos fondos tienen exactamente la misma  $X$  agregada y  $Y$  distintas. ¿Puede un predictor determinista que solo usa esa  $X$  acertar ambas exactamente?
12. Dices «descarto redes porque solo sirven para imágenes» y «descarto SVM porque solo clasifica». Corrige ambas afirmaciones.

## D Defender una elección

13. Explica en un minuto por qué comparar ridge, KNN y boosting no es una selección arbitraria.
14. Nombra una técnica incompatible con tu objetivo actual y otra compatible que solo estás postergando.
15. ¿Qué resultado te haría preferir ridge a boosting? ¿Qué hallazgo te haría revisar  $X$  antes de agregar modelos?

Criterio de dominio. Saber repetir «lineal, vecinos y árboles» no basta. Debes explicar cómo cada método produce una estimación, qué supuesto necesita, qué puede fallar y cómo lo comprobarías.

## PRÁCTICA · DATOS FICTICIOS

## 32 · Respuestas de fundamentos

1. Supervisado, regresión, fondo-mes. Hay respuestas observadas para aprender una regla numérica. No hay una muestra independiente por cada instrumento de la cartera.
2. Cambia la naturaleza de la respuesta y deben justificarse los cortes. Corresponde estudiar clasificación, potencialmente ordinal, y métricas que respeten el orden y los costos de equivocarse. Spearman no ajusta un modelo ni basta para evaluar calibración o categorías.
3. No. Falta una regla que conecte grupos y Y, aprendida sin consultar la prueba. Puede construirse una referencia con promedios por grupo, pero entonces se evalúa ese procedimiento completo.
4. Predice 8. Falta conocer variables, escalas, distancias, calidad de datos, fechas y procedencia de los vecinos. El promedio es sencillo; la comparabilidad es el punto decisivo.
5. Residuo 3; pérdida absoluta 3; pérdida cuadrática 9. Las dos primeras conservan la unidad de Y; la pérdida cuadrática tiene unidades al cuadrado.
6. Los objetivos son 80 y 62; se prefiere la segunda solución entre esas dos. Eso no demuestra superioridad fuera de muestra. Si  $\lambda$  se selecciona entre candidatos, esa selección se hace con validación; el procedimiento se evalúa después en prueba.
7. Predice 8,4. La tasa reduce el tamaño de cada actualización y regula el aprendizaje; permite construir la aproximación gradualmente, aunque no elimina el sobreajuste por sí sola.
8. El preprocesamiento utilizó la distribución de datos reservados. Debe aprenderse dentro de cada entrenamiento, junto con el resto de pasos ajustables.
9. Puede mostrar aproximación temporal en fondos conocidos. No acredita transferencia a fondos no vistos; además, la superposición de Y exige cautela al interpretar independencia.
10. Sobreajuste, entre otras posibilidades. Hay que revisar complejidad y diferencias de distribución, no seleccionar por ajuste de entrenamiento.
11. No. Con idéntica entrada entrega idéntica salida. Hace falta información adicional para distinguir esos casos, o aceptar incertidumbre residual.
12. Las redes también hacen regresión tabular y existe SVR. Se pueden postergar por alcance, carga de ajuste y falta de una necesidad concreta demostrada, no por esas supuestas incapacidades.
13. Ridge prueba una estructura global lineal regularizada; KNN, similitud local; boosting, reglas no lineales e interacciones. Se comparan bajo el mismo protocolo y contra referencias simples.
14. Refuerzo queda fuera de la tarea de estimación actual; random forest es compatible y se posterga por alcance.
15. Preferiría ridge si su desempeño es comparable o más estable y su menor complejidad resulta útil. Revisaría X si fallan todos, hay vecinos distantes, errores asociados a cobertura o carteras indistinguibles con Y muy diferentes.

## PRÁCTICA · PROYECTO

## 33 · Ejercicios sobre tu investigación

Responde primero sin mirar el capítulo siguiente. Estos ejercicios combinan cifras documentadas con razonamiento metodológico; no requieren ejecutar código.

### 1. Denominadores

¿Por qué 63.633 no coincide con  $31.973 + 659 + 1.761$ ? ¿Qué representa la diferencia? ¿Por qué 18,7 millones de cuotas no es el tamaño de entrenamiento?

### 2. Volatilidad

Si la desviación semanal es 0,02, calcula  $Y1$ . ¿Qué denominador usa la varianza muestral con 260 retornos? ¿Cuántas anclas hacen falta?

### 3. Cola izquierda

Si  $Y1=10$  y el percentil de retornos estandarizados es  $-1,98$ , calcula  $k$  e  $Y2$ . ¿Qué pasa si  $s=0$ ? ¿Podrías calcular  $k$  directamente desde el percentil de retornos porcentuales sin estandarizarlos?

### 4. KNN real

Con dos vecinos de  $Y=8$  y  $Y=10$  a distancias 2 y 8, compara media uniforme con ponderación inversa por distancia. ¿Cuál de esas formas corresponde a la configuración de 15 vecinos informada?

### 5. Selección de variables

¿Por qué se excluyeron `cobertura_emisor` y `cobertura_pct`? ¿Qué tiene de especial que la regla se aplicara en TRAIN? ¿Retener vencimiento medio significa que todos sus valores eran observados?

### 6. Métricas y decisión

Boosting tuvo menor MAE en VALID, pero KNN tuvo 95 subestimaciones graves frente a 139. ¿Hay contradicción? ¿Cómo cambia la respuesta si el costo de subestimar es especialmente alto?

### 7. Auditoría

Dos factores alternativos cambian un retorno diario en 1,78529 pp. ¿Puedes concluir que  $Y1$  cambia en esa cantidad? ¿Puedes afirmar que afecta directamente TEST 2024-2026?

### 8. Alcance de resultados

Un IC por RUN excluye cero para la mejora relativa. ¿Demuestra independencia entre fondos, validez económica de  $Y$  y pronóstico futuro? ¿Cuatro escenarios con iguales filas y métricas prueban cuatro robusteces diferentes?

### 9. Transferencia

¿Por qué un fondo no visto en TRAIN puede seguir siendo un fondo maduro? ¿Qué diferencia eso de aplicar a un fondo realmente nuevo?

### 10. Próxima decisión

¿Qué verificarías antes de agregar una red neuronal: factores y ventanas, otra búsqueda en TEST o ambas? Justifica.

## PRÁCTICA · SOLUCIONES

## 34 · Respuestas sobre tu investigación

1. Las tres particiones activas suman 34.393. Quedan 29.240 filas en reserva por grupo de RUN o fecha no admitida. Las cuotas son insumos temporales por serie, no filas independientes de aprendizaje.
2.  $Y1=100 \times \sqrt{52} \times 0,02 \approx 14,4222 \%$ . La varianza muestral divide por 259. Se requieren 261 anclas para 260 retornos consecutivos.
3.  $k=\text{máximo}(1, -1,98/-1,65)=1,2$ ;  $Y2=12 \%$ . Si  $s=0$ ,  $Y1=0$  y  $Y2$  indefinida. No usar el percentil sin estandarizar: cambia el objeto y la escala del ajuste.
4. Media uniforme: 9. Ponderación inversa:  $(8/2+10/8)/(1/2+1/8)=8,4$ . La corrida utilizó ponderación por distancia con 15 vecinos, no el promedio uniforme del ejemplo básico.
5. Eran constantes: 31.973 valores presentes y un único valor distinto. La decisión no se tomó observando TEST. Vencimiento medio tenía 22.063 presentes y 9.910 ausentes; su inclusión exige imputación y una explicación de cobertura.
6. No hay contradicción. MAE y número de grandes subestimaciones son funciones distintas del error. Si subestimar cuesta más, debe fijarse un criterio acorde antes de seleccionar. No cambiar retrospectivamente la función para favorecer al ganador preferido.
7. No. La volatilidad depende de toda la ventana, de agregación semanal y de serie seleccionada. El retorno de 2008 no forma parte directa de ventanas terminadas en 2024-2026, pero puede afectar entrenamiento antiguo. Hay que reconstruir vínculos evento-ventana-partición.
8. No. El bootstrap describe incertidumbre bajo su esquema y condicionado a datos y predicciones. No arregla medición ni prueba futuro. Escenarios con idéntica muestra no constituyen comparaciones de cohortes realmente depuradas.
9. No visto significa que sus filas no se usaron al ajustar. Puede tener más de cinco años de historia. Un fondo recién creado no tiene Y quinquenal propia; transportarle una regla aprendida en fondos maduros requiere evidencia y condiciones adicionales.
10. Primero verificaría mediciones, ventanas y diseño; no usaría TEST como buscador de hiperparámetros. Una red es viable, pero no soluciona factores mal interpretados ni exposición ausente. Si se amplían modelos, se necesita un protocolo de selección y una evaluación independiente.

**RECUERDA:** comprender el proyecto implica saber decir qué se calculó, qué se documentó y qué todavía no se ha probado.

## MEMORIA · CONCEPTOS

## 35 · Lo que debo explicar de memoria

1. Aprendizaje supervisado: aprender una relación a partir de ejemplos con características y respuesta conocidas.
2. Regresión: estimar una respuesta cuantitativa; no exige una relación lineal.
3. Clasificación: estimar categorías o sus probabilidades.
4. Característica: atributo utilizado como entrada del modelo.
5. Parámetro: cantidad aprendida al ajustar; hiperparámetro: elección que regula ese ajuste.
6. Generalización: desempeño en observaciones que no se utilizaron para aprender o seleccionar el procedimiento.
7. Sobreajuste: adaptación a particularidades del entrenamiento que no se sostienen fuera de él.
8. Regularización: restricción de flexibilidad para controlar inestabilidad y sobreajuste.
9. Ridge: regresión lineal con penalización del cuadrado de los coeficientes, normalmente sin penalizar el intercepto.
10. KNN: estimación mediante respuestas de observaciones cercanas en el espacio de características.
11. Gradient boosting: combinación secuencial de modelos que aproxima direcciones de reducción de la pérdida.
12. Interacción: el efecto predictivo de una variable depende del valor de otra.
13. Fuga de información: uso durante ajuste o selección de información que no debería estar disponible en esa etapa o aplicación.
14. MAE: promedio de errores absolutos; RMSE: raíz del promedio de errores cuadrados.
15. Validación temporal: evaluación que respeta el orden de disponibilidad de la información.
16. RUN-mes: unidad analítica del estudio; no equivale a posición ni a serie-día.
17. Y1: volatilidad retrospectiva anualizada de 260 retornos semanales; Y2: ajuste secundario de cola.
18. As-of: información realmente disponible en la fecha de aplicación, no solo referida a esa fecha.
19. Robustez: estabilidad frente a cambios pertinentes efectivamente ejecutados; no repetición de etiquetas.
20. Equivalencia regulatoria: exige convenciones y evidencia adicionales; no se deduce de una buena métrica.

## DEFENSA ORAL · ARGUMENTO Y OBJECIONES

## 36 · Defender la elección en voz alta

Mi problema es aproximar una medida numérica de volatilidad histórica usando composición de cartera. Por eso es regresión supervisada sobre observaciones fondo-mes. La representación es agregada y no contiene el look-through completo: un buen algoritmo puede aprovechar esa información, pero no inventar exposiciones ausentes.

Comparé ridge, KNN y histogram gradient boosting porque representan hipótesis diferentes. Ridge evalúa una relación lineal regularizada; KNN formaliza similitud local; boosting permite umbrales e interacciones. No sostengo que sean los únicos métodos válidos: random forest, elastic net, GAM o SVR son alternativas que pueden incorporarse con una hipótesis y un protocolo definidos.

En el experimento documentado se usaron 26 variables seleccionadas en TRAIN, con fondos y períodos separados. Boosting obtuvo menor MAE de TEST que las referencias y los otros dos modelos, pero esa ventaja corresponde a configuraciones concretas y a una muestra seleccionada. No demuestra superioridad universal.

La medida objetivo es retrospectiva. Persisten controles de factores, moneda, trazabilidad y disponibilidad, y las sensibilidades reportadas no acreditan una cohorte depurada distinta. Por tanto los resultados son exploratorios y no se presentan como pronóstico futuro, solución probada para fondos recién creados ni indicador oficial de CMF.

### **Si te preguntan «¿por qué no random forest?»**

Es una alternativa directa y pertinente. Su ausencia inicial delimita el alcance; no se ha demostrado que sea inferior. Incorporarlo serviría para contrastar otro ensamble flexible, pero no reemplaza resolver calidad de datos y evaluación.

### **Si te preguntan «¿por qué ridge y no solo regresión lineal?»**

Los porcentajes y atributos relacionados pueden producir inestabilidad. Ridge restringe el tamaño conjunto de los coeficientes. Su valor debe evaluarse fuera de muestra; regularizar no garantiza ganar ni convierte coeficientes en causalidad.

### **Si te preguntan «¿qué aporta el estudio?»**

Documenta cuánto puede aproximarse  $Y_1$  con una representación observable de  $X$  bajo un experimento definido. Aporta también una cadena de datos y límites verificables. La ausencia de una certificación regulatoria no invalida esa pregunta estadística, pero delimita la conclusión.

## BIBLIOGRAFÍA • GUÍA DE LECTURA

## 37 • Dónde estudiar los métodos

[ISL] James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. (2023). An Introduction to Statistical Learning: with Applications in Python. Springer. DOI: 10.1007/978-3-031-38747-0. Se consultó el PDF consolidado adjunto. Es la edición Python; no se confunde con la segunda edición del texto con aplicaciones en R.

[ESL] Hastie, T.; Tibshirani, R.; Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Segunda edición, Springer. El archivo adjunto puede ser una impresión posterior. Se usa paginación interna.

[BISHOP] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. §3.1.4: mínimos cuadrados regularizados. Puede estudiarse después de la introducción de ISL.

Los libros sostienen mecanismos, supuestos y evaluación; no las cifras de fondos chilenos. Las páginas siguientes son impresas, no el número del visor. La revisión de referencias fue dirigida a estos temas, no una certificación exhaustiva de todos los textos.

| Tema                         | ISL: primera lectura | ESL: profundización          |
|------------------------------|----------------------|------------------------------|
| Aprendizaje y generalización | §§2.1-2.2            | Cap. 2 y 7                   |
| KNN                          | §3.5, pp.111-115     | §§2.3.2 y 2.5                |
| Ridge                        | §6.2.1, pp.240-243   | §3.4.1, desde p.61           |
| Lasso y selección            | §§6.2.2-6.2.3        | §3.4.3                       |
| Elastic net                  | Contrastar L1 y L2   | §3.4.3 y §18.4               |
| PCA / PCR                    | §6.3.1               | Cap. 3                       |
| Árboles                      | §8.1.1, desde p.331  | Cap. 9                       |
| Random forest                | §8.2.2, pp.346-347   | Cap. 15                      |
| Boosting                     | §8.2.3, pp.348-350   | §§10.10-10.12, pp.358-365    |
| Fuga por selección           | §5.1 como base       | §7.10.2, pp.245-246          |
| GAM / SVR / redes            | Ampliación posterior | §9.1 / §12.3.6 / §§11.3-11.4 |

## FUENTES • TRAZABILIDAD

## 38 • Fuentes del proyecto y jerarquía

[PROY-X] Hito\_CMF\_XY\_V1\_Memoria\_Tecnica.docx. Memoria técnica al 12-10-2026. Documenta adquisición e integración; sus cifras exploratorias anteriores no se sustituyen silenciosamente por las del experimento posterior.

[PROY-Y] RESOLUCION\_METODOLOGICA\_Y5\_OP\_V1.md. Decisión del 13-10-2026. Fija Y1, Y2, anclas, abstenciones, continuidad y limitaciones. Prevalece sobre decisiones de Y aún abiertas en la memoria anterior.

[CIERRE] INFORME\_CIERRE\_HITO\_XY\_V1.md. Informe de cierre experimental del 18-10-2026, versión consultada 1. §§3-5: X, Y y cohortes; §6: particiones y modelos; §7: métricas; §§8-10: controles y fuentes. Se recuperó y leyó completo para la auditoría precedente. Es respaldo documental, no acceso a bases ni predicciones.

[DICTAMEN] Dictamen\_Validaciones\_Seminario\_II\_2026-10-19.md, versión consultada 1. II.2: selección de variables; II.3: FA/FR; II.4: moneda; II.5: disponibilidad; II.6: trazabilidad; II.7: robustez. III: protocolo pendiente.

[AUDITORÍA] AUDITORIA.md, 23-10-2026, elaborada en esta revisión. Contrasta el manual web con resolución, informes y libros; verifica texto, sumas y ejemplos. No reejecuta los modelos ni abre la base original.

[PROY-I] Informe\_MVukusic.pdf. Informe de Seminario I del 2-10-2026. Antecedente del problema. Sus resultados piloto no se transfieren como evidencia del panel posterior.

### Referencia regulatoria: alcance explícito

Los documentos del proyecto citan una nota técnica CMF de junio de 2026 y, entre otras fuentes, las circulares 1.333 y 1.850. Este manual no verifica la vigencia, la sucesión histórica de esas normas ni la equivalencia de los factores en todos los períodos. La expresión B se presenta como referencia documental del proyecto, no como dictamen de cumplimiento.

Antes de afirmar homologación se necesita identificar documento primario, versión, páginas, definiciones de  $\sigma$ ,  $k$ ,  $w$ , moneda y aplicación temporal. No atribuir carácter oficial a Y1 o Y2 por semejanza formal.

RECUERDA: cronología no significa aprobación automática. La resolución fija decisiones, el cierre reporta resultados y el dictamen documenta controles; ninguno sustituye las fuentes originales cuando se exige reproducción independiente.

## 39 · Qué quedó corregido y qué no

Esta edición integra las correcciones que podían resolverse editorialmente con la evidencia disponible. Se revisó la presentación del PDF y se conservó la distinción entre ejemplos, resultados documentados y verificaciones pendientes.

### Correcciones incorporadas

- Y2 tiene estandarización, interpolación, umbral de estabilidad y distinción respecto de B.
- Se separan cálculo por serie-moneda y unidad analítica RUN-mes.
- Se concilian cohorte, particiones y reserva; no se presentan conteos antiguos como vigentes.
- Se distingue KNN uniforme didáctico de KNN-15 ponderado por distancia.
- Se informa  $\alpha=100$  y configuración de histogram boosting, clipping y ausencia de búsqueda exhaustiva documentada.
- Se incorporan resultados VALID/TEST, incertidumbre por RUN y heterogeneidad anual.
- Se restituyen tablas, notación legible y una figura científica con datos ficticios identificados.
- Las alternativas viables se postergan por alcance, sin inventar evidencia de inferioridad.

### Controles que siguen abiertos

No se reconstruyeron 261 anclas y 260 retornos por clave activa, no se certificaron factores ni moneda funcional, no se recuperaron fechas reales as-of, no se recalcularon predicciones ni bootstrap del experimento y no se ejecutó una depuración efectiva de cohortes.

El gráfico didáctico sí se calculó para esta edición, pero usa datos sintéticos y otros parámetros; no constituye nueva evidencia sobre los fondos.

### Criterio para avanzar

Ya puedes estudiar y defender el razonamiento de selección: pregunta → datos → respuesta → supuestos de cada familia → comparación → límites. Para reforzar la conclusión empírica, el siguiente paso es cerrar los controles con evidencia y después valorar ampliaciones del experimento.

**RECUERDA:** comprender una técnica incluye reconocer qué necesitarías observar para dejar de preferirla. El objetivo es construir una explicación revisable, no memorizar una defensa inmune a los datos.